基于 Rough 集约简算法的中文文本自动分类系统

盛晓炜 江铭虎
(清华大学中文系计算语言学实验室 北京 100084)
(中国科学院自动化研究所模式识别国家重点实验室 北京 100080)

摘 要： 现有的文本自动分类离不开文档向量的构造，向量的分量与文档中的特征项相对应。这种向量通常高达几千维甚至数万维，计算量相当大，因此需要对向量进行约简。而传统的基于频率的阈值过滤法往往会导致有效信息的丢失，影响分类的准确度。该文将 Rough 集理论引入自动分类，并提出了一种新的文档向量约简算法。实验证明该算法不仅能有效缩减文档向量的规模，而且相比传统的阈值法信息损失小、准确率更高。

关键词： 自动分类，Rough 集，决策表，约简算法
中图分类号： TP391 **文献标识码：** A **文章编号：** 1009-5896(2005)07-1047-06

Automatic Classification of Chinese Documents Based on Rough Set and Improved Quick-Reduce Algorithm

Sheng Xiao-wei Jiang Ming-hu
(*Lab of Computational Linguistics, Dept of Chinese Language, Tsinghua University, Beijing* 100084, *China*)
(*State Key Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract Much of the previous automatic Text Classification (TC) methods are closely connected with the construction of document vectors. With each term corresponding to a unit in the vector, this method maps the document vectors into a very high dimensional space, possibly of tens of thousands of dimension, which results in a massive amount of calculation. Since the traditional algorithms based on frequency and threshold filtering may often lead to the loss of effective information, this paper presents a new system for TC, which introduces rough set theory that can greatly reduce the document vector dimensions by reduction algorithm. The empirical results prove to be very successful, for it can not only effectively reduce the dimensional space, but also reach higher accuracy while losing less information compared with usual reduction methods.

Key words Automatic classification, Rough set, Decision table, Reduction algorithm

1 引言

随着网络的迅猛发展，电子文本的信息量也急剧膨胀，为了能更好地管理和检索这些信息，对文本预先进行分类成为必不可少的一环。目前的文本分类系统都依赖于各种向量模型[1,2]，在训练阶段要为每篇文档构造一个文档向量，然后通过向量聚类得到分类向量。在对新文档进行分类时，将其文档向量和各分类向量进行余弦夹角等计算，最终确定该文档所属类别。

文档向量通常的构造方法为：计算文档中的特征项的权值(通常由特征项频率 TF，反文档频率 IDF 等频率信息得到)[2]，然后将这些权值填入一个由全部特征项构成的向量中，未出现的特征项对应的分量为 0。由于文档中可能出现的特征项数量很多，因此这种文档向量通常高达几千甚至数万维，带来了很大的运算量。导致了现有的分类系统在效率上难以适应 Internet 上信息量迅猛膨胀的要求。

为了降低文档向量的维数，很多系统在频率统计的基础上使用了阈值过滤的方法[3]，即将文档向量中低于阈值的分量全部去除。这样做虽然能降低向量的维数，却不可避免地丢失了一些有用的信息，特别是对于分类很重要的低频词(比如某些类别中的专有名词，虽然出现频率很低，但区分类别的作用却很大)，最终影响到了检索的精度。

80 年代初诞生的 Rough 集理论[4,5]能够有效地解决文档向量维数过大的问题，目前该理论已经应用到了邮件分类和网页分类中[6,7]，取得了一定的效果。但 Rough 集的数学计算量是较大的，很难应用于大批量的文本分类。本文将

2004-02-19 收到，2004-08-05 改回
教育部优秀青年教师资助计划，教育部归国人员启动基金，模式识别国家重点实验室开放基金和清华大学基础研究基金资助课题

Rough 集理论及决策表的约简[8]引入自动分类，提出了一种适合自动分类的快速约简算法——IQR 算法。实验结果表明该算法能在不丢失有用信息的情况下有效降低向量维数，相比阈值过滤法有更高的准确率；而且该算法避免了传统 Rough 集约简算法中的盲目搜索，有利于快速求得约简结果，使得 Rough 集理论能适用于大批量文本分类的需求。

