综述与评论 Review and Comment

负关联规则在 Web 文档分类中的研究*

石芙芙，董祥军，陈修宽
（山东轻工业学院 信息科学与技术学院，山东 济南 250353）

摘 要：对 Web 文档进行分类可以较好地解决网上信息杂乱的现象，介绍了 Web 文档分类的相关知识以及关键技术，并对目前的分类方法进行了总结，对 Web 文档分类中关联规则挖掘研究现状和主要技术进行了论述，指出了负关联规则在 Web 文档分类中的发展趋势。

关键词：数据挖掘；Web 文档分类技术；负关联规则
中图分类号：TP311 **文献标识码**：A

Research of negative association rules in Web documents classification

SHI Fu Fu, DONG Xiang Jun, CHEN Xiu Kuan
(School of Information Science and Technology, Shandong Institute of Light Industry, Jinan 250353, China)

Abstract: Web documents classification technology can be a better way to solve information disorderly phenomenon online. This paper introduces the relevant knowledge and key technology of Web documents classification, summarizes the main methods of it and expatiates the present situation and main technology of association rules in Web documents classification, points out the development tendency of negative association rules in Web documents classification.

Key words: data mining; Web classification method; negative association rules

随着科学技术的发展，Internet 正在以令人难以置信的速度迅速发展，使得万维网成为一个拥有大量资源的数据库。这个蕴含着丰富资源的信息空间为数据挖掘研究提出了新的挑战，存放的这些信息绝大多数都是以文本的形式存在的，如何在浩瀚的文本信息中挖掘到潜在的知识是一个有待解决的问题。为了帮助人们有效地组织和管理海量的 Web 信息，Web 文档分类技术应运而生。

1 Web 文档分类的概念及分类方法

1.1 Web 文档分类的概念

Web 挖掘[1]是指从大量 Web 文档的集合 C 中发现隐含的模式 P。如果将 C 看作输入，将 P 看作输出，那么 Web 挖掘的过程就是从输入到输出的一个映射 $\zeta: C \rightarrow P$。一般地，Web 挖掘可以简单的分为三类[2]：Web 内容挖掘(Web content mining)，Web 结构挖掘(Web structure mining)和 Web 使用记录的挖掘(Web usage mining)。

文档分类是 Web 内容挖掘中一项非常重要的任务，它是根据页面的不同特征，将其划分为事先建立起来的不同类，属于有指导的机器学习问题。其目的是让机器学会一个分类函数或分类模型，该模型能把 Web 文本映射到已存在的多个类别中的某一类，使检索或查询的速度更快，准确率更高。这样，用户在浏览 Web 文档时，就不会因为纵横交错的超级链接而“迷路”。

Web 文本分类是一个典型的有教师的机器学习问题，一般的可分为数据预处理、构选分类器和文档分类 3 个阶段。数据挖掘主要是针对结构数据库，如数据仓库中的数据，然而，在网络中大部分的信息是存储在文本数据库当中，即所谓的半结构化数据 (semi-structure data)，它既不是完全无结构的，也不是完全结构的，因此需要对文本进行预处理，抽取代表其特征的元数据，这些特征可以用结构化的形式保存。

训练算法的工作是对训练文档集合中每篇文本对应的词表进行统计，计算出类别向量矩阵，同时进行归一化，最后保存训练得到的向量表，即得到了分类知识

* 基金项目：山东省自然科学基金(Y2007G25)；山东省优秀中青年科学家奖励基金项目(2006BS01017)

