

基于模糊粗糙集的 Web 文本分类

孙海虹, 丁华福

(哈尔滨理工大学 计算机科学与技术系, 黑龙江 哈尔滨 150080)

摘要:网络信息的多样性和多变性给信息的管理和过滤带来极大困难,为加快网络信息的分类速度和分类精度,提出了一种基于模糊粗糙集的 Web 文本分类方法。采用机器学习的方法:在训练阶段,首先对 Web 文本信息预处理,用向量空间模型表示文本,生成初始特征属性空间,并进行权值计算;然后用模糊粗糙集算法来进行信息过滤,用基于模糊粗糙集的属性约简算法生成分类规则;最后利用知识库进行文档分类。在测试阶段,对未经预处理的文本直接进行关键属性匹配,经模糊粗糙因子加权后,用空间距离法分类。通过试验比较,该方法具有较好的分类效果。

关键词:机器学习;模糊粗糙集;Web 文本分类;属性约简

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2010)07-0021-04

Web Document Classification Based on Fuzzy - Rough Set

SUN Hai-hong, DING Hua-fu

(Department of Computer Science and Technology of Harbin University of Science and Technology, Harbin 150080, China)

Abstract: The diversity and variability of network information brings great difficulty to information management and information filtering. Put forward a method to Web document classification based on fuzzy - rough set in order to improve the speed and accuracy of network information classification and use machine learning method for training and testing Web document. In the training process, firstly, representing preprocessed Web documents by vector space model, forming initial attribution features space and conducting weight value computing. Then, conducting information filtering and reducing attribution feature space by fuzzy - rough set algorithm, forming classification rules. Finally, classifying documents by multiple knowledge bases. In the testing process, matching key attributes directly and computing weight value by the fuzzy - rough factor, then classifying document by space distance method. The experiment results and the comparison with others show that this Web document classification has good classification performance.

Key words: machine learning; fuzzy - rough set; web document classification; attribution reduction

0 引言

随着信息技术和 Internet 的迅速发展,网络信息的多样性和多变性导致信息迅速膨胀,信息检索、内容管理及信息过滤等变得越来越重要和困难^[1]。一方面,互联网和企业信息系统每天都不断产生大量文本数据,这些文本资源中蕴含着许多有用信息;另一方面由于技术手段的落后,用户从 Web 上海量、动态、异构的丰富信息资源中快速、有效地查找自己感兴趣的信息从而获取潜在的有价值的知识十分困难。因此,人们迫切需要研究出有效的方法和手段从大规模文本信息资源中提取符合需要的简洁、精炼、可理解的知

识^[2]。Web 文本分类是 Web 文本挖掘的重要技术和内容,是信息过滤领域的关键问题,它是指按照预先定义的主题类别,为 Web 文档集中的每个文档确定一个类别^[3]。这样,用户就不仅能够方便地浏览 Web 文档,而且可以通过限定搜索范围来使文档的查找更为容易,从而更好地帮助用户把握文档信息,集中精力处理用户所感兴趣的文档。然而,面对大量的 Web 文档,如何有效地组织这些文档,将 Internet 这个浩大的分布式信息空间的无序状态有序化,已成为热点问题。

文中提出了一种基于模糊粗糙集(Fuzzy - rough set)理论的 Web 文本分类方法。采用机器学习的方法:在训练阶段,首先对 Web 文本信息预处理,用向量空间模型表示文本,生成初始特征属性空间,并进行权值计算;然后用模糊粗糙集算法来进行信息过滤,用基于模糊粗糙集的属性约简算法生成分类规则;最后利用知识库进行文档分类。在测试阶段,对未经预处理

收稿日期:2009-09-14;修回日期:2009-12-17

基金项目:国家自然科学基金重点项目(60736014)

作者简介:孙海虹(1984-),女,硕士,研究方向为机器学习、文本分类;丁华福,研究员,研究方向为机器学习。

的文本直接进行关键属性匹配,经模糊粗糙因子加权后,用最小空间距离法分类。通过试验比较,该方法具有较好的分类效果,对于收集到的海量文本能够迅速区别文本类型的最小属性集,大大降低了关键词向量空间的维数;缩小了问题的规模;简化了分类过程。