2 Rough 集理论简介

2.1 Rough 集基本概念

Rough 集(Rough sets)理论是由波兰数学家 Pawlak 于 20 世纪 80 年代初提出的一种分析不确定、不完备知识的集合理论。其研究对象是由多值属性描述的知识系统，形式化定义如下[9]：

定义 1 一个知识系统 S 可以定义为一个 4 元组：$S=<U, A, V, f>$。其中 U 是对象集合，即论域；A 是对象的属性集合；V 为属性的值域；$f:U\times A\to V$ 是一个属性函数，即对 $\forall x\in U, \forall a\in A$ 均有 $f(x,a)\in V$ 为对象 x 的 a 属性的值。

如果将对象作为行，属性作为列，就构成了一个表格，表格中的值就是对象的属性值。通常的知识都可以用这种表格的形式来表示，我们称之为知识表。

一般来说，知识表中包含的信息是不足以明确区分每个对象的(也就是对象不可辨)，特别是在属性值不完备的情况下。而属性值不完备是一种普遍的现象，因为我们往往无法精确地、毫无遗漏地确定对象的每一个属性，这造成了知识的不确定与不完备。这种不可辨的对象相对于其属性则构成了一个 Rough 集。而对象的不可辨关系则是 Rough 集上最重要的关系之一。

定义 2 在 S 中，对于任意属性子集 $P\subseteq A$，均可定义一个不可辨关系 $\mathrm{IND}(P)$：

$$\mathrm{IND}(P)=\{(x,y):x,y\in U,\ \forall p\in P\ \ f(x,p)=f(y,p)\} \tag{1}$$

在不产生混淆的情况下可以用 P 代替 $\mathrm{IND}(P)$。

显然 $\mathrm{IND}(P)$ 是集合 U 上的一个等价关系，因此集合 U 可以根据等价关系 $\mathrm{IND}(P)$ 进行等价类的划分。

定义 3 集合 U 对于等价关系 $\mathrm{IND}(P)$ 的划分记为 $U/\mathrm{IND}(P)$，其计算公式如下：

$$U/\mathrm{IND}(P)=\otimes\{U/\mathrm{IND}(p)\mid p\in P\} \tag{2}$$

其中运算符 $\otimes$ 定义为：$A\otimes B=\{X\cap Y\mid \forall X\in A, \forall Y\in B, X\cap Y\neq\phi\}$。$U/\mathrm{IND}(P)$ 也可简记为 U/P。

为了能进行精确的数学计算，Rough 集中引入了一些传统集合来对自身进行描述，这些集合包括下近似集和正域等。

定义 4 对于样例 $E\subseteq U$，E 的下近似集 $\underline{PE}$ 和正域 $\mathrm{POS}_P(E)$ 分别定义为

$$\underline{PE}=\bigcup\{X\mid X\in U/\mathrm{IND}(P), X\subseteq E\} \tag{3}$$

$$\mathrm{POS}_P(E)=\underline{P}(E) \tag{4}$$

2.2 决策表

将 Rough 集理论运用到知识获取中，往往需要先进行实例学习。为此 Rough 集理论在知识表的基础上引入了决策表的概念。通常决策表同知识表一样也包含了大量实例记录，但决策表在知识表的基础上还多一条决策属性。知识获取的目的就是对实例库进行分析，确定哪些属性对决策是有贡献的，哪些属性是不重要或者冗余的。引入决策表后，假定决策属性集为 $Q\subseteq A$，则同样存在划分 U/Q。该划分的正域的计算方法为