库。分类算法(也可称为识别算法)则依据训练得到的分类知识库，并用一定的算法对待分类文本进行分类。Web 文档分类的步骤如图 1 所示。

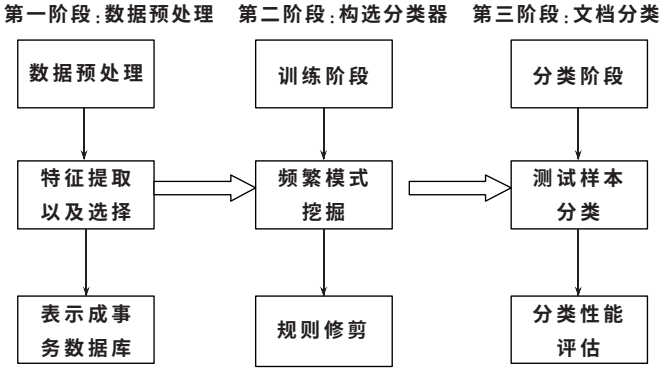

图 1 Web 文档分类的步骤

1.2 Web 文档分类方法

目前关于 Web 文档分类的研究已经取得了很大的进展，并提出了一系列的分类法。常见的文档分类方法可以分为三类[3]：一类是基于 TFIDF 方法，包括 TFIDF 算法、K 近邻算法等；另一类则是基于统计学分类方法，如贝叶斯算法、最大熵算法等；第三类是基于知识学习的方法，如决策树(Decision Tree)C4.5 等算法。

近年来，Web 文档分类已成为众多领域研究者的热门研究课题，研究者们从不同的角度把越来越多的知识引入该领域。

向量空间模型 VSM(Vector Space Model)将所有文本都表示成具有某种相同结构的数据，它是 Salton 等人于上世纪 60 年代末首先提出的，最早应用于信息检索领域。而 Joachims 最早将 SVM 方法用于文本分类[4]。饶文碧等人[5]提出了一种扩展的基于 VSM 的 Web 文本分类，在传统的训练和分类算法的基础上，加入了一个反馈判定的算法，这种方式更加贴近真正意义上的机器学习，使得该算法具有一定程度上的认知自主性和不断学习的能力。由于 Web 文档往往具有不确定的特征，使得利用模糊集合理论对信息检索过程的不确定性建立模型成为可能，雷景生在参考文献[6]中提出了一种基于模糊相关技术的 Web 文档分类方法。张志强、郑家恒提出基于加权类轴的方法[7]，利用了 Web 文本的标记信息在文档中的重要程度不同，使用了 Web 标记的三级加权处理，使向量的维数有所下降。胡和平、易高翔提出了基于容错粗糙集的方法[8]，利用容错关系来表示文档，利用特征词协同出现的价值，丰富特征词对 Web 文档的描述。

2 关联规则在 Web 文档分类中的研究

2.1 文本预处理

由于 Web 上的内容大都是半结构化的，通常要进行结构化的表示。表示方法为：一篇文档有类标签 C，词条为 $T(T=\{t_1,t_2,\cdots,t_m\})$，用事务 $D=\{C,t_1,t_2,\cdots,t_m\}$ 来表示该文档模型，则 Web 文档集可表示成事务集，其项集由文档词条和文档所属类别组成。

在数据清理阶段，采用移除停用词和计算 TF/IDF 的方式对文本文档进行清理。需要移除对构造关联分类器价值不大的词条，形成清洁文档，包括一些副词、连词、某些代词等对分类意义不大的词。其中 TF/IDF 的公式如下：

$$TF/IDF_i=TF(W_i,Doc)\times\log\frac{D}{DF(w_i)}$$

其中，$TF(W_i,Doc)$ 指单字 Wi 在文档 Doc 中出现频度，D 为文档总数，$DF(W_i)$ 为单字 W_i 在其中出现至少一次的文档数目。

2.2 关联规则的产生

经过文本预处理后，就要寻找频繁项集，借鉴了传统的 Apriori 算法，找到了一种很适合处理文本文档事务集的方法：对每个候选项集定义一个 $tidlist$ 的结构，项集 I_i 的 $tidlist$ 由包含事务的 tid 组成。$1-itemset$ 通过搜索文档事务集 D 得到，候选 k 项集的 $tidlist$ 可由该候选 k 项集的那 2 个 $(k-1)$ 项集的 $tidlist$ 求交集产生。

