基于粗糙集理论的中文文本主客观性研究

李龙澍,张晓红,赵志伟
(1. 安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2. 安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:随着互联网的发展,各种文本信息资源量急剧增加。如何从海量信息中挖掘出这些有用的并进行分类、获取其中潜在的信息,已经成为数据挖掘、知识发现和中文文本处理等领域的一项最新的研究课题。粗糙集理论是一种对不确定数据进行分类的数学工具,在保持基本的分类能力不变的情况下,进行属性和知识约简,从而减少数据挖掘的原始数据,提高知识发现的效率。将粗糙集属性约简算法运用于文本主客观性研究中,提高了文本主客观性判断的效率。

关键词:文本处理;主观性;粗糙集;属性约简
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2011)06-0112-04

Research of Chinese Sentences Subjectivity and Objectivity Based on Rough Set Theroy

LI Long-shu, ZHANG Xiao-hong, ZHAO Zhi-wei
(1. Ministry of Education Key Laboratory of Intelligent Computing and Signal Processing, Anhui University, Hefei 230039, China;
2. School of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: As the development of internet, many kinds of text information have increased sharply. How to get and classify the information and finally get the potential information has become a new research subject in data mining, knowledge discovery and text processing areas. Rough set theory is a mathematical tool used in classification of indefinite data. The application of the rough set theory for image classification is to utilize equivalence relation classes through attribution reduction and decision rule reduction, and to obtain the reduction of knowledge and improve the efficiency of date mining in the case of keeping the same ability for classification. Apply the attribution reduction of rough set theory to the research of subjectivity in text processing to improve the efficiency to judge the text subjectivity.

Key words: text processing; subjectivity; rough set theory; attribution reduction

0 引 言

近年来,在研究不完整数据和不精确知识的表达、学习、归纳等方法的基础上,波兰华沙理工大学的科学家帕克拉(Z. Pawlak)基于"知识(人的智能)就是一种分类能力"的观点,于1982年开创性地提出了粗糙集理论[1](Rough Set Theory)。粗糙集理论具有很强的定性分析能力,能有效地表达不确定的或不精确的知识,善于从数据中获取知识,并能利用不确定、不完善的经验知识进行推理等,因此在知识获取、规则生成、决策分析等领域获得了广阔的运用,特别是在数据挖掘领域,取得了巨大的成功。

随着Internet的迅猛发展,人们在网络上从事的活动越来越多,但是这些活动能够成功进行的前提是必须了解这些活动的相关信息。这些规则的来源就是大量的网络文本资源,如,淘宝网商品的品论详情、宝贝详情、成交记录等。从这些数据资源中获取信息,首先要分清哪些是主观性评论、哪些是客观性描述、哪些是事实的记录,例如文献[2]的研究结果,这样才能有助于人们对不同性质的评论做出相应的判断。

粗糙集在一些文本自动分类研究中获得了较好的效果[3~5],这些研究主要集中于属性约简算法。目前国内对中文文本主客观性的研究还处于初步阶段,例如文献[6],运用粗糙集理论提取中文文本中的名词性句子;文献[7]提出一种提取中文观点句中评价对象和评价词主观匹配关系的方法。分析观点句中评价词和评价对象的词性、词语位置,通过句法分析获取语义特征,将2类特征应用于最大熵模型,提取观点句的主观关系。文献[8]对比句子的主观性和关系从句的

收稿日期:2010-11-17;**修回日期**:2011-02-21
基金项目:安徽省自然科学基金(090412054)
作者简介:李龙澍(1956-),男,安徽亳州人,教授,博士生导师,研究方向为知识工程、软件分析与测试。

客观性这两个不同的语义特征，总结英、汉语句子主观性的语言实现（如语音调节、情态副词、句法移动、句末语气词等）不能出现在关系从句中的语言事实，并评述 Cinque（1999，2004）、Rizzi（1997，2002）和 Chomsky（1998，2001）等对主观性相关内容的句法研究。最后在最简方案理论框架下分析携带[-subjectivity]（[-主观性]）特征的关系从句与携带[+subjectivity]（[+主观性]）特征的句子之间的句法联系与区别。文献[9]选取了 Sougou 实验室提供的中文新闻稿作为数据进行测试，在粗糙集理论的基础上对中文文本的特征进行选择。