1 Web 文本表示与特征属性空间生成

1.1 向量空间模型

文中采用向量空间模型表示预处理后的 Web 文档。假设文章中词条出现的先后次序是无关紧要的,每个特征词对应特征空间的一维,将文本表示成欧氏空间的一个向量。

其核心概念可以描述如下:

特征项:组成文档的字、词、句子等。Document = $D(t_1, t_2, \dots, t_k, \dots, t_n)$, 其中 t_k 表示第 k 个特征项,作为一个维度。

特征项的权重:在一个文本中,每个特征项都被赋予一个权重,以表示特征项在该文本中的重要程度。

向量空间模型(VSM, Vector Space Model):在舍弃了各个特征项之间的顺序信息之后,一个文本就表示成向量,即特征空间的一个点。如文本 d_i 表示为:

$$V(d_i) = (W_{i1}, W_{i2}, \dots, W_{ik}, \dots, W_{im})$$

其中, $W_{ik} = f(t_k, c_j)$ 为权值函数,反映特征 t_k 决定文档是否属于类 c_j 的重要性^[4]。

1.2 特征属性空间生成与权值计算

引入权值计算的目的是提高查询的查全率与查准率。自从提出向量空间模型以来,出现了很多种权值计算函数。采用如下的公式计算权值:

$$W(t, d) = \frac{tf(t, d) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{i \in d} [tf(t, d) \times \log(N/n_t + 0.01)]^2}} \quad (1)$$

其中 $W(t, d)$ 表示词 t 在文档 d 中的权重, $tf(t, d)$ 表示词 t 在文档 d 中的词频, N 为训练文本的总数, n 为训练集中出现 t 的文档数目,分母为归一化因子。

2 模糊粗糙集的基本概念

粗糙集理论和模糊理论是处理两种不确定性(粗糙性和模糊性)的不同的数学方法,二者是互为补充的,而不是相互排斥的^[5]。模糊粗糙集是将模糊集与粗糙集结合使用的理论,是一种处理不精确或模糊知识的数学工具。最近几年,人们将它应用于机器学习、模式识别、数据挖掘等很多领域,取得了一定的进展,并且还有很大的研究的价值和空间。本节讨论一下模糊粗糙集的基本概念。

2.1 信息表达与知识分类观点

信息系统 S 通常可以表示为:

$$S = (U, A, F)$$

其中 U 为对象集,即 $U = \{x_1, x_2, \dots, x_n\}$, U 中的每个元素 $x_i (i \leq n)$ 称为一个对象,若 $X_i (i \leq k)$ 为 U 的子集,且 $X_i \neq \emptyset (i \leq k)$, $X_i \cap X_j = \emptyset (i \neq j)$, $\bigcup_{i=1}^k X_i = U$, 则称 $\{X_i | i \leq k\}$ 为 U 的划分。 A 为属性集,即 $A = \{a_1, a_2, \dots, a_m\}$, A 中的每个元素 $a_l (l \leq m)$ 称为一个属性,分为条件属性与决策属性两类,决策属性用 d 表示,此时信息系统表示为 $S = (U, A, F, d)$ 。 F 为 U 与 A 之间的关系集,即 $F = \{f_l: U \rightarrow V_l (l \leq m)\}$, 其中 V_l 为 $a_l (l \leq m)$ 的值域。

设 R 是集合 U 上的二元关系,如果它是自反、对称和传递的,则它是 U 上的等价关系。 U 上的等价关系 R 必然产生 U 上的一个划分, U 上的一个划分由 U 上的一个等价关系 R 产生, U 上的等价关系与 U 上的划分一一对应。设 $P \subseteq R$, 且 $P \not\subseteq \emptyset$, P 中所有等价关系的交集 $\cap P$ 也是一个等价关系,称为不可分辨关系,记为 $\text{ind}(P)$ 。在实际应用中,通常可用等价关系代替分类^[3]。

2.2 粗糙的成员关系与边界观点

粗糙集延拓了经典集合论,把用于分类的知识嵌入集合内,作为集合组成的一部分^[6]。一个对象是否属于集合 X ,可分为三种情况:对象 x 肯定属于集合 X ;对象 x 肯定不属于集合 X ;对象 x 可能属于集合 X ,也可能不属于集合 X 。粗糙集认为知识的粒度性造成用已有知识不能精确表示某些概念和造成对象之间的不可分辨^[7],这就产生了所谓的关于不精确的“边界”观点。

