

# 基于粗糙集约简算法的配置文本聚类方法研究

唐启涛,张 燕,彭利红

(长沙医学院 计算机科学与技术系,湖南 长沙 410219)

**摘要:**当前的网络设备的配置文本日趋复杂,在网络设备出现故障时采集到的数据量也随之倍增。但是,并不是所有的配置文本信息都是有用的。文中提出了通过对配置文本应用基于 DAG 思想的配置元集无关性算法,实现消除配置文本中的冗余信息,只保留配置文本中有用的信息。对于每一次网络设备配置故障诊断,因所采用通信信道、采集设备的不同致使获得的信息无法保证它的完备性和正确性,因此,想获得理想的故障诊断结果通过传统的方法行不通。在对配置文本信息进行了预处理后,为了便于对网络设备配置文本故障进行智能、快速的诊断,提出了一种基于粗糙集约简算法的配置命令文本聚类方法,实现按功能的不同对预处理后的设备配置命令文本进行分类。最后,利用 Simulink 仿真软件比较配置文本归类与不归类在故障诊断时的差别。

**关键词:**网络设备;配置元集;故障诊断;粗糙集;有向无环图

**中图分类号:**TP301

**文献标识码:**A

**文章编号:**1673-629X(2015)11-0105-05

**doi:**10.3969/j.issn.1673-629X.2015.11.021

## Research on Clustering Method of Configuration Text Based on Rough Sets Reduction Algorithm

TANG Qi-tao, ZHANG Yan, PENG Li-hong

(Department of Computer Science and Technology, Changsha University of Medical,  
Changsha 410219, China)

**Abstract:** Configuration text of the current network equipment is becoming more complicated, the amount of information acquisition is more and more in network equipment malfunction. However, not all information of configuration text is useful. The configuration element of ideologies of independent algorithm is put forward based on DAG used in configuration text, which eliminates redundant information, retains only the useful information in the configuration text. For the every configuration fault diagnosis of network equipment, because the acquisition equipment, communication channel and other factors make the information obtained cannot guarantee the complete and correctness, therefore, the fault diagnosis results with traditional methods is not ideal. After the configuration text information is preprocessed, in order to facilitate the intelligent and rapid diagnosis for the network equipment configuration text fault, present a method of configuration text clustering based on rough sets reduction algorithm, realizing classification of network equipment configuration text command by different function after pretreatment. Finally, compare the configuration text difference of text categorization and not classified in fault diagnosis by using Simulink simulation software.

**Key words:** network equipment; configuration element set; fault diagnosis; rough set; DAG

## 0 引言

由于当前的网络设备功能越来越强大,其对应的配置文本信息量与日俱增。在各种网络设备的配置文本中,并不是所有的信息都是有用的,存在很多冗余信息。如何消除网络设备配置文本的冗余信息,国内外专家学者提出了很多好的方案,典型的有基于有穷自动机理论的配置元集无关性算法<sup>[1]</sup>。该算法能对配置

文本在一定程度上消除冗余,但不彻底,生成的状态树中,存在相同的元素。为此,文中提出了一种基于 DAG 思想的配置元集无关性算法,进一步消除在生成的状态树中存在的相同元素,在保证配置文本命令能正常识别的情况下,使各种元素在一个状态树中只出现一次。同时,在日常的网络设备故障诊断中,由于采集设备、通信信道、采集的时间等原因使得诊断时获得

收稿日期:2015-02-14

修回日期:2015-05-19

网络出版时间:2015-11-04

基金项目:湖南省教育科学技术研究项目(14C0114)

作者简介:唐启涛(1975-),男,硕士,讲师,研究方向为计算机网络及信息安全。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20151104.0952.062.html>

的结果信息无法保证其完备性和正确性,所以,单纯的想通过传统的方法获得理想的故障诊断结果是行不通的<sup>[2]</sup>。

基于这些原因,国内外专家学者提出了诸多不同的解决策略,其中基于决策树的方法能很好地处理冗余属性问题,而且在配置文本的分类上提供了方便,获得了很多专家的重视<sup>[3]</sup>。但是,一旦面对海量数据信息想通过应用决策树进行最优约简时,因很难找到合适的映射函数进行模型训练,所以对于当前的各种网络设备的应用难度较大。对属性集的全部约简中,包含属性数量最小的属性集称之最优约简<sup>[4]</sup>。为了找到一个能适应当前网络设备配置文本聚类的可行性算法,如何优化现有的粗糙集约简算法就成为了一个迫切需要解决的问题。针对这个问题,不少学者也提出了很多相关的约简算法。例如基于特征选择的属性约简算法、基于互信息的属性约简算法<sup>[5]</sup>、归纳属性约简算法等<sup>[6]</sup>。这些约简算法的基本思路是从粗糙集的核心出发,应用信息熵与启发式算法,算出各种特征属性的特征值,再把已有信息集合中的添加信息增益最大的属性挑选出来,最终构造最优约简,但这些算法当前还存在着无法进行全局最优寻找,计算复杂度高等不足<sup>[7]</sup>。

