第32卷 第17期 计 算 机 工 程 2006年9月
Vol.32 № 17 Computer Engineering September 2006
·专题论文· 文章编号：1000—3428(2006)17—0089—03 文献标识码：A 中图分类号：TP391

# Web文本褒贬倾向性分类研究

柴玉梅，熊德兰，昝红英
(郑州大学信息工程学院，郑州 450052)

**摘 要：**分析了网页内容褒贬色彩的客观性和褒贬倾向性分类的可行性，介绍了Web文本褒贬倾向性分类的原理和实现方法，并将已有的特征选择方法与褒贬特征提取技术结合起来，使用几种分类算法实现了名人网页的褒贬倾向性分类，达到了较好的分类效果。
**关键词：**褒贬倾向性分类；褒贬特征提取；分类

# Research on Web Text Appraisable Classification

CHAI Yumei, XIONG Delan, ZAN Hongying
( College of Information Engineering, Zhengzhou University, Zhengzhou 450052)

**【Abstract】** This paper analyzes the objectivity of appraisal of content in Webpages and feasibility of appraisable classification, and introduces theory and implement method of Web text appraisable classification. Combined existing technique of feature selecting with the method of appraisable feature extracting, the thesis implementes appraisable classification in some celebrities' Webpages by several classification algorithm, and achieved preferable effects.
**【Key words】** Appraisable classification; Appraisable feature extracting; Classification

Web文本自动分类是Web数据挖掘的研究热点之一，它能有效地组织和管理Web资源、提高信息检索效率。网页自动分类可以使用文本分类的相关方法。目前，英文文本自动分类已经取得了很好的成绩，提出了多种分类算法，还建立了Reuters等标准的分类语料和统计的评价方法[1]。文献[2,3]分别介绍了使用网页中的标记和元数据等信息进行网页分类的方法以及使用多分类器进行分类的基本方法和步骤，并在实验中证明算法的有效性。

目前，国内在中文文本分类和中文网页分类领域也进行了大量的研究。但现有的Web文本分类大多是根据网页所涉及的主题进行分类，如将网页分为政治类、军事类、经济类等，而根据网页中作者对所描述内容的观点、态度等主观情感进行分类的研究很少。网页内容的褒贬倾向就是明显反映作者观点、态度的感情色彩之一，网页文本褒贬倾向性分类是未来多角度、立体性文本分类的一个重要研究方面[4]，具有广泛的应用前景。

本文结合已有的文本自动分类技术，探讨Web文本褒贬倾向性分类的基本原理和实现方法。介绍了网页内容的褒贬客观性和Web文本褒贬倾向性分类的可行性，并给出了分类的主要过程；描述了褒贬倾向性分类中的几个关键问题，并提出了后期工作的设想。

## 1 褒贬倾向性分类的定义和工作过程

### 1.1 褒贬倾向性分类的可行性

语言的感情色彩是客观存在的，因而，网页内容通常不是单一的事件或人物的描述，它还传递了网页作者本人或他所代表的集团(派别)的立场、观点、情感态度等信息。名人是人们上网搜索的热点之一，名人相关网页是增长速度最快的网页种类之一。名人的相关介绍、新闻事件、社会评论等都会不同程度地带有作者的情感倾向，传媒、普通民众、名人自己都需要及时地了解这些正面或反面的报道和评论。

名人网页进行褒贬倾向性分类就是通过分析网页中语言的感情色彩，从褒贬倾向角度对网页进行分门别类。根据网页情感态度的不同，将网页分为正面褒扬类、中立类、反面贬斥类3种，对网页褒贬倾向的基本划分标准如下：

(1)[正面褒扬类]：名人相关的正面报道，对所描述的内容持肯定语气，带有明显的称赞、颂扬、赞赏或哀悼、惋惜等意味；

(2)[中立类]：对名人或名人相关事件的报道客观公正，无个人评价性说明；

(3)[反面贬斥类]：名人相关的负面报道，对所描述的内容持否定、质疑、讽刺等语气或鄙视、批评、痛斥等色彩；

名人网页褒贬倾向性分类就是将汉语语言理解融入到文本分类技术中，通过对已知褒贬网页集合(即训练集)的分析理解，得到一个明确的褒贬分类规则，去确定未知的网页集合(即测试集)的类别。可以用数学公式表示如下：

$f:A\rightarrow B$

其中，A为待分类的网页集，B为褒贬倾向性分类后的网页类别集。

### 1.2 分类过程

本文提出的名人网页褒贬倾向性分类工作的主要过程如图1所示。从图中可以看出，整个工作分为训练过程和分类过程两个部分，这与通常的ATC技术大致相同。所不同的是，提出了褒贬特征提取方法。因为在实际网页中，褒贬特征信息出现很少，有的网页甚至没有褒贬词语的出现，因而按照自动分类中常规的特征选择方法很容易被当作无关信息而过

