基于 Rough Set 的电子邮件分类系统*)

李志君　王国胤　吴　渝
(重庆邮电学院计算机科学与技术研究所　重庆400065)

摘　要　随着电子邮件的广泛使用,通过它进行不良信息传播的事件不断发生,电子邮件分类问题成为了网络安全研究的热点。本文通过对电子邮件头进行分析,运用 Rough Set 理论中相关的数据分析技术,建立了电子邮件分类系统的模型,并进行了实验测试,得到了满意的结果。

关键词　电子邮件,分类,Rough Set,数据挖掘,网络安全

An E-mail Classification System Based on Rough Set

LI Zhi-Jun　WANG Guo-Yin　WU Yu
(Institute of Computer Science & Tech.,Chongqing Univ. of Posts and Telecom.,Chongqing,400065)

Abstract　With the widespread using of email, junk mail becomes a severe problem both socially and technically. Email classification has become an important topic in network security researches. Through analyzing email headers with rough set technology, an email classification model is designed. It is proved to be effective through simulation.

Keywords　E-mail, Classification, Rough set, Data mining, Network security

1　引言

网络的产生,对人们的生活、工作产生了巨大的影响。随着互联网技术的发展,电子邮件在人们生活中扮演着越来越重要的角色。人们之间大量的交流都通过电子邮件的形式来进行,根据 IDC 调查,2000年全球日平均发送邮件超过100亿封。但是网络同样也存在着阴暗面,一些不良网站以及一些别有用心的人为蓬勃发展的网络蒙上了一层阴影。由于电子邮件协议的一些漏洞,许多垃圾邮件几乎是无孔不入、到处横飞。因此对电子邮件进行分类,帮助个人用户和邮件服务器摒除一些可疑邮件和垃圾邮件是必要的。

目前,国内外有很多科研机构在进行有关邮件分类方面的研究。邮件过滤也可以看作是一种分类,将邮件分成正常邮件和垃圾邮件两类。现在流行的一些邮件收发工具都给用户提供了一些过滤功能,这些过滤功能一般都是提供一些简单的规则,如:若邮件的发件人包含某一用户,则进行某一个动作;若邮件长度大于某一指定的长度,则进行某一动作;若邮件带有附件或者附件的名称为某一指定名称,则进行某一动作等等。虽然可以实现一定的功能,但是这些都是一些很粗略的呆板的过滤,这些规则必须由用户来自行指定,对用户的专业技能的要求较高。而且,当垃圾邮件以另外一种形式出现的时候,又必须手工对规则进行改动,不能进行智能处理。因此,很多研究者致力于可自动进行邮件分类(过滤)的工具的研究,如基于 Bayesian classifier[1~3],以及基于 broosting-trees[4],co-training[5]等等。这些方法大部分都是基于文本分类的,即根据一些关键词,将邮件分到几个子类中。但是,文本的大词汇量变成了文本分类的瓶颈。

邮件可分为用户正常通信的邮件和垃圾邮件。垃圾邮件又主要可以分为两类:一类是各种商业广告邮件;另一类则是可疑邮件,可疑邮件一般都刻意隐藏了发信人的真实信息,其隐蔽性强,危害性通常比一般的商业广告更大。这种分类具有一定的不确定性和不一致性,如某些邮件的特征具有一定程度的相似性,但有些属于正常邮件,有些却属于垃圾邮件。

Rough Set 理论是波兰科学家 Z. Pawlak 在1982年提出的一种处理含糊和不精确性问题的一种新型数学工具[6],主要应用在对不精确、不确定和模糊信息的表达与处理上。它从新的视角出发对知识进行了定义,知识被认为是一种将具体的或抽象的对象进行分类的能力,知识构成了某一感兴趣领域中各种分类模式的一个族集,这个族集提供了关于现实的显事实,以及能够从这些事实中推导出隐事实的推理能力[7]。针对电子邮件分类这种具有一定的不确定性和不一致性的分类问题,本文将从 Rough Set 数据分析理论出发来研究。这既给该问题的解决提供了一个新的方法,也是 Rough Set 理论应用于解决计算机网络安全问题的一次尝试。

2　电子邮件原理

2.1　传输机制

电子邮件的传送采用了一种特殊的方法:信件并不是直接送往目的地,而是采取"存储一转发"的方式,经过几台邮件服务器进行中转。如图1,通常一封邮件的发送是从用户发出,通过发送邮件客户端程序送到 SMTP 服务器,再经过一些邮件服务器进行转发,发送到收件人信箱所在的服务器,最后,由 POP3服务器程序配合帐号、口令接收信箱的邮件。

