基于粗糙集的Web结构挖掘

周 勇,刘 锋
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:Web站点是由许多Web页面构成的信息系统,随着网络的飞速发展,Web挖掘得到了越来越多的研究。如何从Web中找到与用户查询主题相关的权威页面,是Web结构挖掘的一个重要研究方向。粗糙集理论作为一种有效处理模糊和不确定信息的数学工具,由于其不需要任何先验知识,在数据挖掘领域取得了广泛的应用。文中概述了Web结构挖掘的有关概念,基于粗糙集理论,定义了Web结构挖掘的数据模型,并给出了基于粗糙集的Web结构挖掘的实现流程,分析说明了该方法的性能。

关键词:Web挖掘;Web结构挖掘;粗糙集
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2008)03-0151-03

Web Structure Mining Based on Rough Set Theory

ZHOU Yong, LIU Feng
(School of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: Web site is an information system composed of many Web pages, with the rapid development of the Internet, more and more research works are come out with the Web mining. How to find the authoritative pages interrelated to the themes of users' query is an important research aspect about the Web structure mining. As an effective mathematical tool to deal with vague and uncertain information and having no use for any preliminary information, rough set theory has been widely applied to data mining. The conceptions of Web structure mining are introduced. Based on rough set theory, a data model of Web structure mining is defined. An implementation flow about the Web structure mining based on rough set theory is proposed and the performance is analyzed.

Key words: Web mining; Web structure mining; rough set

0 引 言

数据挖掘是将人工智能技术和数据库技术紧密结合发展出的一门新的技术,利用计算机从庞大的数据中智能地、自动地抽取价值的知识模式,以满足人们不同应用的需要。随着网络的飞速发展及信息量的指数级增长,Web挖掘逐渐成为数据挖掘、人工智能和信息检索领域的一个研究热点。一般来讲,Web挖掘可以分为3类:Web内容挖掘,Web结构挖掘和Web日志挖掘[1,2]。Web结构挖掘作为Web挖掘3大分支之一,越来越受到广大研究人员的重视。所谓Web结构挖掘就是指通过分析不同网页之间的超链接结构,网页内部的可用HTML、XML表示成的树形结构,以及文档URL中的目录路径结构等,发现许多蕴涵在Web内容之外的,对人们有潜在价值的模式和知识的过程。已有的Web结构挖掘算法主要有PageRank和HITS等[3,4]。由于Web结构挖掘具有数据量异常庞大、数据缺乏统一结构、动态信息、面向不同用户群等特点,使得现有方法都存在一定的不足。粗糙集理论作为一种处理不精确和不确定信息的有效工具,由于其不需要数据集合之外的任何先验知识,在数据挖掘领域取得了广泛的应用[5]。文中基于粗糙集理论,提出一种有效的Web结构挖掘方法,针对特定主题抽取Web数据模型中的属性和有效规则,从而发现与用户查询主题相关的重要Web页面。

1 粗糙集的基本概念

下面简要介绍文中所用粗糙集的基本概念[6]。

定义1:一个知识表达系统S定义为$S = < U, C, D, V, f >$,其中U表示对象的集合,即论域;$R = C \cup D$,是属性的集合,其中C表示条件属性集,而D表示决策属性集;$V = \bigcup_{r \in R} V_r$是属性值的集合,即属性的值域集,其中$V_r$是属性$r \in R$的值域;$f$是信息函

收稿日期:2007-06-10
基金项目:国家自然科学基金(60273043);安徽省教育厅自然科学研究项目(KJ2007B153)
作者简介:周 勇(1967-),男,安徽合肥人,硕士,讲师,研究方向为机器学习、Web挖掘。

数，$f: U \times R \to V$，即 $f(x,R) \in V_r$，它指定了 U 中每一对象 x 的属性值。

定义 2：一个决策表定义为 $DT = (U, C \cup D, V, F)$，其中 U 表示对象的集合，即论域；$R = C \cup D$，是属性集合，C 表示条件属性集，D 表示决策属性集；$V = \bigcup_{r \in R} V_r$ 是属性值的集合，$(x,R) \in V_r$，它指定了 U 中每一对象 x 的属性值。

定义 3：令 X 是 U 中根据条件属性 C 可定义的分类，Y 是 U 中根据决策属性 D 定义的分类，对于每个 $X_i, Y_j \in U$，定义一个函数：$D_x: \mathrm{Des}_C(x_i) \to \mathrm{Des}_D(y_i): x_i \cap y_i = \varnothing$，对于 $x_i \in X, y_i \in Y$ 函数 D_x 称为决策表 T 中的决策规则。$\mathrm{Des}_C(x_i)$ 表示基于属性集 C 对象 x_i 的特征描述。