这篇文章在研究现有的中文文本主观性识别方法[10-12]的基础上，通过提取大量主观和客观句子，运用分词程序对这些句子首先进行分词处理，然后得出影响这些主观性句子主观性的结构模型（主观性模型）。并且在进行主客观判断之前，如果运用粗糙集的约简算法，对主客观识别模型进行知识约简，在不影响最终的判断结果的前提下，减少主观句判断的执行时间，提高效率。

1　粗糙集理论

1.1　知识和概念

设 U 是所有研究对象组成的非空有限集合，称作一个论域。论域 U 的任何一个子集 $X \subseteq U$，称作 U 的一个概念或是范畴。论域 U 的任何子集簇（概念簇）称作关于 U 的抽象知识，简称知识。

1.2　知识库

给定一个论域 U 和 U 上的一簇等价关系 S，称二元组 $K=(U,S)$ 是关于论域 U 的一个知识库或近似空间。

知识库表示了论域上由等价关系导出的各种各样的知识，即划分或分类模式，同时代表了对论域的分类能力，并隐含着知识库中概念之间存在的这种关系。

1.3　不可分辨关系

给定一个论域 U 和 U 上的一簇等价关系 S，若 $P \subseteq S$，且 $P \neq \varnothing$，则 $\cap P$（P 中所有等价关系的交集）仍然是论域 U 上的一个等价关系，称作 P 上的不可分辨关系，记作 $\mathrm{IND}(P)$，而且

$$\forall x \in U, [x]_{\mathrm{IND}(P)} = [x]_P = \bigcap_{\forall R \in P} [x]_R$$

这样，$U/\mathrm{IND}(P) = \{ [x]_{\mathrm{IND}(P)} \mid \forall x \in U \}$ 表示与等价关系 $\mathrm{IND}(P)$ 相关的知识，称为知识库 $K=(U,S)$ 中关于论域 U 的 P - 基本知识（P - 基本集）。

$\mathrm{IND}(P)$ 的等价类也成为知识 P 的基本概念或基本范畴。

1.4　集合的上近似和下近似

给定知识库（近似空间）$K=(U,S)$，其中 U 为论域，S 为论域 U 上的等价关系簇，则 $\forall X \subseteq U$ 和论域 U 上的一个等价关系 $R \in \mathrm{IND}(K)$，定义子集 X 关于知识 R 的上近似和下近似分别为：

$$\underline{R}(X) = \{x \mid (\forall x \in U) \wedge ([x]_R \subseteq X)\} = \cup \{Y \mid (\forall Y \in \frac{U}{R}) \wedge (Y \subseteq X)\}$$

$$\bar{R}(X) = \{x \mid (\forall x \in U) \wedge ([x]_R \cap X \neq \varnothing)\} = \cup \{Y \mid (Y \in \frac{U}{R}) \wedge (Y \cap X \neq \varnothing)\}$$

下近似 $\underline{R}$ 或正域 $\mathrm{pos}_R(X)$ 是由那些根据知识 R 判断肯定属于 X 的论域 U 的元素的集合；

上近似 $\bar{R}(X)$ 是由那些根据知识 R 判断肯定属于或是可能属于 X 的论域 U 中的元素的集合。

1.5　知识的约简

知识约简中涉及两个最基本的概念：约简与核。

定义 1 给定一个知识库 $K=(U,S)$ 和知识库中的一个等价关系簇 $P \in S$，$\forall R \in P$，若

$$\mathrm{IND}(P) = \mathrm{IND}(P - \{R\})$$

成立，则称知识 R 为 P 中不必要的，否则 R 为 P 中必要的。

如果对每一个 $R \in P$，R 都是 P 中必要的，则称 P 为独立的，否则称 P 是依赖的或是不独立的。

如果 P 是独立的，$\forall G \in P$，则 G 也是独立的。

定义2（知识的约简）给定一个知识库 $K=(U,S)$ 和知识库上的一簇等价关系 $P \subseteq S$，对于任意的 $G \in P$，若满足以下两条：