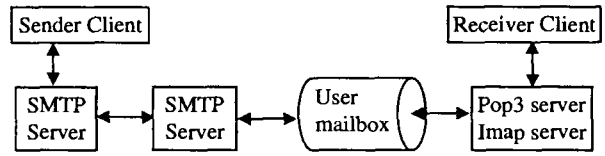

图1　Internet 邮件传送示意图

*)基金项目:国家自然科学基金(No. 69803014);攀登一特别支持费;高等学校骨干教师资助计划(No. GG-520-10617-1001);教育部重点项目;重庆市应用基础研究项目;重庆市教委科学技术研究项目(No. 020511,No. 020505)。李志君　硕士研究生,主要研究领域为网络安全;王国胤　博士,教授,主要研究领域为 rough set 理论,神经网络;吴　渝　博士,教授,主要研究领域为数据挖掘、小波分析和多媒体技术。

2.2 电子邮件的格式

电子邮件格式是由 RFC822所定义的，是半结构化的文本文件，包括邮件头和正文。其中，邮件头中含有 FROM，SENDER，REPLY-TO，TO，CC，BCC，MESSAGE-ID，DATE，RECEIVED，RESENT，SUBJECT，COMMENTS 等关键信息。

一封普通邮件的邮件头如图2所示。

```
Received: from localhost (localhost[127.0.0.1])
        by bjmx2.163.net (Postfix) with SMTP id D92D-
        61CD3810C
        for (lzj@163.net); Fri, 6 Sep 2002 14:03:20 +0800
        (CST)
Received: from optel.com.cn (unknown[211.101.180.116])
        by bjmx2.163.net (Coremail) with SMTP id TwEA-
        AChFeD2DFbR0.1
        for (lzj@163.net); Fri,06 Sep 2002 14:03:20 +0800
        (CST)
Received: from xiaowang[218.70.96.59] by optel.com.cn
  (SMTPD32-6.06) id A529A4400CC; Fri,06 Sep 2002 14:03:21
+0800
Date: Fri,6 Sep 2002 14:3:55 +0800
From: <xiaowang@optel.com.cn>
To: lzj <lzj@163.net>
Subject: hello!
X-mailer: FoxMail 4.0 beta 1[cn]
Mime-Version: 1.0
Content-Type: multipart/mixed;
Message-Id: <200209061403285.SM01000@ xiaowang>
```

图2 普通邮件的邮件头

3 基于 Rough Set 的分类模型与算法

本文将所有的邮件分为：正常邮件、广告邮件、可疑邮件三种，通过对这三类邮件分别选取一定数量的邮件进行特征提取，得到以上几类邮件对应的特征集。Rough Set 理论主要把关系数据库作为研究对象，通常把关系数据库抽象为一个信息系统。一个信息系统 S 是一个四元组[6,7]：

$S=(U,R=C\cup D,V,f)$，其中 $U=\{x_1,\cdots,x_n\}$ 是有限的对象集合即论域，$C=\{a_1,\cdots,a_k\}$ 为条件属性的集合，D 表示为决策属性集，$V=\{v_1,v_2,\cdots,v_k\}$ 是属性的值域集，v_i 是属性 a_i 的值域，f 是信息函数(information function)，$f:U\times R\rightarrow V$。一般只考虑只有一个决策属性的情况，对于多决策属性问题可以化为单决策属性问题处理。

电子邮件分类系统的粗糙集模型定义如下：

定义1 一个电子邮件分类系统是一个四元组 $S=(U,R=C\cup D,V,f)$。其中，U 是邮件的集合；R 为属性的集合，其中 C 为邮件特征的集合，D 表示决策属性，即邮件的分类，V 是属性值的集合，f 是信息函数，表示 U 上每个对象 x 的属性值。

定义2 U 上的不可分辨关系：假设属性集 $P\subseteq R$，对象 $X,Y\in U$，对于每个 $a\in P$，若有 $f(X,a)=f(Y,a)$，则称 X 和 Y 是不可分辨的，即：$ind(P)=\{(X,Y)\in U:\forall a\in P,(f(X,a)=f(Y,a))\}$

