第33卷 第15期 计 算 机 工 程 2007年8月
Vol.33 No.15 Computer Engineering August 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)15—0178—03 文献标识码：A 中图分类号：TP18

模糊综合评判理论在网页自动分类中的应用

吕英杰，叶 强，李一军
（哈尔滨工业大学管理学院，哈尔滨 150001）

摘 要：互联网的发展使网页分类技术成为了研究热点，针对传统的基于统计的分类技术只能提供精确分类的情况，该文运用模糊综合评判理论进行文本分类，根据自然语言的复杂性和理解的不确定性，使分类后的网页以一定的概率分属于各个类别，真实地反映了网页信息。实践表明，用户可以根据所得概率进行灵活处理，以满足各种实际需要。

关键词：网页分类；特征提取；模糊综合评判

Application of Fuzzy Comprehensive Evaluation Theory in Automatic Web Page Categorization

LV Ying-jie, YE Qiang, LI Yi-jun
(School of Management, Harbin Institute of Technology, Harbin 150001)

【Abstract】Web page categorization has become popular with the fast growing of Internet. Because of the disadvantages of traditional statistical approach, this paper proposes a categorization method based on fuzzy comprehensive evaluation. Due to the complexity of natural language and uncertainty of comprehension, the Web page belongs to every class with its probability, which can reflect the true information. According to the probability, users can deal with the Web pages flexibly to satisfy their needs.

【Key words】Web page categorization; extraction of features; fuzzy comprehensive evaluation

近年来，许多学者在文本分类模型和分类算法上进行了大量研究，提出了诸如朴素贝叶斯分类(naive bayes)等多种基于统计的文本分类技术[1,2]，分类准确率达到80%以上。由于自然语言在描述和理解方面具有高度的不确定性和模糊性，因此给文本类别的识别带来一定的模糊性。网页的设计比较随意，通常包含大量广告、注释和版权等信息，同一个网页也有可能包含多个主题。这些情况会导致网页分类的不确定性，运用传统的、基于统计的文本分类技术进行精确的分类，效果并不理想。

本文使用模糊数学理论对中文网页进行分类，使分类后的网页不再非常明确地属于某一类或不属于某一类，而是以一定的隶属程度属于各个类别。这样的模糊分类能提供更多、更真实的信息。采用模糊技术进行分类的最大优势在于：分类结果可以反映出分类过程中的不确定性，有利于用户根据结果进行决策。在具体处理中，可以根据最大隶属度原则把网页归入到隶属程度最大的类别中完成分类，也可以通过设定阈值把含有多主题的网页归入相应的多个类别中，通过阈值有效地过滤出与任何给定类别均不符的异常网页。经过多种处理手段，可以使归类结果更有效地反映网页的真实信息。

1 网页归类算法

1.1 训练网页

为了使计算机能自动地进行网页分类，必须先对网页文本进行训练。其中主要步骤包括：选取样本网页，分词处理，特征提取等。预先定义好网页类别后，在每一类别中选取一定数目的网页，对网页信息进行分词处理。根据每个词在样本网页各部分出现的次数，计算每个词在该网页中的相对词频，可以运用TF-IDF公式计算该词的相对词频，公式如下：

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

其中，$W(t,\bar{d})$为词 t 在网页 $\bar{d}$ 中的相对词频；$tf(t,\bar{d})$为词 t 在网页 $\bar{d}$ 中的绝对词频；N 为训练网页的总数，n_t 为训练网页集中出现 t 的网页数。

网页中的词数量很大，不同词对反映网页主题作用也是不同的，像有些高频词“的”等，虽然出现频率很高，但对网页主题却没有什么反映。因此，必须对网页进行特征词提取。目前特征提取的方法很多，如基于信息增益、互信息、信息熵等多种方法[3,4]，根据有关研究表明，使用互信息来提取特征词，不需要像使用其他方法那样首先去除高频词，原因如下：

(1)高频词的概念没有明确定义，不好界定，人为设定可能会漏掉对分类有意的词。

(2)对那些确实对分类意义不大的高频词，通常互信息都比较低，会被自动滤除。

“互信息”的计算公式如下：

$$\mathrm{I}(W,C)=\log\frac{\mathrm{P}(W\wedge C)}{\mathrm{P}(W)\times\mathrm{P}(C)}$$

其中，P(W)为出现词 W 的文档数与文档总数的比值；P(C)为类别 C 文档数与文档总数的比值；P($W\wedge C$)为词 W 与类别 C 同时出现的概率。