文中在分析了粗糙集约简特点的基础上,提出了一种基于粗糙集约简算法的网络设备配置文本归类方法,能够在现有的网络环境及设备条件下,获得相对最优的约简组合,从而为获得正确的网络设备故障诊断结果提供了保障。

## 1 DAG 概述

一个无环的有向图称作有向无环图 (Directed Acycline Graph, DAG)<sup>[8]</sup>。DAG 图是一类特殊有向图。可以利用有向无环图描述含有公共子式的表达式, DAG 是一种描述表达式的有效工具。例如下述表达式  $(a + b) * (b * (c + d) + (c + d) * e) * ((c + d) * e)$  可以用 DAG 图来表示,也可以用二叉树来表示<sup>[9]</sup>。通过认真观察表达式可发现有一些子表达式是相同的,如  $(c + d) * e$  和  $(c + d)$  等。当用二叉树来描述表达式时,子表达式也重复出现,如图 1 所示。当改用有向无环图表示时,因可以实现对相同子表达式的共享,表示方式更简洁,使用的存储空间更小,如图 2 所示。

在工程有向图中就是使用有向图表示一个工程,用有向边表示活动开展的先后关系,例如有向边  $< V_i, V_j >$  表示活动  $V_j$  必须在活动  $V_i$  完成之后才能进行,用顶点表示活动,通常把这种有向图称为 AOV 网络,它可以有效提高工程效率<sup>[7]</sup>。

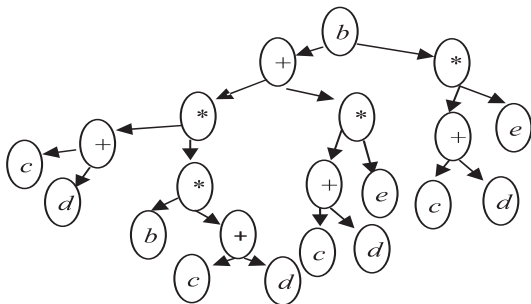


图 1 用二叉树描述表达式

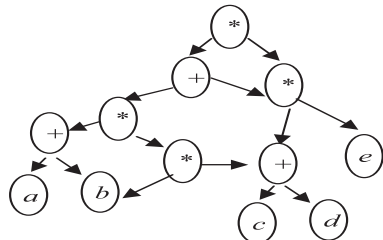


图 2 描述表达式的有向无环图

## 2 配置元集

网络设备的配置文档是一种有序的结构化命令集合,信息元就是配置文本中结构化的命令,通常把配置文本作为一个有序元集,通过这个元集实现网络设备的所有操作<sup>[10]</sup>。可以利用有向无环图 DAG 理论识别出配置文本的信息元。网络设备的配置文本是一种控制结构和语法规则都较为自由、松散的文档。配置文本中常用的单词符号有四种<sup>[11]</sup>:

- (1)标识符:用来表示各种变量名,如网络设备名、用户名等;
- (2)界符:如方括号、感叹号、百分点等;
- (3)关键字:由生产厂家预先定义的具有特殊意义的符号;
- (4)整型常数:用来表示端口号、IP 地址等类型的整数。

通过以上符号的组合构成相应的网络设备的配置命令,在网络设备配置文本中,通常单词之间有空格字符隔开,命令之间有换行符区分,这些为后续工作提供了便利。

因由不同厂家生产的同一种网络设备具有不同的配置命令集合,即使是同一厂家生产的不同型号的网络设备产品也存在着一定的差异,在所有的网络设备配置文本中没有一套统一的配置命令集,即不同的产品使用的配置命令集合都有所不同<sup>[12]</sup>。为了能有效识别出各种配置元集,应用于不同型号和不同版本的网络设备的配置管理中,实现配置命令管理的智能化,文中提出了基于 DAG 理论的配置元集无关性算法。它能应用于不同类型和多种型号网络设备配置文本,在算法中,把配置元集与有向状态图进行一一对应,从