显然，ind(P)是一个等价关系。这样，属性集 P 可以认为是用等价关系(在该属性集上的取值相等)表示的一个知识的名称。因此，一个电子邮件过滤系统可视为一个知识库系统，当有邮件需要分析时，根据 ind(P)判断该邮件对应的知识库规则，按规则进行分类。

根据定义1，属性集合分为邮件特征集合和决策属性集合，在分析大量邮件头信息的基础上，结合邮件收发协议，我们定义了如下邮件特征集合：

$C=A0\cup A1\cup A2\cup A3\cup A4\cup A5\cup A6\cup A7\cup A8\cup A9\cup A10$

其中，这些特征的含义如下：

A0：邮件中继次数，即邮件头中“Received”标签的个数；

A1：收件人个数；

A2：邮件路由信息中的中断次数，邮件路由信息中断定义为：若前一条“Received”标签中接收站点的域名和 IP 与后一条“Received”标签中发送站点的域名和 IP 均不相同，则被视为路由信息中断；

A3：“Received”项中的各个域名与其 IP 不匹配的次数；

A4：“Received”中发送站点项缺少域名的次数；

A5：“Received”中的 by 项缺少域名的次数；

A6：“Received”中的 from 项缺少 IP 的次数；

A7：“From”中的原始发送地址与“Received”中的原始发送地址是否一致；一致则取值为1，否则取值为0；

A8：“To”中的目的地址与“Received”中的实际收信人的地址是否一致；一致则取值为1，否则取值为0；

A9：“Delivered-To”项和“To”项是否一致，若一致则取值为1，否则取值为0，若不存在“Delivered-To”项，则这一属性值的缺省值为1。

A10：“Return_Path”项和“From”项是否一致，若一致则取值为1，否则取值为0，若不存在“Return_Path”项，则这一属性值的缺省值为1。

决策表的决策属性描述为：普通邮件取值为1，广告邮件取值为2，可疑邮件取值为3。表1为一个邮件决策表示例。

表1 邮件分类决策表

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	D
3	1	1	0	0	1	0	0	1	0	0	2
1	3	0	0	0	0	0	1	0	1	1	1
2	1	0	0	0	0	0	1	1	1	1	1
6	1	4	0	4	3	1	3	3	1	1	3
2	1	1	0	1	1	0	0	1	1	1	1
4	1	1	0	0	1	0	0	0	0	1	3
6	1	4	0	4	3	1	0	0	1	1	2
2	1	1	0	2	1	0	0	1	1	1	1
1	1	0	0	0	0	0	0	0	1	1	2

基于 Rough Set 理论，从决策表中寻找决策规则的一般步骤[7]如下：

1) 数据预处理：包括删除重复记录、决策表补齐、数据离散化。经过比较，发现采用改进的贪心算法1[8]进行离散化，在邮件分类问题上可以得到比较好的效果。

2)删除多余属性，求出属性约简。在属性约简中，采用基于条件信息熵的属性约简 CEBARKNC 算法[9]进行属性约简，在邮件分类问题上可以得到比较好的效果。

3)删除多余的属性值，得到值约简。采用启发式值约简算法[10]进行约简，在邮件分类问题上可以得到比较好的效果。

4)根据值约简求出逻辑规则。结合各个属性的逻辑意义，对属性约简得到的规则进行分析，得到相应的逻辑规则。并根据邮件传输原理对这些逻辑规则进行分析和验证。

4 算法实验及结果分析

我们收集到了包含普通邮件、广告邮件以及可疑邮件三类邮件共13167封作为实验的数据集合，其中普通邮件为7586封，广告邮件为4831封，可疑邮件为750封。随机选取其中10%

的邮件作为训练数据，其中普通邮件为764封，广告邮件为463封，可疑邮件为90封，共1317封邮件。按照如前所述的属性描述逐个提取各邮件的特征信息，得到了包含1317个记录的邮件信息决策表，然后再对这个决策表进行规则提取。采用平台是重庆邮电学院计算机科学与技术研究所开发的 RIDAS (Rough Set Intelligent Data Analysis System)，该系统集成了有关 Rough Set 的30余种经典算法[11]。

在属性提取过程中，每封邮件的每个属性都对应唯一确定的值，因此得到的邮件决策表不存在缺失信息，故不需要对其进行数据补齐。

邮件决策表中，数据类型都是整型；虽然数据的离散化程度较高，但若不进一步离散化处理，所产生的规则的每个条件属性将都是一些数据点，导致系统的未识别率偏高；故仍需要先进行数据的离散化处理。