经特征词提取后，每一个样本可以用一组特征词来表示：

作者简介：吕英杰(1981－)，男，硕士，主研方向：文本挖掘；叶 强，副教授；李一军，教授、博士生导师

收稿日期：2006-08-20 **E-mail：**luyingjie982@163.com

$X_i = ((x_1, w_1), (x_2, w_2), ..., (x_m, w_m))$

其中，w_m 为特征词 x_m 的相对词频。对每一类中的所有网页取其平均样本，即统计出表示该类网页的所有特征词以及该词在该类网页中的平均词频，该类网页可表示为

$C = ((x_1, \overline{w_1}), (x_2, \overline{w_2}), ..., (x_k, \overline{w_k}))$

其中，$\overline{w_j} = \frac{\sum_{i=1}^{n} w_{ij}}{n}, j \in k$；$k$ 表示该类中特征词的数目；n 为该类所有的网页数目，w_{ij} 表示在网页 i 中特征词 j 的平均词频。

1.2 构造模糊评判矩阵

模糊评判矩阵的构造是以特征词为基础的。特征词对网页主题的反映，不仅体现在词频上，还与特征词的位置、词性等多种因素有关，而各种因素的影响程度也各不相同。因此，可以构造 2 级模糊综合评判矩阵，将与特征词相关的各因素进行等级划分，这样既能突出主要因素的决定作用，又避免了次要因素因其权重过小而被“淹没”在众多的评价因素值中，使分类算法更好地反映网页的真实信息。步骤如下:

(1)构造因素集

1)1 级因素集: 根据网页中的位置不同进行划分，将因素集划分为 u_1(标题)、u_2(正文第 1 段)、u_3(其它段落)等 3 个因素，即 $U = (u_1, u_2, u_3)$。

2)2 级因素集: 在各个位置中选取一定数目的特征词作为 2 级因素集。即

$u_i = (u_{i1}, u_{i2}, ..., u_{ik})$

其中，$u_i \in (u_1, u_2, u_3)$；u_{ik} 是在位置 u_i 内选取第 k 个特征词作为评价因素。

(2)确定因素集的权重

在 1 级因素集中，各个位置对反映文章主旨的影响程度由大到小依次为

标题>正文第一段>其他段落

权重分别定为 A1, A2, A3。即 A=(A1, A2, A3)。数值由专家根据实际情况评定。

在 2 级因素集中，可以根据以下两方面因素确定各个特征词的权重:

1)特征词的词性。相关研究表明，名词、动词等实词对文章主旨的反映程度要大于形容词和副词等词性的特征词，因此，可将各词性的权重分别定为名词(b1)、动词(b2)、形容词(b3)、副词(b4)、介词(b5)、连词(b6)、助词(b7)，具体数值由专家根据实际情况评定。

2)在特征词上所标注的 html 标签。像一些<h1></h1>、、等标签所标注的特征词，在网页文本中起到强调的作用，自然对网页主题的反映程度比较大，这些因素必须在特征词的权重计算中加以考虑，具体权重分配策略就不再详述了。

综合以上 2 个因素，可求出每个特征词的权重。再对每个位置上的所有特征词的权重进行归一化处理，确定出 2 级因素集的权重，即

$Ai = (a_{i1}, a_{i2}, ..., a_{ik})$

其中，$Ai \in (A_1, A_2, A_3)$；a_{ik} 是特征词 u_{ik} 的权重。

(3)构造评判集

在本算法中，评判集就是要进行网页分类所划分的各个类别，表示为

$V = (V_1, V_2, ..., V_n)$

如果要将新闻网页分为体育、科技、财经 3 类，则评判集为 V=(体育,科技,财经)。

(4)构造模糊评判矩阵

选取一个从 U 到 V 的模糊映射函数，构造因素集中的各个特征词对每一种网页类别的隶属函数。由于本文是通过比较待归类网页特征词的绝对词频与样本网页中该特征词的平均词频之间的关系来反映该网页隶属于各网页类别的模糊程度，因此选用升半梯形的模糊函数来确定一个模糊评判矩阵 $\boldsymbol{R}$，即

$$\boldsymbol{R} = (r_{ij})_{n\times m} = \begin{pmatrix} r_{11} & r_{12} \cdots & r_{1m} \\ \vdots & & \vdots \\ r_{n1} & a_{n2} \cdots & a_{nm} \end{pmatrix}$$

其中

$$r_{ij} = \begin{cases} 0 & x_i < a \\ \dfrac{x_i - a}{w_{ij} - a} & a \le x_i < w_{ij} \\ 1 & x_i \ge w_{ij} \end{cases}$$

其中，x_i 为特征词 i 在待归类网页中的绝对频数，W_{ij} 为特征词 i 在第 j 类网页中的平均词频。经过以上步骤，就可以构造出对待分类网页的模糊评判表。