而实现不同配置元集对应不同状态图,当系统在相应的网络设备上运行时,只要找出配置元集对应的状态图,就能快速、有效地识别出配置文本中的所有命令。

3 基于 DAG 思想的配置元集无关性算法

用  $C$  表示输入某具体元集的集合,原始数据由识别器提供,这个集合的初值为 NULL,集合  $C$  中信息元的个数是动态变化的,具有一定的随机性。信息元与信息元之间由特定符号隔开,每个信息元中的数字、可选配置项、标识符等用特定符号标示出来,通常可以将一条含有标识符、整型数字、关键字、可选配置项的信息元命令表示如下:

关键字 | 可选配置项 \* 标识符 \* | 数字 %

结合命令信息元的表示形式,可以进一步简化表示方式,比如集合  $C$  中输入的“数字”用 \$ 表示,输入的标识符用 # 表示,对于并列的可选配置项用 “|” 表示,可以实现将网络设备的配置命令信息元根据给定的规则拆分成多条命令信息元,再按要求集合  $C$  中加入网络设备的配置元集。具体操作流程如下:

把状态图  $T$  中的 root 看作图根,算法简要描述如下<sup>[13]</sup>:

- (1) 预处理:除  $C$  中所有关键字、标识符、整型数字及可选配置项外,其他多余全部删除;
- (2) 假设  $C$  不为空,顺序读取  $C$  中的一条命令  $L$ ,用符号 % 把每条命令隔开;否则,结束程序;
- (3) 假设  $L$  不为空,依次把  $L$  中的单词  $w$  读取出来;假设为空,则跳转到(2)继续执行;
- (4) 假设  $w \neq '|'$ ,则在图  $T$  的基础上重新构建一个新图  $T$ ,该图以 root 为根,把每个单词作为一个节点,在图中没有重复的兄弟节点,用表达式可以表示为  $T = T + w$ ;否则跳转到(5)继续执行;
- (5) 假设符号  $w = '|'$ ,而且是第一次出现,则假设  $q$  为指向  $w$  的前一个节点,然后跳转到(3)继续执行;假设第 2 个期待的单词为 |,则将符号 | 后的单词和  $q$  指向的单词建立连接,当  $q$  为空时,跳转到(3)继续执行。

假设现有以下几条命令<sup>[14]</sup>:

- (1) 对 ospf 路由重新分配: redistribute ospf 1 metric 10000 100 255 1 1500
- (2) 对直连的路由重新分配: redistribute connected subnet;
- (3) 对 rip 路由重新分配: redistribute rip metric 50000 500 255 1 1500。

在集合  $C$  中存储上述命令时的表达式可以简化为:

redistribute connected subnet% redistribute ospf 1

metric 10000 100 255 1 1500 % redistribute rip metric 50000 500 255 1 1500%

将上述命令预处理为:

redistributeconnectedsubnet% redistributeospfYmetric \$ % redistribri buteripmetric \$ %

配置命令信息元应用算法优化处理后可以得相应的状态图,如图 3 所示。

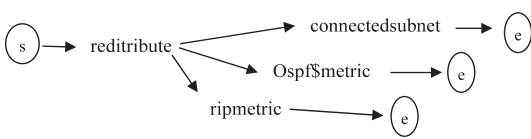


图 3 算法举例

图中的圆圈表示命令的开始和结束,当圆圈中有字母  $s$  时,代表一条命令的开始,当圆圈中有字母  $e$  时,代表一条命令的结束。

结合上述算法,在读取完集合  $C$  中的符号后,将会形成一个包含若干个叶子节点,一个根节点,所有子节点均不相同的配置文本命令状态图;当把信息元命令继续添加到集合  $C$  中时,能够实现动态更新所形成的配置命令状态图。针对不同的配置文本命令集,根据算法将生成相应的配置命令状态图。

4 粗糙集概述

波兰学者 Pawlak Z 在 1982 年提出了 RS 理论,它是一种表达不完整性 and 不确定性的数学工具,适合于处理不一致、不完整、不精确等不完备信息,可以应用该工具发现不完备信息中隐含的知识<sup>[14]</sup>。粗糙集理论的主体内容是在分类能力保持不变的情况下,根据给出的知识库,通过应用相应的约简算法实现对知识的约简,最终把概念的分类规则导出。