定义 4：P 和 Q 为 U 中的等价关系，Q 的 P 正域记为 $\mathrm{POS}_p(Q)$，即 Q 的 P 正域是 U 中所有根据分类 U/P 的信息可以准确地划分到关系 Q 的等价类中去的对象集合。

定义 5：设 $DT = (U, C \cup D, V, F)$ 为一决策表，$P \in C$，若 $\mathrm{POS}_p(D) = \mathrm{POS}_c(D)$，且 P 相对于 D 独立，则称 P 为 C 的一个约简。注意：约简不一定唯一。

定义 6：核即为所约简的交，它是决策表中最重要的属性组成的集合。

2 基于粗糙集理论的决策分析算法

按照粗糙集理论，决策分析就是给定知识表达系统的条件属性和决策属性，求出最小决策算法，即怎样利用有用特征和有用数据产生决策规则。其具体步骤可分为如下几步：

(1)进行条件属性的简化，即从决策表中消去某些列；

(2)消去重复的行；

(3)消去每一决策规则中的属性的冗余值，即属性值的约简；

(4)输出决策规则。

文中给出一种基于差别矩阵的实现方法。

首先进行属性的约简。定义差别矩阵，给定决策表 $DT = (U, C \cup D, V, F)$，差别矩阵定义为：$M(DT) = (C_{ij})_{n*n}$，$n = \mathrm{card}(U)$，其中 $C_{ij} = \{a \in C \mid a(X_i) \neq a(X_j)\}$ $i,j = 1,2,3,\cdots,n$，否则 $C_{ij} = \varnothing$。

具体算法为：

输入：$DT = (U, C \cup D, V, F)$，

输出：约简 B，$\{B \in C \mid \mathrm{POS}_B(D) = \mathrm{POS}_c(D)\}$。

(1) 构造 $M(DT)$。

(2) 求核 C_o。统计 $M(DT)$ 中每个元素所含属性的个数，找出非零且最小的元素，其对应的 C_{ij} 中的属性即为核属性。

(3) $B \leftarrow C_o$。

(4) $M(DT) = M(DT) - \{C_{ij} \mid C_{ij} \cap B \neq \varnothing\}$，即将差别矩阵中所有与 B 相交不空的属性集赋空。

(5) 判断 $M(DT)$ 是否等于空集，若真，转到(7)，否则转(6)。

(6) 计算每个 $c \in C - B$ 在 $M(DT)$ 中出现的次数，将出现次数最多的属性加入 B，转(4)。

(7) 此时得到的 B 即为属性的一个约简，删除不属于 B 的那些属性列。

对于属性值的约简，同样首先定义决策差别矩阵，给定决策表 $DT = (U, C \cup D, V, F)$，决策差异矩阵定义为：

$M'(DT) = (C_{ij})_{n*n}$，$n = \mathrm{card}(U)$，其中 $C_{ij} = \{a \in C \mid a(X_i) \neq a(X_j)$ 且 $d(X_i) \neq d(X_j)\}$ $i,j = 1,2,3,\cdots,n$，否则 $C_{ij} = \varnothing$。

具体算法为：

(1) 构造 $M'(DT)$。

(2) 求核 C_o。同样地，统计 $M'(DT)$ 中每个元素所含属性的个数，找出第 i 行非零且最小的元素，其对应的 C_{ij} 中的属性即为第 i 条规则的核属性 C_o。

(3) $B \leftarrow C_o$。

(4) $M'(DT) = M'(DT) - \{C_{ij} \mid C_{ij} \cap B \neq \varnothing\}$，其中 i 为定值，即将决策差异矩阵中第 i 行所有与 B 相交不空的属性集赋空。

(5) 判断 $M(DT)$ 的第 i 行是否为空，若真，转到(7)，否则转(6)。

(6) 计算每个 $c \in C - B$ 在 $M'(DT)$ 第 i 行出现的次数，将出现次数最多的属性加入 B，转(4)。

(7) 删除第 i 条规则中不属于 B 的那些属性值。

具体实现时，应用循环语句，依此算法对每一条规则进行处理。

3 基于粗糙集的 Web 结构挖掘方法

所谓 Web 结构挖掘就是指通过分析不同网页之间的超链接结构，网页内部的可以用 HTML、XML 表示成的树形结构，以及文档 URL 中的目录路径结构等，发现许多蕴涵在 Web 内容之外的对人们有潜在价值的模式和知识的过程。目前，Web 用户主要是使用搜索引擎在互联网上检索信息，但目前的搜索引擎往往返回给用户成千上万个检索到的页面，且其中很大一部分是重复的或与用户检索要求不相关的内容。为了解决这些问题，高效率地利用 Web 提供的资源，必须充分利用 Web 独有的一些特点。

