# 基于粗糙集理论的文本分类算法研究

林 珣[1,2] 李志蜀[2] 周 勇[3]
(西南财经大学经济信息工程学院 成都 610071)[1] (四川大学计算机学院 成都 610064)[2]
(华兴职业技术学院 成都 610071)[3]

**摘 要** 文本分类是中文信息处理的重要研究领域。给文本分配一个或多个不同的类别,可提高文本检索和存储的处理效率。粗糙集是一种不需要任何先验信息的分类方法,通过对文本分词、过滤掉停用词之后把剩余的词语作为特征项,然后把文本用向量空间模型表示出来,将文本集转化成不带决策属性的信息系统,用粗糙集理论中核心内容属性约简实现对文本的分类。实验表明,该方法的查准率和查全率都有所提高。

**关键词** 文本分类,粗糙集,约简
**中图法分类号** TP391 **文献标识码** A

## Text Classification Algorithm Study Based on Rough Set Theory

LIN Xun[1,2] Li Zhi-shu[2] ZHOU Yong[3]
(School of Economic Information Engineering, Southwestern University of Finance and Economics(SWUFE), Chengdu 610074, China)[1]
(School of Computer, Sichuan University(SCU), Chengdu 610064, China)[2]
(Huaxing Vocational and Technical College, Chengdu 610071, China)[3]

**Abstract** Text dataset is transformed to information system without attribute of decision making and the core content of attribute reduction has been applied to text classification. Experiment shows that the precision rate and recall rate are enhanced in this method; furthermore, it does not require any a priori information.

**Keywords** Text classification, Rough set, Reduction

文本分类(Text Categorization)是中文信息处理重要的研究领域,其目标是在分析文本内容的基础上,给文本分配一个或多个比较合适的类别,从而提高文本检索、存储等应用的处理效率。在常用的文本分类算法中,如支持向量机方法、K近邻方法、朴素贝叶斯方法、决策树方法,每个文本都用维数特别高的向量来描述,其向量维数通常高达上万维,即使处理能力最强的计算机也难以处理[1]。

很多学者运用粗糙集理论中的约简方法,约去不重要的信息,生成文本的分类规则。不过,文献[2-5]在应用粗糙集理论进行文本分类时,利用人工把文本集划分成不同的类,并且把每个文本归于不同的类,进而形成决策属性,把文本集形成一个决策信息系统,再利用粗糙集中的约简方法。文献[6,7]利用模糊聚类的方法把文本集聚类,把文本归属于类的结果作为决策属性,生成决策信息系统,然后再利用粗糙集对决策表进行约简。这两种方法都存在着人为地把信息系统转化成决策系统的缺点,本文把文本集直接转化成信息系统,运用粗糙集理论对信息系统直接约简,找到文本集特征词约简后的特征词核集合,进而得到文本集的分类器。

## 1 基本概念

**定义 1**[9] 称四元组 $S=\langle U,A,V,f\rangle$ 为信息系统,其中 $U=\{u_1,u_2,\cdots,u_n\}$ 是具有 $n$ 个元素的非空集,称为对象空间,$U$ 中的元素称为对象,$A=\{a_1,a_2,\cdots,a_m\}$ 也是一个非空有限集,$A$ 中的元素 $a$ 称为属性,$V=\bigcup_{a\in A}V_a$,$V_a$ 是属性 $a$ 的值域,$f:U\times A\to V$ 是一个信息函数。

从信息系统的定义可以得到:

(1)信息函数为每个对象的每个属性赋予一个信息值,即 $\forall a\in A,x\in U$,有 $f(x,a)\in V_a$;

(2)如果在对象集中没有重复元素,那么一个信息系统就是一个关系数据库。

**定义 2** 在信息系统 $S=\langle U,A,V,f\rangle$ 中,任意属性子集 $p\subseteq A$,称二元关系 $ind(P)=\{(x,y)\in U\times U\mid \forall a\in P, f(x,a)=f(y,a)\}$ 为属性子集 $P$ 的不可分辨关系。如果 $(x,y)\in ind(P)$,那么称 $X$ 和 $Y$ 是 $P$ 不可分辨的。

很显然,对于 $\forall p\subseteq A$,不可分辨关系 $ind(P)$ 是等价关系。符号 $U/ind(P)$(简记为 $U/P$)表示不可分辨关系 $ind(P)$ 在 $U$ 上导出的划分。