采用改进的贪心算法1的离散化算法[8]，得到断点集如表2所示。

经过属性约简，属性 A3被约简掉了。其实属性 A3是一个非常重要的属性，但是由于域名的动态性和网络资源的限制，我们在目前的实践中没有检验域名与 IP 的匹配信息。

表2 离散化处理后的各个字段取值

属性编号	离散后的分段数	分段描述
A0	6	[*,0.5),[0.5,1.5),[1.5,2.5),[2.5,3.5),[3.5,5.5),[5.5,*)①
A1	2	[*,1.5),[1.5,*)
A2	3	[*,0.5),[0.5,2.5),[2.5,*)
A3	1	[*,*)
A4	3	[*,0.5),[0.5,1.5),[1.5,*)
A5	3	[*,0.5),[0.5,1.5),[1.5,*)
A6	2	[*,0.5),[0.5,*)
A7	2	[*,0.5),[0.5,*)
A8	2	[*,0.5),[0.5,*)
A9	2	[*,0.5),[0.5,*)
A10	2	[*,0.5),[0.5,*)

注：①如属性 A_0所示，范围[*,0.5)用数字0表示，范围[0.5,1.5)用数字1表示，以此类推，则式子 $A0_1$表示当属性 A0取值范围为[0.5,1.5)，式子 $A0_{123}$则表示属性 A0的取值范围为[*,3.5]

经过值约简，我们从包含这1317个邮件特征信息的邮件信息决策表中得到了包含105条规则的规则库，下面对所得到的规则举例说明：

rule1：$A0_1 \wedge A8_1 \rightarrow D_1$

rule2：$A2_0 \wedge A10_0 \rightarrow D_2$

rule3：$A0_3 \wedge A1_0 \wedge A2_1 \wedge A4_1 \wedge A5_1 \wedge A6_0 \wedge A7_0 \wedge A8_0 \wedge A9_1 \wedge A10_1 \rightarrow D_3$

根据各个属性的逻辑含义上述三条规则的实际意义为：

rule1：如果邮件路由信息中的中断次数为1且填写的收信地址跟系统记录的原始发信地址符合，则该邮件为一封普通邮件。

rule2：如果邮件路由信息中的中断次数为0且填写的回复地址跟发信人不符合，则该邮件为一封广告邮件。

rule3：如果邮件路由信息中的中断次数为3，收信人的个数为1，邮件路由信息中的中断次数为1或者2，"Received"中发送站点项缺少域名的次数为1，"Received"中的 by 项缺少域名的次数为1，"Received"中的 from 项缺少 IP 的次数为0，"From"中的原始发送地址是否与"Received"中的原始发送地址不一致，"To"中的目的地址与"Received"中的实际收信人的地址不一致，"Delivered-To"项和"To"项一致，"Return-Path"项和"From"项一致，则该邮件为一封可疑邮件。

得到105条规则后，用13167封邮件作为测试数据进行测试，采用多数优先的策略[12]进行规则匹配，以处理规则之间的冲突和不一致，得到结果如表3所示。其中，错误识别包括把普通邮件识别成广告邮件、把广告邮件识别成可疑邮件等，其分布表如表4所示。

表3 识别率分布表

	识别数	百分比
正确识别	11615	88.2%
错误识别	1414	10.7%
未识别	138	1.1%

表4 错误识别分布表

邮件	数量	识别值	识别数	百分比
普通邮件	7856	普通邮件	7563	96.3%
		广告邮件	293	3.7%
		可疑邮件	0	0
广告邮件	4831	广告邮件	4175	86.5%
		普通邮件	596	12.3%
		可疑邮件	60	1.2%
可疑邮件	750	可疑邮件	285	38%
		普通邮件	0	0
		广告邮件	465	62%

5 分类算法在邮件分析系统中的应用

得到分类规则之后，再通过规则配置模块把这些分类规则应用到我们的邮件分析系统 MailAnalysis 中，就可以实现垃圾邮件的自动过滤。MailAnalysis 由以下模块组成：邮件接收模块、邮件路由分析模块、邮件分类模块和规则配置模块。其中，邮件接收模块包括接收邮件头模块、接收邮件模块和配置过滤器模块。邮件路由分析模块将邮件接收模块接收到的邮件头作为输入，对邮件的路由信息进行分析得到邮件从发送地址到最终接收地址的完整路由，进行判断和分析。邮件分类模块对邮件进行分类，并根据分类结果对邮件采取相应的动作(删掉或者下载到本地)。