定义 1: 设 $X \subseteq U$ 且 $x \in U$, 则 x 属于集合 X 的粗糙隶属函数为:

$$\mu_x^R(x) = \frac{\text{card}(X \cap [x]_R)}{\text{card}([x]_R)} \quad (2)$$

$\text{card}(\cdot)$ 表示集合的基数^[8]。可以看出,隶属函数 $\mu_x^R(x) \in [0, 1]$, 这里的隶属关系是根据已有的分类知识客观计算出来的,可以解释为一种条件概率,它能够从全域上的个体加以计算,而不是主观给定的。

定义 2: 设 U 是对象集, R 是 U 上的等价关系。

(1) 称 (U, R) 为近似空间,由 (U, R) 产生的等价类为 $U/R = \{[x_i]_R | x_i \in U\}$, 其中 $[x_i]_R = \{x_j | (x_i, x_j) \in R\}$ 。

(2) 对于任意 $X \subseteq U$, 记 $\underline{R}(X) = \{x_i | [x_i]_R \subseteq X\}$, $\bar{R}(X) = \{x_i | [x_i]_R \cap X \neq \emptyset\}$, 称 $\underline{R}(X)$ 为 X 的下近似, $\bar{R}(X)$ 为 X 的上近似。

(3) 若 $R(X) = \bar{R}(X)$, 称 X 为可定义的集合, 否则称 X 为粗糙集。

定义 3: 设 (U, R) 为近似空间, 对于 $X \subseteq U$, 称 $BN(X) = \bar{R}(X) - R(X)$ 为 X 的边界。

2.3 信息系统的属性约简

定义 4: 设 (U, A, F) 是一个信息系统, 对于 $B \subseteq A$, 若 $R_B = R_A$, 称 B 是划分协调集。若 B 是划分协调集, 而 B 的任何真子集均不是划分协调集, 称 B 为划分约简集。

定义 5: 设 (U, A, F) 是信息系统, $B_k (k \leq r)$ 为所有划分约简集, 记

$$C = \bigcap_{k=1}^r B_k \quad K = \bigcup_{k=1}^r B_k - C \quad I = A - \bigcup_{k=1}^r B_k$$

$a \in C$ 时称 a 为划分核心, C 为划分核心集; $a \in K$ 时称 a 为划分相对必要属性, K 称为划分相对必要属性集; $a \in I$ 时称 a 为划分不必要属性, I 称为划分不必要属性集。根据属性的约简, 可以将一个知识体系中不相关的信息过滤掉, 减少知识体系中的噪声, 有助于提高分类的效率和准确率^[9]。

2.4 模糊粗糙集

模糊粗糙集的主要思想是当等价关系使模糊集合的论域变得粗糙时, 定义此模糊集合的相应上近似和下近似; 或者把等价关系弱化为模糊相似关系, 从而得到一个更具表达力的粗糙模型。

定义 6: 设 (U, R) 是 Pawlak 近似空间, 即 R 是论域 U 上的一个等价关系。当等价类的元素所属的类别不明确时, 等价类便表示为模糊集的形式 $F = \{F_1, F_2, \dots, F_H\}$, $F_j \quad j \in \{1, 2, \dots, H\}$ 是模糊集。给定 R 上的一个模糊划分 θ , 利用上近似 $\bar{\theta}$ 和下近似 $\underline{\theta}$ 的形式表达任一模糊集合 F , 称 $\bar{\theta}(F)$ 和 $\underline{\theta}(F)$ 为模糊粗糙集。定义模糊上下近似如下^[10]: 若 A 是 U 上的一个模糊集合, 则 A 关于 (U, R) 的一对上下近似定义为 U 上的一对模糊集合, 其隶属函数分别定义为:

$$\mu_{\bar{\theta}}(F_i) = \sup_{x \in \theta} \mu_{F_i}(x) * \mu_F(x) \quad \forall x \quad (3)$$

$$\mu_{\underline{\theta}}(F_i) = \inf_{x \in \theta} \mu_{F_i}(x) * \mu_F(x) \quad \forall x \quad (4)$$