**定义 3** 在信息系统 $S=\langle U,A,V,f\rangle$ 中,属性子集 $A_1\subseteq A$,$A_2=A_1\cup\{r\}$,它们导出的等价类分别为:$U/A_1=\{X_1,X_2,\cdots,X_n\}$,$U/A_2=\{Y_1,Y_2,\cdots,Y_m\}$,如果对于任意 $X_i\in U/A_1$ 都存在 $Y_0\in U/A_2$,使得 $X_i\subseteq Y_0$,则称属性 $r$ 为 $A_2$ 中 $A_1$ 不必要的,或者称 $A_1$ 是 $A_2$ 的一个约简。

到稿日期:2010-12-13 返修日期:2011-04-25 本文受国家自然科学基金(60803106)资助。

林 珣(1973-),女,博士生,讲师,主要研究方向为数据挖掘和商务智能,E-mail:linx_t@swufe.edu.cn;李志蜀(1947-),男,教授,博士生导师,主要研究方向为计算机网络、智能控制等;周 勇(1970-),男,副教授,主要研究方向为数据挖掘、粗糙集等。

当然，$A_2$ 的约简并不只有 $A_1$ 一个约简，很有可能有多个约简。

**定义 4** $A_2$ 中所有必要关系组成的集合称 $A_2$ 的核，记作 $core(A_2)$。

## 2 文本集的信息系统描述

### 2.1 文本的向量表示

用简单而准确的方法把文本表示成计算机能够处理的形式是进行文本分类的基础。目前，在信息处理中，文本有向量空间模型、语义网络、框架模型等表示方法，其中向量空间模型得到了广泛的应用。向量的特征项可以选择字、词、词组或概念。根据实验结果，普遍认为选取词作为特征项要优于字和词组；概念特征虽然能更好地表示文本，但是却相对复杂。在本文中，对文本分词、过滤掉停用词之后把剩余的词语作为特征项，然后把文本用向量空间模型表示出来。形式如下：$(w_1, w_2, \cdots, w_n)$，$w_i$ 表示第 $i$ 个特征词语的权重。

### 2.2 文本向量的确定

为了能够让粗糙集进行预处理，必须把文本集中的所有特征词组成一个向量，并确定每个文本所对应向量的每个分量的值，即权重。自从提出向量空间模型以来，出现了很多权重函数，如布尔权重函数、TF-IDF 权重函数、ITC 权重函数、OKAPI 权重函数等等。实验表明，OKAPI 权重计算公式是所表示的文本向量空间中最为合理的权重函数之一。但是 OKAPI 权重计算公式在刻画词语反映文本主题的程度时只考虑了词频因素，而事实上，位置在反映词语强调文本主题时也很重要。研究表明，一篇文章中出现在标题、摘要、正文中的词语对主题的表现能力依次减小，因而用修正的 OKAPI 权重计算公式：

$$a_{ik}=\frac{tf.tl}{0.5+\frac{1.5dl}{avg-dl}}\log\left(\frac{N-df+cdf+0.5}{df\_cdf+0.5}\right)$$

式中，$a_{ik}$ 表示第 $i$ 个特征词在第 $k$ 个文本中的权重，$tf$ 表示第 $i$ 个特征词在训练第 $k$ 个文本中出现的次数，$df$ 表示在训练集中出现第 $i$ 个特征词的文本数，$dl$ 表示第 $k$ 个文本的长度，$df\_cdf$ 表示所有文本长度的平均值，$N$ 表示训练集中的文本数。

### 2.3 文本集信息系统的生成

把训练集中每个文本都表示成一个向量，并且把所有的文本表示成一个信息表。信息表的行就是每个文本向量，信息系统的列是不同文本的特征词的权重，如表 1 所列。

表 1 训练文本集的信息系统

| | $word_1$ | W | $word_{n-1}$ | $word_n$ |
|---|---|---|---|---|
| Tx1 | $w_{11}$ | …… | $w_{1(n-1)}$ | $w_{1n}$ |
| Tx2 | $w_{21}$ | …… | $w_{2(n-1)}$ | $w_{2n}$ |
| …… | …… | …… | …… | …… |
| $Tx_m$ | $w_{m1}$ | …… | $w_{m(n-1)}$ | $w_{mn}$ |

### 2.4 属性值的离散化

粗糙集理论与信息系统相结合的离散化算法中，根据离散化处理时是否考虑到信息系统的具体属性值，可把离散化算法分为“非参照性的离散化算法”和“参照性的离散化算法”。“非参照性的离散化算法”在离散化过程中很少考虑或不考虑信息系统中具体的属性值，而“参照性的离散化算法”是参照信息系统中具体的属性值来进行的。根据离散化过程是否改变信息系统原有的不可分辨关系，可以把离散化算法分为“改变不可分辨关系的离散化算法”和“不改变不可分辨关系的离散化算法”。

