粗糙集理论在 Web 信息过滤中的应用研究

The Study on the Application of Rough set Theory in the Web Information Filtering

徐义峰 （衢州学院 浙江衢州 324000）
陈春明 （桂林电子科技大学图书馆 广西桂林 541004）
徐云青 （衢州学院 浙江衢州 324000）

摘要：粗糙集理论是一种适用于不完整和不确定系统的知识发现的数学工具。本文提出了一种利用粗糙集理论生成规则的 Web 信息过滤技术，通过对其中相应算法的改进并进行测试，发现将该方法应用到 Web 信息过滤中是行之有效的，该方法生成的规则少，提高了过滤分析的实时性和实用性。

关键词：信息过滤 粗糙集 属性约简 决策规则

1 引言

防止不良信息在网上传播，保护网络安全，已成为当今网络安全技术中的一大热门课题。Web 内容分析判别过滤是对用户浏览的网页内容进行综合分析判别。基于此项技术可望获得的内容判别准确率更高，又能避免数据库判别方式的弱点，无需经常性地更新数据库。目前对网页内容分析判别过滤的主要问题是在满足一定准确性的条件下如何提高过滤分析的快速性和实用性，这也是网络信息安全领域急待解决的关键技术之一[1]。通过对粗糙集理论的研究，发现粗糙集理论是一种适用于不完整和不确定系统的知识发现的数学工具[2]。故本文提出了一种利用粗糙集理论生成规则的 Web 信息过滤技术。

2 Web 信息过滤系统的粗糙集模型

本文的重点是对 Web 信息利用粗糙集理论进行属性约简，因此我们提出 Web 信息过滤系统[2]的粗糙集模型[3]：

定义 1：Web 页面过滤系统是一个四元组 $S=(U, R=C\cup D, V, F)$。其中，U 是 Web 页面的集合；R 为属性的集合，其中 C 为 Web 特征的集合，D 表示决策属性，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。

3 基于粗糙集约简的 Web 页面过滤技术

根据定义 1，我们给出规则提取的步骤：(1) 建立相关数据集；(2) 数据预处理，并对连续属性进行离散化；(3) 计算条件属性集的约简；(4) 产生分类规则，进行规则的选择和过滤；(5) 模型评价。

3.1 条件属性集和决策属性集的选取及数据预处理

Web 页面被视为一个文档文件，是否需要过滤可以看成一个分类标签（即决策属性 D），文本是否需要过滤作为决策属性集创建决策表，分别将其编码为 1 和 0。以 Web 页面中提取的特征项作为规则条件属性集，其中条件属性集中包含了网页的布局、PICS（Internet 网内容选择平台）等级评定应用、暗示性条文和文档内容四个方面的特征。事实上，在文本的训练阶段，从训练文本中提取的特征子集维数较高，使用 TF－IDF[4]公式 (1) 从所有出现过的单词中提取权值较高的词条作为特征词构成条件属性集。特征词按权值大小进行降序排列，形成高维数据集。

$$w_k(d)=\frac{tf_{ik}(d)\log\left(\frac{N}{n_k}+0.01\right)}{\sqrt{\sum_{k=1}^{n}(tf_{ik}(d))^2\times\log^2\left(\frac{N}{n_k}+0.01\right)}} \quad (1)$$

其中 $tf_{ik}(d)$ 表示词条 t_k 在文档 d 中出现的频率，N 表示全部样本文档的总数，n_k 表示包含词条 tk 的文

档数。此外还需考虑词条的位置信息，例如文章标题、副标题、关键字中的词条要全部保留下来作为特征项，并赋予较高的权重。

3.2 连续属性的离散化处理

根据 TF－IDF 公式得到属性权值是连续性的，需要对其进行离散化处理。针对所选数据集的特点，我们采用布尔逻辑和粗糙集理论相结合的 NS 离散化算法[4]。该算法是一个全局的有监督的离散化算法，将求离散化区间的问题转变成了寻找初始区间的约简的问题。NS 离散化算法是一种有效的离散化方法，可以获得比较理想的离散化效果。