$$\mathrm{POS}_P(Q)=\bigcup_{X\in U/Q}\underline{P}X \tag{5}$$

定义 5 属性子集 P 对于决策集 Q 的依赖度(degree of dependency)定义为

$$\delta_P(Q)=\frac{|\mathrm{POS}_P(Q)|}{|U|} \tag{6}$$

而对 $\forall x\in P$，属性 x 对于决策 Q 的重要度(significance)定义为

$$\delta_x(Q)=\delta_P(Q)-\delta_{P-\{x\}}(Q) \tag{7}$$

至此，我们即可根据属性对决策的重要度来对决策表进行约简。

3 基于 Rough 集理论的自动分类

3.1 现有向量约简方法的局限

自动分类中分类向量的规模会直接影响整个分类的效率。前面已经提到这种与特征项相对应的向量的规模往往能达到几千甚至上万维。为了确保系统的效率，必须要对向量进行约简。

为了得到分类向量，先要进行语料训练。而何时对分类向量进行约简则有两种选择，可以在得到训练集中的文档向量后对每个向量进行约简，这样合成得到的分类向量(或矩阵)的维数自然也降低了。但这样做在每篇文档上都有信息损失，积累起来将会严重影响分类的准确性。因此更为常见的做法是先对文档向量进行合成得到分类向量，然后在此基础上进行阈值过滤。而传统的阈值过滤往往存在如下局限：

(1) 无论以什么方式对文档向量进行合成，都难免会丢失一些重要信息。比如对文档向量各分量作加权平均，将得到的结果作为分类向量的分量等方法，很容易使得一些属于该类别，但特征项分布比较稀疏的文档淹没在“主流”中，最终无法在类别向量中有所体现。这会影响分类的准确度。

(2) 未充分考虑到特征项权值在各个类别中的分布，会过滤掉一些对分类很重要的低频词。比如政治类中的专有名词或科技类中的某些术语，这些词虽然出现频率很低，对于

分类却有很大区分度。

(3) 对(非禁用词的)高频词不能起到很好的区分作用。按传统阈值过滤的方法，高频词总是能被保留，因为它们的权值一般都相对较大，但有很多高频词是不具备分类能力的。

3.2 Rough集的优点

使用 Rough 集中的决策表进行向量约简能有效避免上面提到的问题。

首先在 Rough 集中不存在要将文档向量合成为分类向量的问题，因为对决策表的操作是作为一个矩阵整体进行的，这样避免了合成过程中的信息丢失问题。

其次 Rough 集对高频词和低频词都能很好地处理。如果我们将文档作为行，特征项以及类别决策作为列，就能得到一个决策表。在对决策表约简时，如果某个特征项的频率在不同类别中波动相当大，说明此特征项对分类贡献度大；反之，如果某特征项在不同类别中分布均匀，无论其频率是高还是低，对于分类都是没有贡献度的，这些特征项即可约简掉。这里，频率大小将不再重要，起决定因素的是频率分布。因此，使用 Rough 集进行约简，既不会滤掉对分类很重要的低频词，也不会保留没有区分度的高频词。

3.3 约简算法的选择

对决策表求最小约简已经被证明是 NP 难题[10]，在实际中往往采用启发式算法来求得一个较优解。目前已经涌现出了很多约简算法，包括基于可识别矩阵的算法[11,12]，基于最佳原理的 QR 算法(Quick Reduct Algorithm)[8]等。这些算法的复杂度一般为 $O\left(|A|^2|U|^2\right)$[12]，而且在工程应用中都取得了好的实际效果。但这些算法都难以适应自动分类的需求，这是因为自动分类的属性矩阵有如下特点：(1)属性规模相当庞大，能达到数万维。(2)为稀疏矩阵，也就是虽然文档集中的特征项数目很多，但每篇文章中只出现了一小部分。(3)非 0 属性值的分布很不均匀，也就是同一特征项在不同文章中的权值往往相差很大。

基于可识别矩阵的算法都需要求出矩阵的属性核[11]。而自动分类的属性集规模庞大，且属性值分布不均匀，这往往会导致属性核是个空集，因此该类方法都不适用于自动分类。基于最佳原理的 QR 算法在搜索局部最优解时需要对属性集进行遍历，而自动分类的属性集规模庞大且为稀疏矩阵，盲目遍历所花费的代价也是相当大的。

不过最佳原理应用到自动分类上是可行的，因为直观上如果某个特征项的分类能力很强，它与其它特征项联合起来的分类能力往往也比较强。因此我们将以最佳原理为基础，设计一种新的适合自动分类的约简算法。

3.4 适合自动分类的约简算法

自动分类中属性表的约简实际上是很宽松的，因为这种约简既不需要是最小约简，也不需要具有完备性。因此可以牺牲完备性，来换取时间上的优势。我们可以事先确定一些对分类重要的属性，并排除一些对分类基本上没有作用的属性。这可以基于如下假设：