由于本文的信息系统只有条件属性没有决策属性，在离散每个属性时，并不考虑其他属性对该属性的影响，同时也不考虑离散化后是否会改变不可分辨关系，因而就采用等距离离散法。文本集信息系统各个特征词的属性值的取值范围为 (0,1)，分别以 0.1,0.2,…,0.9 为断点，把所有属性值都进行离散，并赋予不同的标志符号。

## 3 分类器形成算法描述

### 3.1 文本集信息系统属性约简

在属性值离散化后的文本集信息系统中，根据属性对文本对象进行划分形成等价类，根据定义 3，从信息系统中约掉一些不必要的特征词属性，从而生成文本信息系统特征词属性核集合。

利用逐一删除法生成特征词属性核集合。设文本集信息系统的特征词属性集为 $A=\{word_1, word_2, \cdots, word_n\}$，依次检验 $A$ 中每个特征词 $word_i$ 的必要性，如果 $word_i$ 是不必要的，则形成 $A$ 的子集 $A_{1i}=\{word_1, word_2, \cdots word_{i-1}, word_{i+1}, \cdots, word_n\}$，$(i=1,2,\cdots,n)$，并令 $A_1=\bigcap_{i=1}^{K} A_{1i}$，其中 $K$ 表示可删除特征词属性的个数。以同样的方法，依次检验 $A_1$ 每个特征词属性的必要性，直到每个特征词属性都是必要的为止。此时就得到 $A$ 的特征词属性核集合 $A_0=\{word_1, word_2, \cdots, word_N\}$，其中 $N$ 表示特征词核属性的个数。

### 3.2 分类器的形成

设文本集信息系统的特征词核集合为 $A_0=\{word_1, word_2, \cdots, word_N\}$，不可分辨关系 $ind(A_0)$ 在 $U$ 上导出的划分为：$U/A_0=\{X_1, X_2, \cdots, X_m\}$，这个划分 $U/A_0$ 就是基于粗糙集理论的文本分类器，每个类的特点分别为：

$\rho(X_1)=f(A_0, X_1)$，$\rho(X_2)=f(A_0, X_2)$，……，$\rho(X_m)=f(A_0, X_m)$

## 4 仿真实验

### 4.1 分类实验

从新浪网站(http://www.sina.com/)下载新闻报道 350 篇作为试验语料，用中科院的“分词系统”把 350 篇文章进行分词，并用该系统去掉停用词；用剩下的 4000 个特征词构成向量，针对每个文本生成特征向量，用 350 个文本向量构成信息系统，对文本集信息系统(共有 350 行，4000 列)的属性离散化，生成运用粗糙集理论能够处理的文本集信息系统。

通过粗糙集理论的属性约简，从 4000 个特征词属性集中搜索到含有 304 个特征词的特征词核属性集合，最后应用特征词核属性集合对约简后的文本集信息系统进行分类，把具有 350 个文本的文本集分类，最后分成 3 类。

### 4.2 实验结果分析

因为文本分类就是映射过程，所以评估文本分类系统的标志是映射的准确程度和映射的速度。映射的速度取决于映射规则的复杂程度，而评估映射准确程度的参照物是通过专

(下转第 263 页)

科学,2010,37(9):270-271,293

[5] Pluim J P W, Maintz J B A, Viergever M A. Mutual information based registration of medical images: a survey[J]. IEEE Transactions on Medical Imaging, 2003, 22(8): 986-1004

[6] Maes F, Vandermeulen D, Suetens P. Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information[J]. Medical Image Analysis, 1999, 3(4): 373-386

[7] Stone H, Le M J, McGuire M. The translation sensitivity of wavelet-based registration[J]. IEEE Trans. Pattern Anal. Machine Intell, 1999, 21(10): 1074-1080

[8] Chalfant J S, Patrikalakis N M. Shape registration via the wavelet transform[C]// IEEE International Conference on Shape Modling and Applications. New York, USA, June 2008: 277-278

[9] Xu P, Yao D. A study on medical image registration by mutual information with pyramid data structure[J]. Computers in Biology and Medicine, 2007, 37(3): 320-327

[10] 刘丽,苏敏. 基于小波变换和互信息的医学图像配准[J]. 中国图象图形学报,2008,13(6):1171-1176

[11] Kaneko S, Satoh Y. Using selective correlation coefficient for robust image registration[J]. Patt Recog, 2003, 36(5): 1165-1173