（1）G 是独立的

（2）$\mathrm{IND}(G) = \mathrm{IND}(P)$

则称 G 是 P 的一个约简，记作 $G \in \mathrm{RED}(P)$，其中，$\mathrm{RED}(P)$ 表示 P 的全体约简组成的集合。

1.6　知识范畴的约简

定义 3 给定一个知识库 $K=(U,S)$ 和论域 U 上的一个子集簇

$S_{\mathrm{pos}}(U) = F = \{X_1, X_2, \cdots, X_n\}$，$\forall X_i \in F (i = 1,2, \cdots, n)$，如果 $\cap (F - \{X_i\}) = \cap (F)$ 成立，则称范畴（子集）X_i 在 F 中为不必要的，否则为必要的。

定义 4（知识范畴的约简）给定一个知识库 $K=(U,S)$ 和论域 U 上的一个子集簇

$S_{\mathrm{pos}}(U) = F = \{X_1, X_2, \cdots, X_n\}$，$\forall X_i \in F (i = 1,2, \cdots, n)$，

对于任意的 $G \in P$，若 G 满足以下两个条件：

（1）G 是独立的

（2）$\cap G = \cap F$

则称 G 是 F 的一个约简，记作 $G \in \widehat{\mathrm{RED}}(F)$，其中 $G \in \widehat{\mathrm{RED}}(F)$ 是 F 的全体约简组成的集合。

2 基于粗糙集的文本主观性判断模型

所谓主观性文本是指对于客观事实进行描述的文本。其主要内容是基于断言(allegations)或评论(arguments)的,并且带有个人情感和意向的抒发。一个句子,不论是主观句还是客观句,也不管它是简单句还是复杂句,只要句子中包括表达主观意见的成分,那么这个句子都定义为主观句。

基于主观句的这些特点,用分析句子构成方式的方法无疑难以达到区分主观和客观的目的。因为对于某个主观句,只要在保持其主观性质不变的前提下,可以随意改变其句子的组织方式,添加修饰成分。但是主观句的形式无论如何变化,其关键点:"表达主观思想"不会发生变化,例如,"我特别喜欢电影",在保持句子主观性不变的前提下,改变其表达的方式,可以说成是:"无论别人怎么评价,我都喜欢电影"或是"虽然电视剧的剧情发展的比较合理一些,但是我还是喜欢电影",这三个句子,无论是采用何种表达形式,最终要表达的个人的观点都是"我喜欢电影",分词软件处理后的结果是"rr vi rr"。所以,在句子中,可以判断出只要含有"rr vi rr"这种构成模型的,就认为这个句子是主观句。这也就是这篇论文判断主观句的依据,句子中包含主观句构成模型的,那么这个句子是主观句;否则,这个句子就是客观句。

基于以上思想,首先要做的工作就是提取主观句结构模型,即用以判断一个句子是否是主观句的依据。为此,文中从人民日报、网络论坛、BBS 以及淘宝商品评价上面收集出 1000 条主观句子,运用分词软件抽取出这些主观句的句子结构模式。首先对这些模式设定一定的阈值,如果到达阈值,那么就保存这种模式;否则不保留。这样,在保证查准率的前提下,可以先舍去一部分的句子模式。在剩余的模式中,有时候会包含冗余的模式或是冗余的句子成分,例如:模式一:rr vi n 和模式二:rr d vi n,这种情况下,就要考虑用粗糙集理论来约简实验的初始参数了,也就是运用粗糙集的知识范畴的约简,约简多余的句子成分,例如:将模式二:rr d vi rr 中的 d 约简掉,约简结果为:rr vi n。然后再利用粗糙集的属性约简,约简多余的属性,这里就是约简多余的模式,将模式一和模式二约简为同一条模式:rr vi rr。这样,从原来的两条模式,7 个句子成分的比较匹配到最终一个模式、三个句子成分的比较,在很大程度上能提高程序的执行效率。