这里“*”表示一种运算, 如果 * 取 min 操作, 则上式表示了模糊事件 F 的可能性和必然性程度。 $\mu_{\bar{\theta}}(F_i)$ 是 F 中 F_i 的可能性隶属度, $\mu_{\underline{\theta}}(F_i)$ 是 F 中 F_i 的必然性隶属度^[11]。

3 基于模糊粗糙集的文本分类算法

3.1 训练算法

Step1: 对训练集的 Web 文本进行预处理和文本表示。

Step2: 对训练集进行特征降维、权值计算, 生成目

标特征向量集。

Step3: 将目标特征向量集转化为 $S = (U, A, F, d)$, 明确条件属性 A 与决策属性 d 。

Step4: 属性约简。计算属性在文本中的可能隶属度 $\mu_{\bar{\theta}}(F_i)$ 和必然隶属度 $\mu_{\underline{\theta}}(F_i)$, 以此确定划分核心、相对必要属性和不必要属性, 生成约简属性集。

Step5: 发现分类规则, 生成分类规则库, 利用规则库进行分类。

3.2 分类算法

算法: 模糊粗糙集文本分类算法;

输入: 待分类文本 t ;

输出: 待分类文本类属 x ;

声明: N 为训练集文档数,

C 为类别数,

Distance[N] 为待分类文本与训练集文档的欧式距离。

函数体:

begin

set $i = 1$;

while $i < N$ do

begin

computer Distance[i];

$i++$;

end;

sort(Distance[i]);

build Neighbor-space W ;

for $c = 1$ to C do

begin

$$\tau_c(t) = \frac{1}{|W|} \sum_{y \in W} \tilde{\mu}_t(y) \mu_c(y)$$

end

$$x = \arg \max_c (\tau_c(t));$$

end;

算法中 $\tau_c(t) = \frac{1}{|W|} \sum_{y \in W} \tilde{\mu}_t(y) \mu_c(y)$ 为模糊粗糙隶属度, 其中 $\tilde{\mu}_t(y)$ 为模糊隶属度因子, 而 $\mu_c(y)$ 为粗糙因子。

由于分词算法和特征降维算法需要占据很大的系统资源和时间, 对测试文本不进行分词、特征降维处理, 只对训练文本集进行分词、降维, 根据训练文本集选出的目标特征项集 T , 采用查找匹配法来确定测试文本的特征项, 旨在提高了对待分类文本分类的速度。经过处理后, 期望达到规则中的条件属性数目、规则冲突被尽可能减少, 规则的适应性更强的效果。最后结合传统的单一的分类方法对该算法进行比较, 验证该算法的可行性。

4 实验结果分析

实验文档来源于门户网站 Yahoo(<http://www.yahoo.com>)。实验中使用了机器学习的方法,只考虑英文网页,选取了 20000 个 Web 文本,其中 3/4 作为训练集,1/4 为测试集。

实验结果如表 1 所示。实验结果表明,将该方法可以有效地提高 Web 文档分类的效率和精确度,对比 K 近邻分类算法和模糊 K 近邻分类算法具有明显的优越性。

表 1 四类文本分类算法精确度比较

	K 近邻	模糊 K 近邻	粗糙集	模糊粗糙集
环保	0.711	0.778	0.801	0.821
教育	0.760	0.773	0.799	0.845
军事	0.770	0.758	0.792	0.838
科技	0.921	0.899	0.913	0.931
司法	0.923	0.946	0.906	0.865
体育	0.883	0.857	0.896	0.864
文化	0.977	0.977	0.955	0.968
卫生	0.701	0.767	0.774	0.911
政治	0.718	0.799	0.768	0.903
经济	0.806	0.853	0.881	0.917
平均精确度	0.8170	0.8459	0.8485	0.9038

5 结束语

在研究了模糊粗糙集理论的基础上,提出了一种基于模糊粗糙集理论的 Web 文本分类方法。实验结果表明,该方法大大降低了特征向量的维数,有效地提高了分类的准确率,对比传统分类算法有很大的优势。然而,随着 Internet 信息的不断增长,如何更为有效地提高分类准确率和压缩高维的文本空间是我们面临的