定义 2：初始区间：若 S 中有 m 个连续值属性，对于第 i 个连续值属性记为 $a^i \in C$，a^i 的值域为 Va^i，S 中 a^i 的值的集合是 $a^i(U)=\{a_1^i, a_2^i, a_3^i, \cdots, a_n^i\}$，n 为 a^i 在 U 中值的个数。属性 a^i 的初始区间为 $a_k^i \sim [a_{k+1}^i, a_k^i)$，$0<k<n$。属性 a^i 的初始区间的集合为 $p^i=\bigcup_{i=1}^{n-1} p_k^i$，S 的初始区间的集合 $P(S)=\bigcup_{i=1}^{m} p^i$。

3.3 属性约简

在离散化阶段，我们去除了一些明显的冗余属性，但是还存在大量的不相关或冗余属性。为了最大程度的减少这样的属性，本文使用改进的启发式属性约简算法，以求出最优或近似最优的属性约简。

启发式属性约简算法中，核（CORE）中的特征必须包含在最优或近似最优的属性约简中。计算属性的核的方法可以通过对条件属性集合逐个减去某个属性，计算它对决策属性的正区域是否改变的方法来确定该属性是否是不可缺少的，所有不可缺少的属性构成了核。也可以利用差别矩阵来求核，核是差别矩阵中所有单个元素组成的集合。

寻找近似最优约简 R 的启发式算法过程如下：将 CORE 作为初始化特征子集，通过特征选取标准：

（1）选取特征 a，如果将其加入特征子集 R，使得正域的基 card（POSR∪{a}（D））增加，并且 Max—size（POSR（D）/IND（{R，D}））比加入其他特征时要大（POSR（D）/IND（{R，D}）表示 D 的 R－正域在等价关系 IND（{R，D}）下的划分，在这个划分中，设有某个集合 POSR（D）具有最多的元素个数，记为 Max—size（POSR（D）/IND（{R，D}））；

（2）如果通过标准（1）有两个特征获得同样的值，则选取具有较少特征值的特征。将特征从可省略特征空间中逐个选出加入特征子集，直到最优约简产生。

算法描述：设 R 为所求的特征子集，C 为条件属性集，$R \cup P = C$，D 为决策属性，U 为所有记录的集合，X 为矛盾记录，E 为精确限度。开始令 R＝CORE（c），P＝C－CORE（C），k＝0。

（1）除去所有相容记录：X＝U－ POSR（D）；

（2）if $k \geq E$ where

$$k=\frac{card(POS_R(D))}{card(U)}, \text{then STOP}$$

else if $POS_R(D) = = POS_C(D)$ then STOP

（3）对任一个 $p \in P$，计算：

$v_p = card(POS_{R \cup \{p\}}(D))$

$m_p = Max\text{—}size(POS_{R \cup \{p\}}(D))/IND(\{R, D\})/(R \cup \{p\} \cup D)$

（4）选出使得 $v_p \times m_p$ 最大的特征 p，并且 R＝R∪{p}，P＝P－{p}

（5）跳回第（1）步。

另外，该算法还必须具有增量学习的能力，即当有新的记录加入时，能够实时的更新属性约简和规则。为了满足这一要求，我们加入增量学习的算法：

对新来的记录 u，如果原规则集中的任何一条规则都不能满足规则前件与由记录 u 生成的规则前件相同，则需对 u 求取其最简规则，并将其加入到原规则集中，形成新的规则集；如果原规则集中的规则满足新记录 u，则不需改变原规则集；以上两种情况，在加入新纪录时都不会导致冲突记录。如果加入新的记录 u 产生冲突记录，我们采用高置信度优先法，选取较高置信度的规则，去掉另外与其产生冲突的规则；如果同时具有相同的置信度，则采用多数优先原则，选取覆盖样本数较多的规则，从而避免了规则冲突。

3.4 值约简

值约简是在属性约简的基础上对决策表的进一步简化。本文提出一种基于粗糙集的启发式值约简算法。分析最小值约简，仍然从值核入手。首先得到决策表的值核，再分析值核外的属性。按下面的算法进行取舍：