3 实验分析

3.1 实验数据的选取

对收集的 1000 条主观性文本,进行中文分词处理,提取出其中表达主观性的基本模式,对于每个模式,统计每个模式在统计中出现的比例 θ_n 为:

$$\theta_n = \frac{\text{模式 n 出现的次数}}{\text{查询的句子总数}} \quad (1)$$

当 θ_n 大于等于给定的阈值 η 时,保留模式 n,否则,从模式表中删除该条模式。

阈值的选取会直接关系到实验的最终评估参数:查准率及时间代价。如果阈值取得太大,那么,将过滤大量的模式,影响系统最终的查准率;如果阈值取得过少,那么大量的模式都可以保存下来,系统的查准率相对较高,但是系统的时间复杂度就相对较低了。所以,综合查准率和时间代价,取阈值为 $\eta = 0.048$,经过阈值比较后,保留下来的模式相对于最初提取的模式已经有了很大的约简,但是这种约简实在为了提高系统的时间复杂度的前提下,以牺牲系统的查全和查准率为代价的。此时,可以进一步进行相应的约简,从而达到在保证系统已有的查准率的前提下,提高系统的执行效率。文中采用粗糙集理论的知识范畴约简,约简掉现有模式中冗余的成分,进而采用粗糙集的属性约简,约简掉不必要的模式,约简结果见表 1。

表 1 属性约简后模式表

modelid	p1	p2	p3	p4	p5	p6	p7
1	wt						
2	ww						
3	an						
4	d	a					
5	d	v					
6	ad	v					
7	n	a					
8	vi	a					
9	ng	a					
10	vi	al					
11	rr	vi	n				
12	rr	vi	rr				

实验证明,在保持查准率的前提下,表 1 在执行一百条句子主客观性判断的时间较约减前的执行时间少了 1 秒多,而且没有降低查准率。不难想像,大批量的句子识别,时间效率上无疑会有很大的提高。

3.2 实验结果分析

根据 3.1 约减后的试验模型,现提取中文主客观句子各 100 条,进行中文语句的主客观性识别,由于参考的文献都只是计算了对主观句的查询结果,为了方便比较,文中也列出了主观句的处理结果,识别结果如表 2 所示。

表 2 文中和其他句子主观性判定实验结果对比表

方法	查准率	F 值
Baseline 方法	39.01	51.94
Baseline+基本特征	65.32	64.49
基于最大熵模型的中文观点句主观关系的提取	71.23	68.32
基于粗糙集理论的中文句子主客观性的研究	83.50	84.71

4 结束语

对于句子的划分和成分的提取,文中使用的是计算所汉语词法分析系统 ICTCLAS,使用 ICTCLAS 系统可以直接实现对一个或一段中文文本的词性的划分和提取。

文中在粗糙集理论属性约简的基础上初步研究了中文主客观文本判别的问题。文中,初步判断句子只有主客观之分,没有那种介于主观和客观之间的模糊或是粗糙的概念。粗糙集理论运用于文本,根本作用是用于约简,减少系统的时间代价。通过观察大量的主客观语句,提取出主观语句的结构的特征,根据设定的阈值,选取几率较大的模式,最后用粗糙集约简模式,达到在确保查准率的前提下,提高系统对主客观句子的判定效率的目的。实验表明,基于粗糙集的中文文本主客观性研究无论是在查准率还是在时间代价上,都有了很大的改善。对于那些同存在于主观和客观句中的模型,无论是保留还是约简,都会影响系统在执行时判断主观或是客观的查全率和查准率,后期将继续研究怎么处理这些模型。

参考文献:

[1] 苗夺谦,李道国. 粗糙集理论、算法与应用[M]. 北京:清华大学出版社,2008.

[2] 叶　强,张紫琼,罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法的研究[J]. 信息系统学报,2007,1(1):79-91.

[3] Aas K, Eikvil L. Text categorization: a survey[M]. Norwegian: Norwegian Computer Center, 1999.