Web页与普通文档不同,它所包含的信息由以下3个部分组成:网页正文,网页所含的超文本标记以及网页间的超链接。从广义上讲,Web结构所包含的信息有:

(1)URL字符串中的目录路径结构信息;

(2)网页内部内容的可以用HTML、XML表示成的树形结构;

(3)网页之间的超链接结构。

目前,有很多团体和科研机构对Web结构进行研究,并提出了许多有关Web结构挖掘的算法,如Pitkow对大量的超链接进行了分析和研究[7];Weise用聚类的方法对链接结构进行了分析[8];Spertus通过将链接结构对应成标准关系数据库中的信息,用SQL语句实现对Web的查询[9];Kleinberg通过对Web对应关联矩阵的特征向量计算寻找Authorities页和Hubs页[4];Brin和Page利用页面的inlink和outlink计算Web页的PageRank值,并以此为根据寻找权威页[3];Lempel和Moran则利用马尔可夫链的概念,对Kleinberg的算法进行了改进,淡化了Authorities页和Hubs页之间的关系,提出了一种分析超链接结构的随机算法SALSA。

由于粗糙集理论不需要任何先验知识,且能有效处理不精确和不确定信息,所以利用粗糙集进行Web结构挖掘更简单有效。文中主要考虑针对一个特定的用户查询主题,利用粗糙集算法抽取Web页面及链接点的属性并发现有效规则,从而找到与该主题相关的重要Web页面,以便对Web页面进行分类或排序。下面定义一种数据模型来描述Web页面及其之间的链接关系。

3.1　数据模型

Web可以看作一个有向图 $G=(V,E)$,其中 V 是图中的结点集合,用来表示页面;E 是图中有向边集合,表示页面之间的超链接。结点 V 的入边表示对 V 的引用,出边表示 V 引用其它的页面。页面结点用三元组(PID,P,C)表示。其中PID唯一标记一个页面结点;P 为该Web页的属性集合,$P=\{P_i \mid P_i$ 为属性,$i=1,2,\cdots\}$,包括相对URL、结点类型、页面内容描述、链接点集、其它页面对该结点的引用次数、该结点对其它页面的引用次数、最近某一段时间被访问的次数等;C 为某查询主题中各主题词在该Web页中出现的次数集合,$C=\{C_i \mid C_i$ 为主题词 i 在该Web页中出现的次数$\}$。页面中链接点用三元组(LID,string,TNI)表示,其中LID唯一标记一个链接点,string描述该链接点的信息,TNI是LID指向的页面结点的PID。Web数据模型是一个三元组(PNS,PLS,LS),其中PNS为页面结点集合,PLS为链接点集合,LS为链接集合。

3.2　实现流程

Web结构中的属性除了包括Web页面的属性集本身,还包括一些链接信息。我们的目的就是从这些属性中抽取对分类起决定作用的重要属性,生成分类规则;并根据分类规则查找到符合用户查询主题要求的Web页面,为Web页面分类或排序等后续工作做准备。

(1)对用户查询主题进行分析,并从中抽取主题词 i,$i=1,2,\cdots$;

(2)为每一个Web页面计算主题词 i 在其中出现的次数 C_i,并初始化 $PNS_k.C$ 集;

(3)选取Web数据模型中结点 $PNS_k.P$ 集、$PNS_k.C$ 集、$PNS_k.PLS.string$ 等属性并进行泛化;

(4)生成初始决策表 S:页面结点集PNS为 S 的对象集合,S 的条件属性集为(3)中泛化后的数据,S 的决策属性为该网页是否是反映用户查询主题的重要页面(是或否);

(5)利用粗糙集属性约简算法抽取出重要条件属性,再利用粗糙集规则约简算法对规则进行约简;

(6)根据(5)得到的规则对Web页面进行分析,查找出等于约简规则的条件属性值的页面集合。

通过挖掘Web的结构信息,可以揭示许多蕴涵在Web内容之外的隐含的有用信息,如Web页面的URL可以反映页面的类型,也可以在一定程度上反映页面间在存储位置和内容方面的层次关系,利用粗糙集挖掘与Web页面URL有关的启发式规则,可用于寻找个人主页,或者已经改变了位置的Web页的新位置。

4　结　语

Web结构挖掘作为Web挖掘的一个重要分支,至今还未形成成熟的理论和技术。粗糙集是一种解决多属性决策问题的有效工具,将其应用在Web结构挖掘中以选取重要属性并发现规则,查找出符合用户查询主题要求的权威页面,是非常有效的。Web结构挖掘涉及到大量的计算数据,如何解决海量数据和有限的存储、计算空间之间的矛盾,提高挖掘算法的效率和实时性将是一个值得继续深入研究的问题。

参考文献:

[1] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques[M]. [s.l.]: Morgan Kaufmann Publishers Inc, 2001.