用一个四元组表示的信息系统  $I$ ,其表达形式通常为:  $I = \langle U, ?, A, f \rangle$ 。其中,  $U$  为论域,表示全体对象的集合;  $?$  表示属性的集合;  $A$  表示属性值域集合,  $D$  为决策属性集合,  $C$  为条件属性集合。对于该信息系统  $I$ ,设  $c_{ij}$  表示可辨识矩阵中第  $i$  行第  $j$  列的元素。这样可辨识矩阵可定义为<sup>[15]</sup>:

$$c_{ij} = \begin{cases} \{a_k | (a_k \in A) \wedge (A_k(x_j))\}, D(x_i) \neq D(x_j) \\ 0, D(x_i) = D(x_j) \end{cases}$$

其中,  $i, j = 1, 2, \dots, n$ 。

定义 1:不可分辨关系。

对于某个属性子集  $B = U \times U$ ,如果  $IND(B) = \{(x, y) | (x, y) \in U_2, b \in B(b(x) = b(y))\}$ ,则把  $IND(B)$  称为等价关系<sup>[3]</sup>。

定义 2:决策表。

决策表是一种特殊的信息知识表达系统,它可以



用五元组的方式来表示,具体表达形式为: $S = \langle U, C, D, V, f \rangle$ 。其中,子集  $D$  和  $C$  分别称为决策属性集和条件属性集。把  $U$  是看成对象的集合,  $R$  表示属性集合,属性  $r \in R$  的范围用  $V_r$  表示,其中  $f: U \times R \rightarrow V$  是一个信息函数。决策属性  $D$  和条件属性  $C$  的等价关系  $IND(D)$  和  $IND(C)$  的等价类分别称为决策类和条件类<sup>[16]</sup>。在配置文本的规则分类中,规则分类的前提条件是能从文本中提取出关键词向量,同时,规则的决策用文本所属的类别来表示。

对信息系统进行属性约简是粗糙集理论的一个基本应用,它在分类能力保持不变的情况下,根据给出的知识库,通过应用相应的约简算法实现对知识的约简,最终把概念的分类规则导出。

可省略关系与不可省略关系概念:设  $R$  为一个等价关系集合,且  $r \in R$ ,当  $IND(R) = IND(R - \{r\})$ ,则称  $r$  为  $R$  中可省略的,否则称  $r$  为  $R$  中不可省略的。

定义 3:绝对约简。  
设  $U$  为一个论域,  $Q$  和  $P$  作为定义在  $U$  上的两个等价关系簇,而且满足  $Q \in P$ ,假设  $Q$  是独立的,且  $IND(Q) = IND(P)$ ,则  $Q$  就是  $P$  的一个绝对约简。

- 决策表的约简步骤可以概括为以下三步:
- (1)简化条件属性,简单来说就是针对决策表中没有用的列进行删除。
  - (2)消除决策表中的冗余信息,把表中重复的行删除。
  - (3)消除决策表中列属性的冗余值。

决策表的简化就是化简表中的条件属性,化简后的决策表条件属性变少但其功能不变,化简后的结果可以用作信息系统的分类规则<sup>[17]</sup>。

5 基于粗糙集理论的配置文本归类方法

5.1 基于粗糙集的网络设备配置文本聚类过程

配置文本的聚类过程中,首先需要选择相应的配置文本作为训练文本,在此基础上再进行关键词的提取及相关属性的约简,最终来验证文本的训练结果。具体操作如下<sup>[18]</sup>:

- (1)定义类别集合,从网络设备配置文本已提取的数据库选出 300 个文本命令,其中路由功能、网络端口管理功能、用户权限管理功能各取 100 个训练文本,每个训练类别文本由人工标上唯一的类别标志  $C_1, C_2, C_3$ ;
- (2)对训练文本按配置文本命令进行分词切分,使其变成一个无序词条的集合,然后运用 Zipf 法则滤掉出现频率过低或者过高的词条,再消除无用词选出特征项。文本命令中首次出现的词条对于分类具有重要作用,应全部保留下来;

- (3)构造决策表,把配置文本所属类别作为决策属性,把关键词向量集作为决策表的条件属性;
- (4)将特征项的权值进行离散化处理;
- (5)使用前面提到的决策表约简方法简化决策表,得到的化简结果就是一个规则集合;
- (6)验证训练结果的正确性。

5.2 配置文本聚类的主要技术

决策表的构造过程是将文本所属的类别作为决策属性集,特征项集合作为规则的条件属性集,对于在某一训练文本中不出现的特征项,规定其权值为空集。表 1 为实验中决策表的一部分。