若自动分类中的属性集为 C，则确定一个属性对分类重要的充分(非必要)条件是：(1)该属性在某一类别中总体权值较高。(2)同时该属性在该类别中分布相对均匀。(3)该属性在其它类别中总体权值较低。

其中第(2)条要求属性在某一类别中分布相对均匀是为了避免下述情况的发生。假定测试集中有两个类别，每个类别各有 4 篇文档 $D_1 \sim D_4$，某个属性的权值分布如表 1。

表 1

	D_1	D_2	D_3	D_4
类 α	10	0	0	0
类 β	0	0	0	0

虽然该属性在类 α 中有较高的总体权值，但并不能确定其就是类 α 的分类属性，因为其在类 α 中分布波动很大，我们不能排除类 α 中的文档 D_1 是一个特例的可能性。根据上述 3 个充分条件，我们引入了数学期望和样本方差来衡量属性在类别中的分布和总体权值。并以此为依据对属性集 C 做局部的排序，同时删除对分类无用的属性。

算法 1 属性集 C 的局部排序与筛选算法：

输入　属性集 C (一个矩阵)

输出　局部排序并筛选后的属性集 C'

(1) 将属性集 C 按类别分割为类别矩阵

(2) 对每个类别矩阵，计算每个列向量的数学期望 x 和样本标准差 σ，最终得到两个向量 $\boldsymbol{X}$ 和 $\boldsymbol{P}$

(3) 将不同类别的 $\boldsymbol{X}$ 向量排成一个矩阵 $\boldsymbol{M}$，计算每个列向量的样本标准差，得到向量 $\boldsymbol{T}$

(4) 对 $\boldsymbol{M}$ 中每个列向量求值最大的 $\boldsymbol{X}$，取相应的 σ 值，构成向量 $\boldsymbol{T}'$

(5) 某一属性在 $\boldsymbol{T}$ 和 $\boldsymbol{T}'$ 中的值分别设为 v 和 v'，分为如下情况：

(a) 若 $v \neq 0$，$v' = 0$，将此类属性排在最前，类中以 v 值排序

(b) 若 $v \neq 0$，$v' \neq 0$，将此类属性排在第 1 类之后，类中以 v/v' 值排序

(c) 若 $v = 0$，将此类属性从 C 中剔除(基本上为无分类作用的高频词或出现频率极低的低频词)

(6) 输出局部排序并筛选后的属性集 C'

到第(4)步后，向量 $\boldsymbol{T}$ 能反映各属性对不同类别的权值分布情况，向量 $\boldsymbol{T}'$ 能反映属性在其出现最频繁的类别中的分布情况。根据前面的 3 个充分条件，如果某个属性在 $\boldsymbol{T}$ 中

值越大(也就是在不同类别中波动大)在 T' 中值越小(也就是在某个类别中波动小)，这样的属性就越有可能成为候选。第(5)步则是根据这一点来排序的。

在实际运算中，与 0 的判等会存在计算上的误差，可以使用一个小的阈值 θ，小于该值的数即可认为等于 0。

我们将最佳原理和算法 1 相结合，作为属性选择的启发式算法，由此得到了一种新的约简算法。

IQR 算法 属性集 C 的约简算法(Improved Quick Reduct Algorithm)：

输入 排序并筛选后的属性集 C,决策集 D，$A=C\cup D$

输出 约简后的属性集 R

(1) $R=T=\phi$

(2) 使用筛选算法 1 对 C 进行局部排序和筛选,得到 C'