[2] 韩家炜,孟小峰,王　静,等. Web挖掘研究[J]. 计算机研

(下转第206页)

手术类别=7→术后情况=1

手术类别=8→术后情况=1

手术类别=9→术后情况=1

手术类别=12 and 手术环境=3→术后情况=2

手术类别=4 and 手术环境=3 and 病人性别=2→术后情况=2

手术类别=10 and 手术环境=3 and 麻醉情况=5→术后情况=3

手术类别=4 and 手术环境=2 and 麻醉情况=1→术后情况=3

手术类别=5 and 手术环境=3→术后情况=3

3 结 论

将变精度粗糙集理论的数据挖掘技术应用于手术诊断的知识发现,提出属性约简算法和基于求核值属性的归纳值约简算法,在实际应用中,在智能诊断的知识自动获取方面取得新的进展,具有良好应用前景。因为粗糙集基于约简的思想,能将知识高度浓缩,所以最后得到的规则知识相对较少,也更有效,而且由最终约简后的信息表得到的规则知识非常直观。

参考文献:

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

[2] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Science,1982,11(5):314-356.

[3] Ziarko W. Variable Precision Rough Set Model[J]. Journal of Computer and System Sciences,1993,46(1):39-59.

[4] 黎东英,王应明. 基于可变精度粗集理论的规则挖掘模型[J]. 计算机测量与控制, 2005,13(8):833-839.

[5] 米据生,吴伟志,张文修. 基于变精度粗糙集理论的知识约简方法[J]. 系统工程理论与实践,2004(1):76-82.

[6] 代建华,李元香. 一种基于粗糙集的决策系统属性约简算法[J]. 小型微型计算机系统,2003,24(3):523-526.

(上接第 153 页)

究与发展,2001,38(4):405-414.

[3] Brin S,Page L. The Anatomy of a Large-scale Hypertextual Web Search Engine[C]//In:Proceedings of the 7th International World Wide Web Conference. Australia:[s.n.],1998.

[4] Kleinberg J M. Authoriative Sources in a Hyperlinked Environment[J]. Journal of the ACM,1999,46(5):604-632.

[5] Pawlak Z. Rough Set Theory and Its Application to Data Analysis[J]. Cybernetics and Systems,1998,29(9):661-668.

[6] 曾黄麟. 粗集理论及其应用[M]. 重庆:重庆大学出版社,1998.

[7] Pitkow J E. Characterizing World Wide Web Ecologies[D]. Georgia:Georgia Institute of Technology,1997.

[8] Weise R,Veles B. HyPursuit:A Hierarchical Network Search Engine that Exploits Content-link Hypertext Clustering [C]//In Proceedings of the 7th ACM Conference on Hypertext. Bethesda,Maryland, United States:[s.n.],1996.

[9] Spertus E. Parasite:Mining Structural Information on the Web [C]//In:Proceedings of the Sixth International World Wide Web Conference. Santa Clara, California, United States:[s.n.],1997.

(上接第 202 页)

从测试结果可以看出,文中提出的算法对远距离车辆的检测有较好的效果,并且降低了误检率。下一步工作,将要进一步探索和研究多种类型车辆和车辆遮挡的检测以及车辆 ROI 的精确检测。

参考文献:

[1] Sun Zehang, Bebis G. On-road vehicle detection using Gabor filters and support vector machines[C]// International Conference on Digital Signal Processing. Greece:[s.n.], 2002.

[2] Betke M. Multiple Vehicle Detection and Tracking in Hard Real-Time[C]// IEEE Symposium on Intelligent Vehicles. France:[s.n.],2002.

[3] Bensrhair A, Broggi A. Stereo Vision-based Feature Extraction for Vehicle Detection[C]// IEEE Symposium on Intelligent Vehicles. France:[s.n.]2002.

[4] Srinivasa N. A vision-based vehicle detection and tracking method for forward collision warning[C]// IEEE Intelligent Vehicle Symposium. France:[s.n.],2002.

[5] Bensrhair A, Broggi A. A cooperative approach to vision-based vehicle detection[C]// Proceedings IEEE Intelligent Transportation Conference. Oakland,CA,USA:[s.n.],2001: 209-214.

[6] Graefe V,Efenberger W. A Novel Approach for the Detection of Vehicles on Freeways by Real-time Vision[C]//in Procs. IEEE Intelligent Vehicles Symposium '96. Tokyo,Japan:[s.n.],1996:363-368.

[7] Shannon C E. A mathematical theory of communication[J]. Bell System Technical Journal,1948,27(3-4):379-423; 623-656.