该系统的工作流程如下：用户通过 POP3协议接收邮件，在接收邮件的时候，利用配置好的过滤器，将邮件进行过滤，把用户希望接收的邮件下载到本地；对于用户不确定的邮件，首先只接收邮件的邮件头，并将其作为邮件分类子模块的输入，从而得到该邮件的类型；然后再根据得到的结论决定是否接收这封邮件。

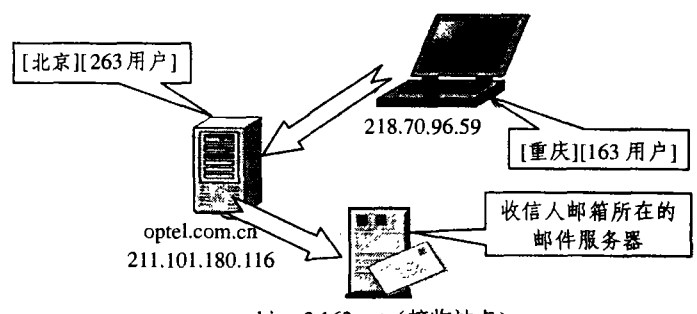

图3 邮件的发送路由

（下转第66页）

公司的各个责任或义务应标识成合同的一个部分。例如，B公司将鞋交付给A公司的责任可以表示成路径表达式contract.delivery。文档的各个条款也可以用路径表达式表示，例如交付日期可以表示成contract.delivery.date。现在条款也就能与提议的消息链接在一起，因此能从提议的条款中得到相关的信息。

结论 谈判是一个复杂的通信过程，它包括文档的交换。在电子谈判中通信交流是以书写的形式进行交换消息的，所涉及的文档是合同的各个版本。不论是文档还是消息是这种交互最基本的元素。我们通过将通信管理和文档管理相结合给B2B电子商务提供一个谈判支持架构。这个方法是以通信理论为基础的，为有效的谈判支持架构提供结构化的消息交换。谈判的结果是合同。我们认为合同版本控制和半结构化表示是必需的。

本文的主要贡献是通信管理与文档管理相结合的架构，它将通信管理与文档管理集成于一体。消息是以结构化的方式交换。文档管理因结合了通信得以加强。由于将消息与文档相结合，就能谈判的所有方面提供有效的支持。虽然我们的方法是以Searle和Habermas理论为基础，但不是说要学习他们的理论之后才能使用基于这个概念模型的实际系统。只是说明，这样的系统需要以用户为中心，获得必需的数据进行有效的设计，以保证实际的系统能提供有效的自动或半自动电子谈判媒体。

虽然本文提供的模型主要还是强调对谈判的有效"支持"，这主要因为是在复杂多问题的谈判中，功能再强、再智能化、再自动的电子谈判媒体也或多或少需要人类的参与，但这个谈判的概念模型为实现电子谈判系统提供了一个设计基础。

参考文献

1 Dumas M, Governatori G, ter Hofstede A H M. A formal approach to negotiating agents development. Electronic Commerce Research and Application, 2002, 1: 193～207

2 Strobel M. An XML schema representation for the communication design of electronic negotiations. Computer Network, 2002, 39: 661～680

3 Schoop M. An Empirical Study of Multidisciplinary Communication in Healthcare using a Language-Action Perspective. In: G. Goldkuhl, M. Lind, U. Seigerroth, eds. Proc. of the Fourth International Workshop on the Language Action Perspective on Communication Modelling (LAP 99) pages 59～72

4 Searle J, Vanderveken D. Foundations of Illocutionary Logic. Cambridge University Press, Cambridge, 1985

5 Schoop M. Habermas and Searle in Hospital: A Description Language for Cooperative Documentation Systems in Healthcare. In: F. Dignum, J. Dietz, eds. Communication Modeling - The Language/Action Perspective; Proc. of the Second International Workshop on Communication Modeling, Computing Science Reports 97-09, The Netherlands, 1997. 117～132

6 Schoop M. A Theoretical Framework for Speech Act Based Negotiation in Electronic Commerce. In: Proc. of the Sixth Research Symposium on Emerging Electronic Markets (RSEEM), Munster, Germany, 1999