(3) do

(4) for 按顺序选取 $x\in(C'-R)$

(5) if $\delta_{R\cup\{x\}}(D)>\delta_T(D)$

$T=R\cup\{x\}$ //找出使 δ 最大的 $R\cup\{x\}$

(6) $R=T$

(7) until $\delta_R(D)=\delta_C(D)$

(8) return R

其中在第(4)步的循环中使用了最佳原理,从搜索局部最优解入手，将集合 R 逐步扩充。而第(2)步中算法 1 的筛选能避免盲目的搜索。算法 1 主要是对矩阵的列向量进行数值计算，其时间复杂度与矩阵规模成线性关系，也就是 $O(|A||U|)$，相比约简算法的复杂度 $O(|A|^2|U|^2)$ 来说影响不大，而且使用算法 1 后，$A=C\cup D$ 的规模会有所缩减，这将会使 IQR 算法的整体效率得到提高。这种效率提高的代价是该算法将无法保证约简的完备性。但完备性对自动分类是不重要的，而且后面的实验将说明 IQR 算法对完备性的影响是很小的。

4 实验结果与分析

我们选用 1999 光盘版的《人民日报》作为语料来源，从中选取了政治、经济、体育、计算机、教育、法律 6 大类共 3000 篇文章。训练集为 6×100 篇文档，测试集为 6×200 篇文档。在通过 TF 构造文档向量后，我们还进行了一步离散化工作，也就是将 TF 值映射到 0～9 这 10 个整数值上。这样可以避免文档向量中 TF 值的变化范围过大，也有利于等价类的划分和运算。

4.1 约简度

我们需要考察使用 IQR 算法到底能在多大程度上缩减文档向量的规模，表 2 是使用不同文档数测试的结果。

其中约简度定义为

$$\text{约简度}=\log\left(\frac{\text{约简前的特征项数}}{\text{约简后的特征项数}}\right) \tag{8}$$

由此可见，使用 Rough 集的约简算法进行文档向量的缩减，效果是非常明显的，能缩减大约 3 个数量级。而使用如此少的特征项是否达到好的分类效果呢？文献[13]已经指出，对于分类而言，的确只需要一个小规模的特征项集就可以得到满意的结果了。后面的实验结果也能证实这一点。

表 2 不同文档数的约简度

总文档数	约简前的特征项数	约简后的特征项数	约简度
6×5=30	4688	4	3.07
6×10=60	6311	5	3.10
6×20=120	8692	8	3.04
6×50=300	12985	13	3.00
6×100=600	16583	15	3.07

4.2 约简效果

除了约简度外，还需要考察约简得到的特征项到底能在多大程度上反映文档的分类。以文档数 6×5 为例，约简后得到的 4 个特征项分别为“领导”、“网络”、“经验”、“经营”，其中 $\delta_{\text{领导}}(D)=0.30$，$\delta_{\text{领导+网络}}(D)=0.57$，$\delta_{\text{领导+网络+经验}}(D)=0.80$，$\delta_{\text{领导+网络+经验+经营}}(D)=1$。这 4 个词在 6×5 篇文档中的频率分布(已经经过了 0～9 映射转换，其中[0,3]为低频，[3,6]为中频，[6,9]为高频)如表 3 所示。

从这个分布表格中可以看出：“领导”一词在政治类中高频出现，而在其他类别中都是低频出现，因此出现“领导”这个词的文章我们可以在很大程度上认为是属于政治类的文章。$\delta_{\text{领导}}(D)=0.30$，也就是只使用“领导”这个词，能在 0.3 的程度上再现原来的分类决策。该词是所有特征项中分类能力最强的，但其 TF 排序却并不高，在第 59 位。可见使用 Rough 集约简，能避开频率因素的影响，挑出对分类决策贡献度最大的特征项。

同样“网络”一词在计算机类中高频出现，在经济类中中频出现，也拥有很强的分类能力，其单独分类能力为 0.2，与领导一词的联合分类能力为 0.57。“经验”一词在政治和教育中中频出现单独分类能力为 0.17，与前面两个词的联合分类能力为 0.80。“经营”一词在经济类中高频出现，在计算机、体育、教育类中中频出现，其单独分类能力为 0.1，与前面 3 个词的联合分类能力为 1。