表 1 信息决策表

	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$	$t_7$	$t_8$
$C_1$	1	3	2	4	8	10	2	*
$C_2$	4	3	1	6	9	3	6	8

其中,数字表示特征词条的权重; $t_1, t_2, \dots, t_8$  为特征词条; $C_1, C_2$  为文本的分类标识;\* 表示空值。

在训练阶段,每个网络设备配置文本都将单独生成一个相关类别的关键词集合,最终通过简化为每个网络设备配置文本生成一条相应的分类规则。通常采用向量空间模型来表示配置文本。在该模型中,网络设备的配置文本通常被看成一个相关无序词条的集合,对每个词条标注上相应的权值,再将配置文本映射成相应的一个特征向量:

$$V(d) = (t_1, \omega_1(d); \dots; t_n, \omega_n(d))$$

其中,  $t_i$  为词条项;  $\omega_i(d)$  为  $t_i$  在  $d$  中的权值。  
选用的词条权值计算方法为 TF-IDF 公式<sup>[18]</sup>:

$$\omega_{id}(d) = \frac{tf_{ik}(d) \log(N/n_k + 0.01)}{\sqrt{\sum_{k=1}^n tf_{ik}(d)^2 \log^2(N/n_k + 0.01)}}$$

其中,  $N$  表示全部样本配置文本的总数;  $tf_{ik}(d)$  表示词条  $t_k$  在配置文本  $d$  中出现的频率;  $n_k$  表示包含词条  $t_k$  的配置文本数。

6 仿真实验结果

Simulink 的一个突出特点是支持图形用户界面,它是 MATLAB 软件中一个软件包,用户通过简单的单击和拖动鼠标的动作就能完成建模操作<sup>[19]</sup>。使用 Simulink 进行系统建模的任务就是如何选择合适的模块并把它们按自己的模型结构连接起来,最后进行调试和仿真。

结合以上提出的算法,利用 Simulink 软件进行了仿真实验,结果如图 4 和图 5 所示。

通过图 4 可以看出,网络设备配置文本在归类后,如果配置文本信息不大,在诊断结果的时间响应方面,

两者的区别不大,随着配置文本的信息量增大,两者的响应时间差别变得更大,配置文本信息量越多,归类后的配置文本在诊断结果响应时间方面更有优势。

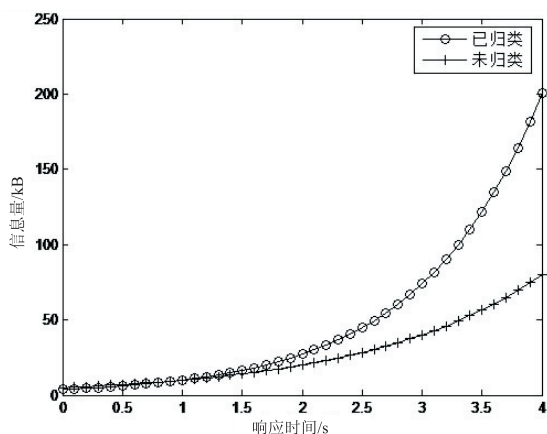


图4 诊断响应时间比较

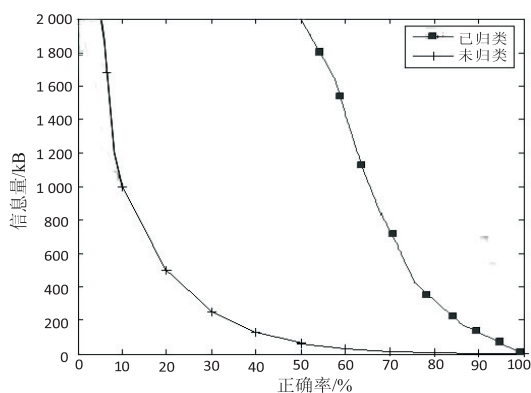


图5 诊断正确率比较

图5显示了归类的配置文本与未归类的配置文本在得出诊断结果的情况下,诊断结果正确率的比较。可以看出,归类后的配置文本在诊断结果正确率方面明显优于未归类的网络设备配置文本。