7 Verharen E. A Language-Action Perspective on the Design of Cooperative Information Agents: [PhD thesis]. Katholieke Universiteit Brabant, Tilburg, The Netherlands, 1997

8 李一军，丁伟，曹荣增. 面向电子商务的谈判支持系统. 管理学报，2001. 6

（上接第60页）

例如，如图2所述的一封普通邮件，通过邮件路由分析模块分析，得到如图3所示的结果，发信人xiaowang首先登陆到他的邮件服务器optel.com.cn，然后通过邮件服务器optel.com.cn将邮件发到收信人lzj@163.net地址所在的邮件服务器bjmx2.163.net。

接收邮件后，先通过邮件分类模块，得到这封邮件11个条件属性值为：(3,1,1,0.1,0.0,1,1,1,1)，然后再按照多数优先策略进行规则匹配，该邮件属性匹配上规则：$A0_{345} \wedge A1_{012} \wedge A2_{01} \wedge A4_1 \wedge A5_0 \wedge A6_0 \wedge A7_0 \wedge A8_1 \wedge A9_1 \wedge A10_0 \rightarrow D1$得到决策，该邮件为普通邮件，按照用户已定义好的策略，将这封邮件下载到本地阅读。

结论 电子邮件分类技术是当前网络信息安全研究的一个热点问题，本文用Rough Set理论进行系统建模和数据分析，取得了较好的效果，是该理论在计算机网络安全领域的一次尝试。影响电子邮件分类的特征很多，本文仅考虑了邮件头里的一些特征。为了得到满意的邮件分类系统，我们还需要进一步完成以下工作：

1)得到A3这一重要属性，属性A3即"Received"项中的各个域名与其IP不匹配的次数是一个非常重要的属性，但是由于域名的动态性和网络资源的限制，我们在目前的实践中没有检验域名与IP的匹配信息；

2) 扩展现在有的条件属性，根据实验结果添加一些对决策有影响的条件属性以降低系统的错误识别率，尤其是将正常邮件判错的错误识别率；

3) 将邮件头的特征和邮件文本的特征结合进行分析，邮件文本的一些特征对分类的判断来说，也是非常重要的，本文只考虑了邮件头的特征，并没有考虑文本的特征，将二者结合起来考虑能够提高系统的正确识别率。

参考文献

1 Lewis D D, Ringuette M. A comparison of two learning algorithms for text categorization. In: Proc. of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 1994. 81～93

2 Androutsopoulos I, Paliouras G, Karkaletsis V, Sakkis G, Stamatopoulos P. Learning to filter spam e-mail: a comparison of a naive Bayesian and a memory-based approach. In: Proc. of the workshop "Machine Learning and Textual Information Access", 4th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000), Lyon, France, Sep. 2000. 1～13

3 Sahami M, Dumais S, Heckerman D, Horvitz E. A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization Papers from the 1998 Workshop: [AAAI Technical Report WS-98-05]

4 Carreras X, Mrquez L. Boosting trees for anti-spam email filtering. In: Proc. of RANLP-01, 4th Intl. Conf. on Recent Advan-ces in Natural Language Processing, Tzigov Chark, BG, 2001

5 Kiritchenko S, Matwin S. Email classification with Co-training. In: Proc. Of CASCON 2001, Toronto, Canada, 2001. 192～501

6 Pawlak Z. Rough Set. International Journal of Computer and Information Sciences, 1982, 11(5): 341～356

7 王国胤. Rough集理论与知识获取. 西安交通大学出版社，2001

8 候利绢，王国胤，聂能. 粗糙集理论中的离散化问题. 计算机科学，2000, 27(12): 89～94

9 王国胤，于洪，杨大春. 基于条件信息熵的决策表约简. 计算机学报，2002, 25(7): 759～766

10 苗夺谦，胡桂容. 知识约简的一种启发式算法. 计算机研究与发展，1999, 36(6): 681～684

11 Wang G Y, Zheng Z, Zhang Y. RIDAS-A rough set based intelligent data analysis system. In: Proc. of the First Intl. Conf. on Machine Learning and Cybernetics, 2002. 646～649

12 Wang G Y, Liu F, Wu Y. Generating rules and reasoning under inconsistencies. In: Proc. of IEEE Intl. Conf. on Industrial Electronics, Control and Instrumentation, Japan, 2000. 2536～2541