表 3 是文档数较少的情况。表 4 是总文档数为 6×100 篇时的约简结果。

表3　6×5篇文档，约简后的4个特征项的分布

文档编号→		I	II	III	IV	V	均值	频率
经济类	领导	0	0	3	0	0	0.6	低
	网络	3	3	3	3	3	3	中
	经验	0	0	0	0	0	0	低
	经营	7	5	6	7	5	6	高
政治类	领导	9	8	9	9	9	8.8	高
	网络	0	0	3	0	0	0.6	低
	经验	0	0	7	7	7	4.2	中
	经营	0	0	3	3	8	2.8	低
计算机类	领导	0	0	0	0	0	0	低
	网络	8	0	9	9	9	7	高
	经验	0	0	0	0	0	0	低
	经营	3	4	0	0	8	3	中
体育类	领导	0	0	0	0	0	0	低
	网络	0	0	0	0	0	0	低
	经验	4	0	4	0	0	2.4	低
	经营	4	7	4	3	4	4.4	中
教育类	领导	0	0	6	0	0	1.2	低
	网络	0	0	0	0	0	0	低
	经验	3	3	3	3	3	3	中
	经营	3	8	8	3	3	5	中
法律类	领导	4	0	4	0	0	1.6	低
	网络	0	0	8	0	0	1.6	低
	经验	0	0	0	0	0	0	低
	经营	0	0	0	0	0	0	低

表4　6×100篇文档约简结果

特征项	δ值	累加δ值	TF	TF排名
运动员	0.0717	0.0717	104	329
计算机	0.0583	0.13	189	130
比赛	0.0283	0.18667	320	47
学科	0.0283	0.24667	60	602
人大常委会	0.0567	0.30833	244	79
领导干部	0.0417	0.36833	196	123
网络	0.0017	0.435	434	30
教师	0.01	0.5133	169	159
依法	0.01	0.605	352	34
经济	0.0017	0.73833	645	10
银行	0.0033	0.81833	469	23
发展	0.0033	0.88167	1314	3
社会	0.0017	0.935	497	20
经营	0.0433	0.98667	289	57
工作	0.0033	1	1332	2

约简结果显示，像“发展”这样的词频率虽高，但分类能力却不如频率相对较低的“运动员”、“计算机”等词。实际上，排在约简结果集前面的几个词都有明确的类别标志：“运动员”、“比赛”属于体育类；“计算机”、“网络”属于计算机类；“人大常委会”、“领导干部”属于政治类；“学科”、“教师”属于教育类；“经济”、“银行”属于经济类；“依法”属于法律类等。而“发展”所属类别则没有这些词明确，其对分类的重要度也应该低于前面这些词。因此Rough集挑选出的特征项都具有很强的分类能力，这样的约简结果使用一般的阈值法是很难得到的。

4.3 分类的精确度

除了约简度外，还需要使用一定规模的语料来测试约简后特征项的分类能力。我们使用准确率作为衡量分类效果的标准。某一类别分类的精确度定义为

$$\text{精确度}=\frac{\text{正确分到该类别的文档数}}{\text{分到某个类别中的文档数}} \tag{9}$$

将各类别的精确度取平均值则是整个分类的精确度。

我们在训练阶段，分别使用基于TF-IDF综合加权的阈值法和IQR算法来筛选特征项。IQR算法筛选出的特征项的权值用其δ值表示。在测试阶段使用余弦夹角进行文档与类别的相似度比较。

TF-IDF阈值法、算法2在取不同数目特征项时精确度比较如表5和图1所示。

表5　取不同数目特征项时，TF-IDF阈值法与IQR算法分类精确度比较

分类方法	特征项数					
	5	10	20	50	80	100
TF-IDF阈值法	39.4	40.6	48.0	76.1	85.4	89.6
IQR算法	51.8	59.2	75.3	88.2	91.3	92.7

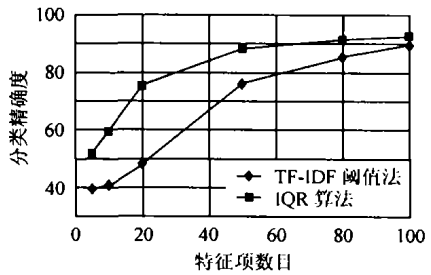

图1　取不同数目特征项时，TF-IDF阈值法与IQR算法分类精确度比较

在IQR算法中，若约简后特征项的数目少于所需数目，则根据δ值大小选取额外的特征项进行补充。从图1中可以看出，在特征项数目少于20时，TF-IDF阈值法的性能很差，这是因为对分类最重要的特征项未必具有频率上的优势，简单的阈值过滤无法保证将这些特征项优先筛选出来。而IQR