（1）对决策表中的条件属性进行逐列考察，如果除去该列后，产生冲突记录，则保留该记录的原属性值；如果没有产生冲突记录但有重复记录，则将该记录

的属性值标记为“*”；对于其他记录，将该属性值标记为“?”。

（2）删除产生表中的重复记录，并考察每条含有属性值为“?”的记录，如果由保留原值的所有属性值可以判断出决策，则将“?”标志改为“*”，否则将“?”改为原值。若某条记录的所有属性均被标记，则将标记“?”修改为原属性值。

（3）删除所有条件属性均被标记为“*”的记录和可能产生的重复记录。

（4）如果产生表中某条规则对的集合是另一条规则对集合的子集，且两条记录决策属性相同，则删除另一条规则。

通过以上四个步骤，规则表中的所有记录为原信息表的值约简结果。规则表中的每一行代表一条决策规则，特征项被当作决策规则的前件，文件所属的类别当作规则的后件，即规则的决策。

4 实验评估

以 UCI 机器学习数据库中的 Iris 数据集为例，使用改进的启发式属性约简算法，并通过值约简，导出决策规则。该数据集共有 150 条记录，4 个条件属性（sepallength，sepalwidth，petallength，petalwidth），分为三个决策类（Iris－setosa，Iris－versicolor，Iris－virginica），三个决策属性值分别为 0、1、2。其中每条记录包含四个条件属性和一个决策属性。首先将数据进行离散化处理，然后使用启发式的属性约简算法，得到 10 条规则，如下所示（注：后面括号中的值分别表示支持度，置信度，同时包含前后件的记录数，包含前件的记录数）：

Rule1：PetalLength = 0 = > Class = 0 （33.3%，100%，50，50）

Rule2：Sepalwidth = 2 and Petallength = 1 = > Class = 1 （5.3%，100%，8，8）

Rule3：PetalLength = 1 and Petalwidth = 0 = > Class = 1 （31%，100%，47，47）

Rule4：PetalLength = 2 and Petalwidth = 1 = > Class = 1 （0.6%，100%，1，1）

Rule5：Sepalwidth = 0 and Petallength = 2 = > Class = 2 （0.6%，100%，1，1）

Rule6：Sepallength = land Sepalwidth = 0 = > Class = 1 （0.6%，100%，1，1）

Rule7：Sepallength = land Petalwidth = 2 = > Class = 2 （25%，100%，38，38）

Rule8：Sepallength = 0 and Petalwidth = 0 and Sepalwidth = 1 and Petallength = 2 = > Class = 1 （0.6%，100%，1，1）

Rule9：Sepalwidth = land Petalwidth = 2 = > Class = 2 （19%，100%，28，28）

Rule10：Sepallength = land Petalwidth = 2 and Petalwidth = 0 = > Class = 2 （2%，100%，3，3）

如果使用 ROSETTA 软件，得到 13 条决策规则，可见，本文提出的基于粗糙集理论的启发式属性约简算法可以得到较少的决策规则。测试时，当有多条规则适用于同一测试记录时，选取置信度高的规则；当置信度相同时，选择支持度高的规则；当支持度仍然相同时，选取首先出现的规则。由于本文选取的数据集中不存在冲突记录，即在所有条件属性全部相同的情况下，不存在不同的决策属性，因此我们得到的规则置信度为 100%。

5 结束语

以上的实验证明了新的启发式属性约简算法的有效性，但是由于样本集的数量有限，本文所能反应的测试结果也许不是非常客观。不过作为一种新的属性约简算法，它确实大大提高了文本过滤的速度，和其他算法比较，准确性也没有下降。因此，将该算法应用到 Web 信息过滤中，能够有效地提高过滤速度，从而实现实时性。

参考文献

1 孟庆春、王汉平，一种基于粗糙集的文本分类规则抽取方法[J]，青岛海洋大学学报，2003(11).

2 Pawlak Z. A rough set view on Bayes´ theorem[J], International Journal of Intelligent Systems, 2003, 18(5):487－498.

3 Tay F, et al. Fault diagnosis based on rough set theory[J], Engineering Applications of Artificial Intelligence, 2003, 16(1):39－43.

4 Mi J S, et al. Approaches to knowledge reduction based on variable precision rough set models [J], Information Sciences, 2004, 159(3):255－272.