**基金项目：**国家“973”计划基金资助项目(2004CB 318102)；河南省自然科学基金资助项目(0211020110)
**作者简介：**柴玉梅(1964—)，女，副教授，主研方向：Web数据挖掘；熊德兰，硕士生；昝红英，副教授
**收稿日期：**2006-04-29 **E-mail：**yulan1021@tom.com

滤掉，但这些褒贬特征具有很强的褒贬区分能力，因此，首先单独将其提取出来。

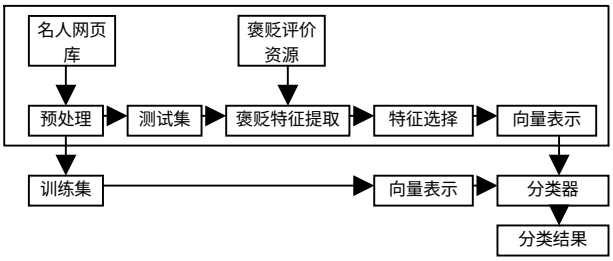


**图 1 网页褒贬倾向性分类过程框架**

## 2 褒贬倾向性分类中的关键技术

### 2.1 网页文本表示

网页与普通的文本有很大的不同，Web 页面中包含了大量的超文本标记、超链接信息以及无用的噪声信息等。因而，Web 文档需要经过一系列的预处理过程才能用于分类。预处理过程包括噪声消除、标签过滤、信息提取、中文分词等步骤，最终将网页表示为计算机可以识别的有效方式。

向量空间模型(VSM)是目前文本分类中使用较多、效果较好的一种文本特征表示方法。它将每个文本表示为特征空间的一个向量，形如：$d_i=\{(t_{i1},w_{i1}),(t_{i2},w_{i2}),...,(t_{in},w_{in})\}$，其中$t_{in}$为特征项，它可以是字、词或短语；$w_{in}$为特征项的权重，表示$t_{in}$在文本中的重要程度。权重是根据特征项在文本中出现频率、位置等信息计算得到的，常用的权重计算方法有TF、TFIDF等。该研究中，采用了一种比较普遍的TFIDF公式：

$$W(t,\vec{d}) = \frac{tf(t,\vec{d}) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t\in\vec{d}} [tf(t,\vec{d}) \times \log(N/n_t + 0.01)]^2}}$$

其中，$W(t,\vec{d})$为特征项$t$在文本$\vec{d}$中的权重，而$tf(t,\vec{d})$为特征项$t$在文本$\vec{d}$中的频率，$N$为训练文本的总数，$n_t$为训练文本集中出现$t$的文本数，分母为归一化因子。该数值越大，特征项$t$反映$\vec{d}$的能力越好；该数值越小，特征项$t$反映$\vec{d}$的能力越差。

### 2.2 特征提取

按照上述 VSM 表示文本后，向量空间的维数往往十分庞大，这就使得分类算法非常低效，而且并不是所有的词语对文本分类都有贡献。为了尽可能提高分类的速度和精度，应去除那些与类别无关或关联性不大的词语，筛选出最具有代表性的词条作为特征项，这就是特征选择的过程。

通常的做法是构造一个评估函数，对特征项集合中的每个特征项进行独立的评估，得到其评估分值(即权值)，然后对所有的特征项按照其权值大小进行排序，最后选择预定数目的特征项作为特征结果。目前，文本分类中常用的特征评估函数有文档频率(DF)、互信息(MI)、信息增益(IG)、期望交叉熵(ECE)、文本证据权(WET)等[5]。

为了提高褒贬分类的精度，提出了依赖褒贬评价资源提取褒贬特征的方法。褒贬特征为网页中出现的褒义词、贬义词以及具有褒贬色彩的短语、语法结构等，褒贬评价资源是指人为构建的褒贬词典和褒贬评价模板，具体方法参见文献[6]。褒贬特征抽取过程就是读取切分后的网页文本中的每个词语，扫描褒贬词典和评价模板，提取具有褒贬色彩的词语或结构。将上述的特征选择过程称为基本特征提取，则两次特征提取过程描述如下(假定基本特征提取方法为 MI)：

Step1 褒贬特征提取(Appraisal Feature Extracting，AFE)

Step1.1 初始条件下，褒贬特征项集合$T1=\Phi$；

Step1.2 读入一个预处理后的网页文本$\vec{d}$；

Step1.3 读取一个完整的句子；

Step1.4 逐个读取中的切分后的一个词语 $t$ 扫描词典，如果 $t$∈褒贬词典，则将 $t$ 加入到 $T1$ 中；将 $t$ 所在的句子与模板匹配，如匹配成功，则将该模板编号加入到 $T1$ 中；