[12] Pluim J P W, Maintz J B A, Viergever M A. Image Registration by Maximization of Combined Mutual Information and Gradient Information[J]. IEEE Transactions on Medical Imaging, 2000, 19(8): 809-814

[13] Wang An-na, Pan Bo, Ma Ji-dong, et al. Deformable image registration base on Gabor wavelet and SVM[C]// IEEE fourth International Conference on Natural Computation. Jinan, China, August 2008: 145-149

[14] 余成波. 数字图像处理及 matlab 实现[M]. 重庆:重庆大学出版社,2003

(上接第 240 页)

家思考判断后对文本的分类结果与人工分类结果越接近,分类的准确程度就越高。分类结果的准确性的度量通常使用查全率和查准率来判定,查全率和查准率反映了分类质量的两个不同方面,两者必须综合考虑[8,10,11]。

查准率是所有判断的文本中与人工分类结果吻合的文本所占的比率,其数学公式可表示如下:

$$precision=\frac{a}{a+b}$$

查全率是人工分类结果应有的文本中分类系统吻合的文本所占的比率,其数学公式表示如下:

$$recall=\frac{a}{a+c}$$

式中,$a$ 为被正确分到该类的文档数;$b$ 为被错误分到该类的文档数;$c$ 为本应属于该类,但没有分到该类的文档数。

表 2 列出基于粗糙集理论的文本分类统计信息。

表 2 基于粗糙集理论的文本分类统计信息

| 类别 | 人工分类 | 自动分类 | 正确分类 | 查全率 | 查准率 |
|---|---|---|---|---|---|
| 财经 | 150 | 160 | 130 | 0.87 | 0.81 |
| 科技 | 80 | 68 | 58 | 0.73 | 0.85 |
| 体育 | 60 | 52 | 48 | 0.80 | 0.92 |
| 娱乐 | 60 | 58 | 55 | 0.92 | 0.94 |

从本次实验的测试结果可以看出,利用本文提出的基于粗糙集理论的文本分类模型对文本进行分类的结果与专家通过人工分类的结果相比,其查全率和查准率都有所提高。

**结束语** 本文利用粗糙集理论中的属性约简技术,把文本集的信息系统中不重要的特征词属性约简掉,找出特征词属性集中的核集合,再利用核集合把文本集进行分类,从而在不利用任何先验信息的基础上对文本集进行分类。通过实验分析得到;该种方法能够使查准率和查全率都有所提高。

## 参 考 文 献

[1] 卢娇丽,郑家恒. 基于粗糙集的文本分类方法研究[J]. 中文信息学报,2005,19(2):66-70

[2] 侯凡,周明全,耿国华,等. 基于粗糙集的文本分类方法在网络科技资源应用集成环境中的应用[J]. 计算机应用与软件,2009,26(3):88-91

[3] 王汉萍,孟庆春,张继军,等. 基于粗糙集的文本自动分类方法的研究[J]. 信息技术,2003,27(8):46-48

[4] Shi Lei, Ma Xin-ming, Xi Lei, et al. Rough set and ensemble learning based semi-supervised algorithm for text classification [J]. Original Research Article Expert Systems with Applications, 2011, 38(5): 6300-6306

[5] 张志飞,苗夺谦. 基于粗糙集的文本分类特征选择算法[J]. 智能系统学报,2009,4(5)

[6] 郑丽英,王海涌,刘丽艳. 基于粗糙集和模糊集理论的文本分类系统的研究与实现[J]. 铁道学报,2007,29(1):45-49

[7] Li Wen, Miao Duo-qian, Wang Wei-li. Two-level hierarchical combination method for text classification[J]. Original Research Article Expert Systems with Applications, 2011, 38(3): 2030-2039

[8] Chen Hui-ling, Yang Bo, Liu Jie, et al. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis[J]. Original Research Article Expert Systems with Applications, 2011, 38(7): 9014-9022

[9] 张文修,吴伟志,梁吉业,等. 粗糙集理论与方法[M]. 北京:科学出版社,2005:18-19

[10] 侯丽娟,王国胤,聂能,等. 粗糙集理论中的离散化问题[J]. 计算机科学,2000,27(12):89-94

[11] 王珍珍. 粗糙集理论的文本分类算法研究及应用[D]. 济南:山东师范大学,2007:38-40

[12] 何薇,徐伟华. 信息检索的粗糙集方法[J]. 重庆理工大学学报:自然科学版,2010,24(9):84-88