巨大挑战,也是下一步研究的重点。

参考文献:

- [1] Craven D, DiPasquo D, Freitag, et al. Learning to Construct Knowledge Bases from the World Wide Web[J]. Artificial Intelligence, 2000, 118(1-2): 69-113.
- [2] 王涛, 孙河山. Web 挖掘技术在搜索引擎中的应用[J]. 信息系统, 2002, 25(4): 296-299.
- [3] 尹世群, 邱玉辉. Web 文本分类关键技术研究[D]. 重庆: 西南大学, 2008: 15-17.
- [4] 曹勇, 吴顺祥. 中文 Web 文本分类研究[D]. 厦门: 厦门大学, 2007: 15-17.
- [5] 陈水利, 李敬功, 王向公. 模糊集理论及其应用[M]. 北京: 科学出版社, 2005.
- [6] Mollestad T, Skowron A. A rough set framework for data mining of propositional default rules[C]// ISMIS-96: Ninth International Symposium on Methodologies for Intelligent Systems. Berlin: Springer-Verlag, 1996: 448-457.
- [7] Yao Y Y. A decision theoretic framework for approximating concepts[J]. International Journal of Man-Machine Studies, 1992, 37: 793-809.
- [8] Hu Xiaohua. Knowledge discovery in databases: an attribute-oriented rough set approach[D]. Regina, Saskatchewan: University of Regina, 1995.
- [9] 张文修, 仇国芳. 基于粗糙集的不确定决策[M]. 北京: 清华大学出版社, 2005.
- [10] Dubois D, Prade H. Putting rough sets and fuzzy sets together [C]// Intelligent Decision Support: Handbook of Applications and Advances of the Rough Set Theory. Boston: Slowinski R ED, Kluwer Academic Publishers, 1992: 203-232.
- [11] 谢克明, 杨静. 粗糙集理论及其在智能控制领域的应用前景[J]. 太原理工大学学报, 1999, 30(4): 338-342.

(上接第 20 页)

子工业出版社, 2007.

- [3] 陈玉芳, 葛隧和. 一个基于 XML 的 WEB 数据收集模型的研究[J]. 计算机工程与应用, 2004, 40(10): 150-152.
- [4] Zhai Y H, Liu B. Structured data extraction from the web based on partial tree alignment[J]. IEEE Transactions on knowledge and Data Engineering, 2006, 18(12): 1614-1628.
- [5] 高丙坤, 成战刚, 李倩. 基于正则表达式的信息滤除算法[J]. 现代计算机, 2008(2): 54-55.
- [6] 胡燕. 基于 Web 信息抽取的专业知识获取方法研究[D]. 武汉: 武汉理工大学, 2007.
- [7] 李嘉佑, 贾自艳, 何清, 等. 基于 Web 挖掘的网页清洗技术[J]. 计算机工程与技术, 2006, 25: 98-101.
- [8] Bowers M. 精通 CSS 与 HTML 设计模式[M]. 刘申, 朱瑜敏, 鲁奇, 译. 北京: 人民邮电出版社, 2008.
- [9] 卢亮, 张博文. 搜索引擎原理实践与应用[M]. 北京: 电子工业出版社, 2007.
- [10] 张海波. 面向主题的网页过滤机制研究[D]. 兰州: 兰州大学, 2007.
- [11] 陈涛. 基于网页关联特征的互联网图像自动标注系统[D]. 杭州: 浙江大学, 2007.
- [12] 李蕾, 王劲松, 白鹤, 等. 基于 FFT 的网页正文提取算法研究与实现[J]. 计算机工程与应用, 2007, 43(30): 148-151.
- [13] 蒲强, 李鑫, 刘启和, 等. 一种 Web 主题文本通用提取方法[J]. 计算机应用, 2007, 27(6): 1394-1395.
- [14] 吴鹏飞, 孟祥增, 刘俊晓, 等. 网页区域分割与识别技术[J]. 现代计算机(专业版), 2006(6): 48-50.
- [15] 朱征宇, 任翔, 苑昆峰, 等. 基于 HTML 语义分析的网页正文提取[J]. 计算机应用研究, 2008, 25(增刊): 177-178.