(5)综合评判

首先进行一级综合评判，根据 $\boldsymbol{B}=A \cdot \boldsymbol{R}$ 可求出

$$\boldsymbol{B} = A_1 \bullet \boldsymbol{R}' = (a_{11}, a_{12}, ..., a_{1n}) \bullet \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ r_{31} & r_{32} & \cdots & r_{3m} \end{pmatrix} = (r_{11}', r_{12}', ..., r_{1m}')$$

同理可求出:

$B_2 = (r_{21}', r_{22}' \ldots r_{2m}')$　$B_3 = (r_{31}', r_{32}' \ldots r_{3m}')$

据此可构造二级模糊评判矩阵，即

$$\boldsymbol{R}' = \begin{pmatrix} B_1 \\ B_2 \\ B_3 \end{pmatrix} \begin{pmatrix} r_{11}' & r_{12}' & \cdots & r_{1m}' \\ r_{21}' & r_{22}' & \cdots & r_{2m}' \\ r_{31}' & r_{32}' & \cdots & r_{3m}' \end{pmatrix}$$

$$\boldsymbol{B} = A \bullet \boldsymbol{R}' = (A_1, A_2, A_3) \bullet \begin{pmatrix} r_{11}' & r_{12}' & \cdots & r_{1m}' \\ r_{21}' & r_{22}' & \cdots & r_{2m}' \\ r_{31}' & r_{32}' & \cdots & r_{3m}' \end{pmatrix} = (b_1, b_2, ..., b_m)$$

可最终得出待分类网页分别隶属于各个类别的程度为 $b_1, b_2, b_3, \ldots, b_m$

(6)结果处理

一般情况下根据“最大隶属度原则”，把待分类网页归入到隶属程度最大的类别中，目标完成、算法结束。如果该网页在几个类别中的隶属度都比较大且程度接近，表明该网页属于多个类别的交叉类，可以同时归入这几个类中，也可以增加评判矩阵中因素集(即特征词)的个数，进一步进行判断直至能明显地区分出该网页对各类别的隶属度差异并进行归类。如果该网页在所有类别中的隶属度均比较小，则该网页有可能不属于任何已有类别，可以提取出来另行研究。

2 性能测试及其评价

2.1 实验语料及测试方法

本实验选取的网页语料来自于新华社的新华网，选取了 300 篇新闻网页作为训练语料，进行手工分类分成财经、科技、军事、体育等 6 大类，每个类别中均包含 50 个网页。再选取 130 个网页作为测试网页，每个类别各有 20 个网页，剩余 10 个网页为兼类网页，即可归入两个以上类别的网页。

训练方法是首先对每个网页进行分词处理，本试验采用的分词工具是中科院的 ICTCLAS 分词软件。进行分词后，为简化处理剔除一元词，再将每类中的所有网页的相同词进行合并，然后计算各个词语的信息量。经过以上处理后，在

每个类别中提取 300 左右的词作为该类的特征词，训练结束。

测试时，将测试网页进行分词后，再与训练词库中的词进行匹配，把匹配成功的词语抽取出来作为该网页的特征词，然后运用前面介绍的模糊综合评判算法进行计算。

2.2 测试结果及评价方法

本实验的归类处理采用最大隶属度原则和阈值法相结合的方法，首先保证测试网页对各个类别的隶属度有超过最低阈值的，否则视为不属于任何类别的另类网页。满足最低阈值后，把测试网页归入隶属度最大的类别中。对可分别归属于两类以上的兼类网页，测试方法是在上述处理方法的基础上，进一步判断最大的隶属度和其他隶属度之间的差值，如果差值小于设定常数，则说明该网页隶属于某几个类别的程度相差不大，把该网页归入相应的多个类中。准确率和召回率是传统的信息检索领域中常用的评价指标，准确率表征的是分类的正确性，召回率表征的是分类的完整性。笔者主要采用这两项指标对分类器性能进行评价。单类网页测试结果及评价指标值如表 1 所示。对于分类错误的网页分析见表 2。10 个兼类网页的测试结果见表 3。

表 1　单类网页测试结果及评价指标值

	财经	科技	体育	教育	军事	娱乐	总计
应有网页数	20	20	20	20	20	20	120
实际网页数	23	22	17	19	19	20	120
正确网页数	17	16	16	18	18	17	102
召回率(%)	85	80	80	95	95	85	85
准确率(%)	73.9	72.7	94.1	94.7	94.7	85	85

表 2　分类错误的网页分析

	财经	科技	体育	教育	军事	娱乐	总计
应有网页	20	20	20	20	20	20	120
错误网页	3	4	4	2	2	3	18
应归于第 2 隶属度的网页数	2	4	3	2	2	2	15