虽然文中算法不能保证网络设备配置文本诊断结果的完全正确,但在网络设备配置文本的诊断结果响应时间和诊断结果正确率方面有了很大提高。

## 7 结束语

在实际的网络设备故障诊断中,待处理的信息量非常大,而其中对于网络设备故障诊断有帮助的数据只是其中很小一部分,通过应用基于 DAG 思想的配置元集无关性算法有效去除了冗余信息文本,在此基础上,进一步应用粗糙集进行约简。文中提出了一种基于粗糙集约简算法的网络配置命令文本归类方法,通过不断调整相应的适应值函数及相关的键参数,使其能适应网络设备配置文本的分类。在 Simulink 仿真环境下比较了配置文本已归类和未归类情况下对网络设备故障诊断的影响,结果显示该算法具有把关键信息从海量信息中提取的能力,表明了算法对于实际的

网络设备出现的故障可以得出有效、正确的故障诊断结果信息,可以优化网络设备故障的智能修复能力。

## 参考文献:

- [1] 张冬慧,孙波,徐照财,等. 文本自动分类关键技术研究[J]. 微计算机信息,2008(6):197-199.
- [2] 杨创新. 基于机器学习的高性能中文文本分类研究[D]. 广州:华南理工大学,2009.
- [3] 常景超. 网络设备配置策略技术研究[D]. 衡阳:南华大学,2008.
- [4] 庞剑锋,卜东波,白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究,2001,18(9):23-26.
- [5] 周卫星,廖欢. 基于K均值聚类和概率松弛法的图像区域分割[J]. 计算机技术与发展,2010,20(2):68-70.
- [6] 张苗,张德贤. 多类支持向量机文本分类方法[J]. 计算机技术与发展,2008,18(3):139-141.
- [7] 龚静,胡平霞,胡灿. 用于文本分类的特征项权重算法改进[J]. 计算机技术与发展,2014,24(9):128-132.
- [8] 杨金柱,刘金岭. 基于词语上下文的文本分类研究[J]. 计算机技术与发展,2011,21(8):145-148.
- [9] 何峰. 一种基于粗糙集理论的文本分类方法[J]. 自动化与信息工程,2006,27(3):1-3.
- [10] 凌维业,贾民平,许飞云,等. 粗糙集神经网络故障诊断系统的优化方法研究[J]. 中国电机工程学报,2003,23(5):98-102.
- [11] 陶志,许宝栋,汪定伟,等. 基于遗传算法的粗糙集知识约简方法[J]. 系统工程,2003,21(4):116-122.
- [12] 于达仁,胡清华,鲍文. 融合粗糙集和模糊聚类的连续数据知识发现[J]. 中国电机工程学报,2004,24(6):205-210.
- [13] 孙秋野,张化光,戴璟. 基于改进粗糙集约简算法的配电系统在线故障诊断[J]. 中国电机工程学报,2007,27(7):58-64.
- [14] 黄文涛,赵学增,王伟杰,等. 基于粗糙集理论的故障诊断决策规则提取方法[J]. 中国电机工程学报,2003,23(11):150-154.
- [15] 刘文,吴陈. 一种新的中文文本分类算法—One Class SVM-KNN 算法[J]. 计算机技术与发展,2012,22(5):83-86.
- [16] 周静,曾国荪. 基于 DAG 图的自适应代码划分优化算法[J]. 计算机工程,2007,33(20):15-17.
- [17] 柴永生,吴秀丽,孙树栋,等. 设备管理信息系统及其关键技术研究[J]. 计算机工程与应用,2004,40(12):212-215.
- [18] 王德年. 网络设备应用技术[M]. 北京:中国铁道出版社,2007.
- [19] 李增智,朱海萍,唐亚哲,等. 一种故障诊断专家系统在网路管理中的设计与实现[J]. 计算机工程与应用,2001,37(17):24-26.

基于粗糙集约简算法的配置文本聚类方法研究

作者：[唐启涛](#)，[张燕](#)，[彭利红](#)，[TANG Qi-tao](#)，[ZHANG Yan](#)，[PENG Li-hong](#)  
作者单位：[长沙医学院 计算机科学与技术系, 湖南 长沙, 410219](#)  
刊名：[计算机技术与发展](#)[ISTIC](#)  
英文刊名：[Computer Technology and Development](#)  
年，卷(期)：2015, 25(11)

引用本文格式：[唐启涛](#). [张燕](#). [彭利红](#). [TANG Qi-tao](#). [ZHANG Yan](#). [PENG Li-hong](#) [基于粗糙集约简算法的配置文本聚类方法研究](#)[期刊论文]-[计算机技术与发展](#) 2015(11)