算法能保证筛选出对分类最重要的特征项，只取20个特征项，就能达到75%以上的准确度。当特征项数目取到50个以上时，准确率能接近90%。

4.4 IQR算法与QR算法的时间及完备性比较

前面已经提到QR算法的盲目搜索是不适用于自动分类的。表6是我们分别对不同数目的文档使用两种算法进行约简所用时间的统计。

表中IQR算法约简所需时间仅为QR算法的1/4到1/5，所以QR算法的盲目搜索是难以适应大规模预料训练的。使用筛选算法，事先能剔除约50%的特征项，而且最后对约简的完备性基本无影响，可见我们在前面所做的关于分类特征项分布特征的假设是完全成立的。

表6 QR算法与IQR算法时间比较

总文档数		6×5	6×10	6×20	6×50	6×100
总特征项数		4688	6311	8692	12985	16583
经筛选算法剔除后的特征项数		593	1188	1701	2851	3730
耗时(s)	QR	119	586	1314	4262	16237
	IQR	31	125	302	989	3336
约简完备	QR	是	是	是	是	是
	IQR	是	是	是	是	是

5 结束语

自动分类系统的分类效果在很大程度上依赖于文档特征项的选取，本文将Rough集理论中的决策表的约简应用到自动分类中，并提出了适合自动分类特点的约简算法。新的算法充分利用了自动分类中文档特征项分布的规律，相比现有的一些约简算法在有明显的时间优势，更能适应网络上大规模文档分类的要求；而新算法相比传统的阈值法，无论是在特征项的选取和约简上还是在分类准确度上都有更好的表现。我们未来的工作将集中于进一步改进筛选算法的筛选能力和约简算法的运算速度，并将所做的工作应用到WWW网络上的自动网页分类中去。

参 考 文 献

[1] Salton G, Wong A, Yang C S. A vector space model for automatic indexing. *Communications of the ACM*, 1975, 18(11): 613 – 620.

[2] Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys*, 2002, 34(1): 1 – 47.

[3] Riloff E, Lehnert W. Information extraction as a basis for high-precision text classification. *ACM Trans on Information Systems*, 1994, 12(3): 296 – 333.

[4] Zdzislaw Pawlak. Rough sets. *International Journal of Computer and Information Sciences*, 1982, 11(5): 341 – 356.

[5] Zdzislaw Pawlak. Rough sets: Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic Publishers, 1991: 15 – 16, 69 – 80.

[6] Chouchoulas A, Shen Q. A rough set-based approach to text classification. In Proceedings of the 7th International Workshop on Rough Sets, Yamaguchi, Japan, November 1999: 118 – 127.

[7] 李滔等. 一种基于粗糙集的网页分类方法. 小型微型计算机系统，2003, 24(3): 520 – 523.

[8] Maudal O. Preprocessing Data for Neural Network based Classifiers: Rough Sets vs. Principal Component Analysis. Project Report, Department of Artificial Intelligence, University of Edinburgh, 1996.

[9] 王国胤. Rough集理论与知识获取. 西安: 西安交通大学出版社, 2001: 133 – 146.

[10] Wong S K M, Ziarko W. On optimal decision rules in decision tables. *Bulletin, Polish Academy of Sciences*, 1985, 33(11/12): 693 – 696.

[11] Skowron A, Rauszer C. The discernibility matrices and functions in information system. In Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht: Kluwer Academic Publishers, 1992: 331 – 362.

[12] 刘少辉, 等. Rough集高效算法的研究. 计算机学报，2003, 26(5): 524 – 529.

[13] Schutze H, Silverstein C. Projections for efficient document clustering. In Proceedings of ACM/SIGIR'97, Conference on Research and Development in Information Retrieval, Philadelphia, USA, 1997: 74 – 81.

盛晓炜： 男，1981年生，硕士，研究方向为自然语言处理.

江铭虎： 男，1962年生，博士后，副教授，主要研究领域为自然语言处理、神经网络信号处理、模式识别与人工智能等.