表 3　兼类网页的测试结果

网页数	财经	科技	体育	教育	军事	娱乐	总计
应有	7	5	2	4	0	2	20
实际	5	6	0	4	0	2	17
正确	3	5	0	3	0	2	13

结果表明，有 6 个网页成功地归入了所属的 2 个类别中，有 3 个网页只归入 1 个类别，1 个网页归入了错误的类别。

2.3 结果分析

(1)在隶属于单个类别的网页分类结果中，总体召回率和准确率均达到了 85.0%，分类效率是比较高的。对分类错误的网页进行分析可以得知，83.3%的分类错误网页应归于其计算得出的第二大隶属度的类别中，因此，可以在今后的研究中关注网页的第 2 隶属类别，以期进一步提高分类正确率。

(2)从兼类网页的分类结果看，10 篇兼类网页其中有 6 篇成功地归入到所属的 2 个类别中，由于兼类问题处理复杂，能取得这样的效率，因此模糊综合评判算法能比较好地处理网页兼类的问题，该技术可以根据用户需要进行灵活处理。

(3)从各个类别来看，无论从单类测试还是在兼类测试，训练词库比较大的类别准确率相对较低。主要是由于大类训练文本篇幅较长，训练后的词库容量较大，使得待分类网页在大类中的隶属程度要普遍大于在小类中的隶属程度，而使一些原属于小类的网页误归入大类中，造成大类中含有较多的“噪声”文本而降低了准确率，这可以通过增大训练样本数量对各个类别的词库容量进行平衡来解决。

3 结论

网页信息的自动分类是互联网信息处理领域中的一项重要研究课题。传统的基于统计的分类方法只能机械地把网页分入确定的某一类别中，不利于用户对网页所要表达的信息进行充分完整的了解，尤其是当网页含有多个主题时，确定的单一分类方式更凸显其劣势。本文为了更完整地反映网页信息，引入了模糊数学的相关理论对网页分类进行研究，通过模糊结果使用户对网页信息有更完整的了解。此外，还对网页分类结果进行灵活的处理，通过最大隶属度原则和设定阈值方式，可以有效地过滤出异类网页和解决兼类网页的合理归类问题，使归类结果科学合理。在实际运用中，网页的类型和主题通常复杂多样，可以针对实际情况和用户目的，进行灵活的处理，这也正体现了本文的研究意义，但还存在一些不足，主要表现在对各因素权重的确定和特征词隶属函数的确定方面，这也是影响分类性能的关键，而这方面目前却没有比较权威的理论支持，只能在反复的试验中总结规律，这也是今后要进一步研究的地方。

参考文献

1 Yang Yiming, Liu Xin. A Reexamination of Text Categorization Methods[C]//Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval. 1999: 42-49.

2 李雪蕾, 张冬荣. 一种基于向量空间模型的文本分类方法[J]. 计算机工程[J], 2003, 29(10): 90-92.

3 Yang Y, Pederen J P. A Comparative Study on Feature Selection in Text Categorization[C]//Proceedings of the 14th International Conference on Machine Learning, Nashville Tennessee, USA. 1997: 412-420.

4 单松巍, 冯是聪, 李晓明. 几种典型特征选取方法在中文网页分类上的效果比较[J]. 计算机工程与应用, 2003, 39(22): 146-148.

（上接第 159 页）

及环 Z_n 上产生安全圆锥曲线的方法，并通过扩展文献[1]中定义的圆锥曲线和加法运算，提出了适用于特征为 2 的有限域上安全圆锥曲线的方程。

参考文献

1 张明志. 用圆锥曲线分解整数[J]. 四川大学学报(自然科学版), 1996, 33(4): 356-359.

2 曹珍富. 基于有限域 F_p 上圆锥曲线的公钥密码系统[C]//第五届中国密码学学术会议. 北京: 科学出版社, 1998: 45-49.

3 曹珍富. RSA 与改进的 RSA 的圆锥曲线模拟[J]. 黑龙江大学自然科学学报, 1999, 16(4): 15-18.

4 孙　琦, 朱文余, 王　标. 环 Z_n 上圆锥曲线和公钥密码协议[J]. 四川大学学报(自然科学版), 2005, 42(3): 471-478.

5 DAI Zongduo, YE Dingfeng, PEI Dingyi, et al. Cryptanalysis of ElGamal Type Encryption Schemes Based on Conic Curves[J]. Electronics Letters, 2001, 37(7).

6 Schneier B. Applied Cryptography: Protocals, Algorithms, and Sourcecode in C[M]. 2nd ed. John Wiley & Sons Inc, 1996.