Step1.5 读取下一个句子，转 Step1.4，若该文本结束，则读取下一个文本，转 Step1.3；

Step2 基本特征提取(MI)

Step2.1 初始条件下，特征项集合 $T2$ 中包含类别 $Ci$ 中出现的所有特征词；

Step2.2 对于每个词 $t$，计算它与类别 $Ci$ 的互信息$MI(t,Ci)=\log(\frac{P(t\mid Ci)}{P(t)})$

Step2.3 对于类别 $Ci$ 中的所有特征词，根据上面计算后的互信息量数值大小进行排序；

Step2.4 选取一个特征向量维数阈值，进行向量维数的压缩，同时根据分类实验结果适当地调整该阈值，进而确定最佳值；

Step3 将两次特征提取的结果合并到总特征项集合中，即$T=T1\cup T2$。

### 2.3 分类算法

分类算法是分类技术的核心，目前存在多种基于向量空间模型的文本分类算法，例如，简单向量距离分类法、最近 K 邻居、贝叶斯方法、支持向量机、神经网络，最大平均熵等。分类过程一般分为训练和分类两个阶段，本研究中实现的领域内名人网页褒贬倾向性分类过程描述如下：

(1)训练阶段

1)定义类别集合 C={C1,C2,C3}；

2)经过人工标记和预处理后的训练网页文本集 U={U1,U2,…,Un}；

3)统计 U 中出现的褒贬特征矢量 T1 和普通的特征词矢量 V(Ui)；

4)使用某种评估函数进行特征选择，确定每个类的特征矢量 V(Ci)；

(2)分类阶段

1)对于测试网页文本集合 D={D1,D2,…,Dn}中每个待分类文本 Dk，计算其特征矢量 V(Dk)与每个类别矢量 V(Ci)的相似度 $\mathrm{Sim}(Dk,Ci)$，其计算公式为

$$\mathrm{Sim}(Dk,Ci) = \frac{V(Dk)\bullet V(Ci)}{\|V(Dk)\|\times\|V(Ci)\|}$$

2)选取相似度最大的一个类别 $\arg\max_{Ci\in C} sim(Dk,Ci)$ 作为 Dk 的类别。

## 3 实验及结果分析

### 3.1 实验数据集说明

实验所选用的数据集是北京大学计算语言学研究所部分研究人员进行名人网页相关度评价研究[7]时所收集的名人网页。该网页集合包括政府、科教、工商、文化、传媒、演艺、体育等 7 个领域的 8 000 多个名人网页，网页格式及名人相关信息比较全面，基本能满足实验需求。

通常，单纯地评定一个人物或事物的好坏优劣是不够准确的，褒贬倾向的判定需要有一定的限制。因此，本文所做的名人网页褒贬倾向性分类工作是在特定的领域范围内进行的，实验中使用的各领域名人网页的数量如表 1 所示。

**表 1 人工分类后的实验数据集**

| 类别标记 | 名人所属的领域 | | | | | | |
|---|---|---|---|---|---|---|---|
| | 政府 | 科教 | 工商 | 文化 | 传媒 | 演艺 | 体育 |
| C1（褒扬） | 786 | 692 | 138 | 165 | 342 | 452 | 221 |
| C2（中立） | 894 | 543 | 182 | 275 | 495 | 508 | 453 |
| C3（贬斥） | 132 | 125 | 84 | 84 | 182 | 325 | 94 |
| 总计 | 1812 | 1360 | 404 | 524 | 1019 | 1285 | 768 |

从表 1 可以看出，各类数据样本集是不够均匀的，类别 C2 的网页数量最多，类别 C3 的网页数量最少，且明显低于前两个类别，只有总数 13%左右。尤其是政府、科教、文化等领域，且褒扬和中立倾向的区分也太不明显。

### 3.2 实验结果及其分析

在 Windows XP 系统下采用 Microsoft Visual C++ 6.0 编程环境中对上述理论进行实验验证。实验中，采用 5-折交叉确认法，按照 4∶1 的比率选取训练集和测试集，分别使用简单距离向量法、KNN 和 Naïve Bayes 算法进行分类实验。3 种分类算法下实验结果如表 2 所示。

**表 2 3 种算法分类结果的 Micro-F1 值**

| 分类算法 | 政 府 | 科 教 | 工 商 | 文 化 | 传 媒 | 演艺 | 体育 |
|---|---|---|---|---|---|---|---|
| 简单向量距离分类法 | 0.425 | 0.357 | 0.291 | 0.358 | 0.521 | 0.506 | 0.489 |
| KNN | 0.253 | 0.540 | 0.417 | 0.562 | 0.654 | 0.667 | 0.574 |
| Naïve Bayes | 0.384 | 0.456 | 0.366 | 0.331 | 0.557 | 0.542 | 0.556 |