[4] Lahtinen T. Automatic indexing : an approach using an index term corpus and combining linguistic and statistical methods: [D]. Finland: University of Helsinki, 2000.

[5] Hochsztain E. A granular approach for analyzing the degree of affability of website[C]//Lecture Notes in Computer Science. [s. l.]:[s. n.],2002:479-485.

[6] 张雪英. 基于粗糙集理论的文本自动分类研究[D]. 南京:南京理工大学,2005.

[7] 樊　娜,蔡皖东,赵　煜. 基于最大熵模型的观点句主观关系提取[J]. 计算机工程, 2010,36(2):4-6.

[8] 杨彩梅. 关系化———一种识别句子主观性语言实现的形式手段[J]. 现代外语,2007,30(1):1-10.

[9] Yang Ying, Liu Xin. A reexamination of text categorization methods [C] / /Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR99). Berkeley, Cal:[s. n.],1999.

[10] 姚天昉,娄德成. 汉语语句主题语义倾向分析方法的研究[J]. 中文信息学报,2007,21(5):73-79.

[11] 李　钝,曹付元,曹元大,等. 基于短语模式的文本情感分类研究[J]. 计算机科学,2008,35(4):132-134.

[12] 吴月萍,陈玉泉. 基于 Web 的概念属性抽取的研究[J]. 中国管理信息化,2009,12(10):98-101.

(上接第 111 页)

表现优秀的虚拟化解决方案。文中通过一个实例介绍了 KVM 在虚拟化过程中的具体应用,其灵活的网络拓扑结构、简便的硬件配置方案、集中统一管理可满足于大多数数据中心虚拟化实践。当然 KVM 也有很多不足,对一些虚拟化扩展特性,如泛虚拟化支持、虚拟机动态迁移、图形化管理界面等新功能正在进一步研究和开发当中。

参考文献:

[1] 何　禹,胡宇鸿,王一波. 虚拟化技术在校园网数据中心的应用[J]. 电子科技大学学报, 2007,16(12):1461-1464.

[2] 周俐军,林泽东,刘伟科. 基于 VMware 的高校数据中心虚拟化管理探究[J]. 中国管理信息化, 2009,12(16):65-66.

[3] 李馥娟. 虚拟机技术在复杂网络实验中的应用[J]. 实验技术与管理, 2009,26(12):79-83.

[4] 王建军. VMware 虚拟机技术在计算机机房管理中的应用[J]. 科技信息, 2009(1):96-97.

[5] 徐　红,刘　羽. 计算机专业虚拟实验教学环境的改革与实践[J]. 实验技术与管理, 2009,26(2):90-92.

[6] 董秋生,黄　文,马骏涛,等. 服务器虚拟化技术在数字图书馆服务器整合中的应用[J]. 情报理论与实践, 2009,32(1):119-122.

[7] 刘荣发. 服务器虚拟化技术在图书馆数字化服务中的应用[J]. 现代图书情报技术, 2007(4):79-83.

[8] 刘爱军,耿国华. 基于 x86 的虚拟机技术现状、应用及展望[J]. 计算机技术与发展, 2007,17(11):250-253.

[9] 董耀祖,周正伟. 基于 X86 架构的系统虚拟机技术与应用[J]. 计算机工程, 2006,32:71-73.

[10] 陈文智,姚　远,杨建华,等. Pcanel/V2——基于 Intel VT-x 的 VMM 架构[J]. 计算机学报, 2009,32:1131-1139.

[11] 顾晓峰,王　健. 基于 Intel VT-x 的 XEN 全虚拟化实现[J]. 计算机技术与发展, 2009,19(9):242-245.

[12] Kivity A, Kamay Y, Laor D, et al. Kvm: the Linux virtual machine monitor[C] // The 2007 Ottawa Linux Symposium. Ottawa, Ontario, Canada: Cisco Unwires Linux Symposium, 2007:225-230.

[13] Russell R. virtio: towards a de-facto standard for virtual I/O devices[J]. SIGOPS Oper. Syst. Rev., 2008,42(5):95-103.