从表 2 可以看出，褒贬倾向较明显的传媒、演艺和体育领域分类效果较好，而网页褒贬感情较单一的政府、工商等领域分类效果不明显。同时，KNN 算法效果明显优于另外两种，平均约为 0.49，这说明在实验样本不平衡时，采用 KNN 分类效果较好。

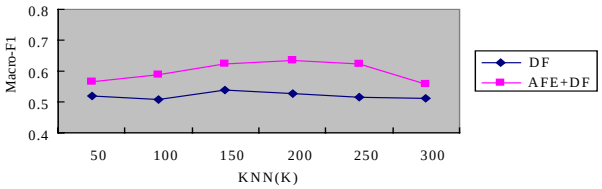


**图 2 DF 特征选择方法下加入褒贬特征提取后的分类结果**

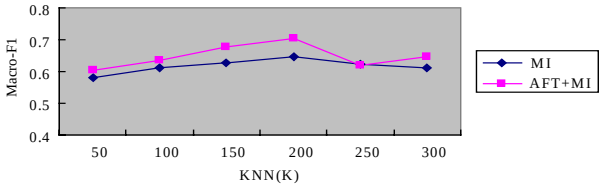


**图 3 MI 特征选择方法下加入褒贬特征提取后的分类结果**

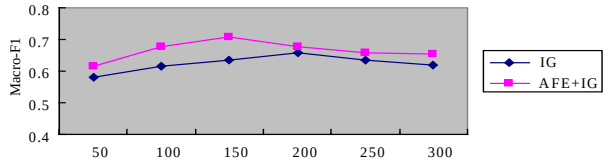


**图 4 IG 特征选择方法下加入褒贬特征提取后的分类结果**

选取上述褒贬色彩较明显的传媒领域内名人网页，使用分类效果较好的 KNN 分类算法，对不同特征选择方法下使用褒贬特征提取(AFE)后的分类效果进行对比实验。DF、MI、IG 3 种特征选择方法下的实验结果图 2、图 3、图 4 所示，其中，X 轴表示 K 的不同取值，Y 轴表示 5 轮分类结果的 Marco-F1 值。

由图 2、图 3、图 4 可以看出：在不同的特征选择方法下，使用褒贬特征提取技术能将分类结果提高 7%～13%。这表明：通用的特征选择方法并不能很好地评估文本中褒贬色彩的词语或结构，使用褒贬特征提取技术，能够将那些在文本中出现较少而含有重要褒贬信息的特征项单独区分开来，从而提高分类的精度。

总体来说，分类结果一般在[0.4，0.7]之间，而政府、工商等领域的分类结果则更低。分析认为：褒贬性是感情色彩的一部分，除了提取的字、词、短语之外，文本中的句式、修辞方式、标点符号等都可能是作者情感的体现，需要更精确的分析提取方法；另一方面，人工分类也不可避免地带有个人的主观色彩，使得分类结果比较存在一定的偏差。

## 4 结束语

本文提出了从情感角度进行文本分类的新方法，探讨了领域内名人网页褒贬倾向性分类的主要工程和具体实现方法，结合实验结果分析了不同特征选择方法和分类算法下的分类性能。文中提供的褒贬感情划分依据和人工分类文档对后续的研究也将有很大的帮助，并为其他类似的研究提供了可以借鉴和模拟的范例。

今后我们将从褒贬色彩的多样性、复杂性等方面考虑，分析不同因素对分类性能的影响，进一步完善褒贬特征提取技术，提高分类精度。同时，尝试在其他应用背景下的褒贬倾向性分类，如新政策法规的民众反映、产品质量评价、电视文艺作品的评论等。

## 参考文献


1 Yang Yiming, Liu Xin. A Re-examination of Text Categorization Methods[C]. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999: 35-39.

2 Ghani R, Slattery S, Yang Y. Hypertext Categorization Using Hyperlink Patterns and Meta Data[C]. Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001: 178-185.

3 Hwanjo Yu, Chang K. Heterogeneous Learner for Web-page Classification[C]. Proceedings of the 2002 IEEE International Conference on Data Mining, IEEE Computer Society, 2002: 538-545.

4 杨建良,王永成.文本自动分类的研究与发展[Z]. http://hu003.chianlibs.net/zhaiyao.asp?titile=6165, 2005.

5 Lewis D D. Feature Selection and Feature Extraction for Text Categorization[C]. Proceedings of Speech and Natural Language Workshop. San Francisco: Morgan Kaufmann , 1992-02: 212-217.

6 熊德兰,柴玉梅,咎红英. 基于内容的名人网页褒贬性评价[J].平顶山工学院学报, 2005, 14(4): 47-49.

7 咎红英, 苏玉梅. 名人网页的相关度评价[J]. 中文信息学报, 2003, 28(5): 27-33.