在文档分类中产生的关联规则数量可能是巨大的，为了减少噪音信息和分类时间，所以要对文本关联规则进行剪枝，马光志等人定义了正负相关的定义：

定义 规则 $T\Rightarrow C$(T 为词条项集，C 为类标签)的作用度($lift$)定义为可信度对期望可信度的比值，即 $I(T\Rightarrow C)=P(C|T)/P(C)$。如果，$I(T\Rightarrow C)$ 的值大于 1，则说 T 和 C 正相关，表明每一个的出现都蕴涵另一个的出现；如果结果值等于 1，则它们相互独立，无相关性；如果小于 1，表明它们负相关。

所以在剪枝的时候就根据定义选择正相关的规则，去掉独立和负相关的规则。然后剔除那些特殊并且置信度较低的规则，最后应用数据库候选覆盖来选取最显著规则子集。

2.3 利用文本关联规则进行文本分类

常见的方法是将新文档分配到与之规则匹配最多的类或是分配给与第一个规则匹配的类别。参考文献[9]提出了一种折中方法，综合考虑匹配规则数和置信度的影响。将规则表按类标签分组，设定一个置信度阈值，将低于该阈值的匹配规则从分组中剔除，接着将分组按置信度的和排序。

在进行类识别时，遵循以下两条规则：(1)优先选择置信度和最大分组的标签作为该文档的类标签；(2)在置信度和相等的情况下，选择匹配规则最多组的标签作为该文档的类标签。

3 负关联规则的研究现状及主要技术

关联规则挖掘是数据挖掘研究的一个重要的、高度活跃的领域。传统的关联规则 AR(association rule)是 A

$\Rightarrow B$ 的形式，用于挖掘顾客事务数据库中项集间的关联关系，最初由 R.Agrawal 等人于 1993 年首先在参考文献[11]中提出，以后诸多研究人员对此进行了大量研究，其工作主要是对原有的算法进行改进，以提高算法挖掘规则的效率。关联规则挖掘实质上就是在满足用户给定的最小支持度的频繁项集中，找出所有满足最小置信度的关联规则，具体分为两步：(1)找出所有的频繁项集；(2)用频繁项集产生关联规则。

作为 $A \Rightarrow B$ 型关联规则的一个重要补充，Wu Xin Dong 等将 $A \Rightarrow \neg B$、$\neg A \Rightarrow B$、$\neg A \Rightarrow \neg B$ 三种形式的关联规则称为负关联规则 NAR(Negative AR)，而 $A \Rightarrow B$ 型的关联规则相应地称为正关联规则 PAR (Positive AR)[12]，并且给出了一个 PR 模型以及能够同时挖掘正关联规则和负关联规则的算法[13]，该算法以传统的Apriori 算法为基础来挖掘频繁项集和非频繁项集，在挖掘频繁项集中正关联规则的同时，能够清楚地挖掘非频繁项集中的 $A \Rightarrow \neg B$、$\neg A \Rightarrow B$ 以及 $\neg A \Rightarrow \neg B$ 型负关联规则。

Zhu 等人在参考文献[14]中提出了一种快速有效的基于 FP-tree 的负关联规则挖掘算法 MNARA，该算法不但可以发现所有的负关联规则，而且由于整个过程只需扫描事务数据库两次，算法是有效和可行的。

Dong 等人提出了一种 PNARC 模型[15]，利用了支持度-置信度框架，采用相关性检验方法，不仅能够同时挖掘出频繁项集中的正、负关联规则，而且能够检测并删除相互矛盾的规则；随后又给出一种基于多置信度和 χ^2 检验的挖掘正负关联规则的方法，并且提出了一种 PNARMC 算法[16]，该算法不仅能够正确地产生正负关联规则，而且能灵活地控制关联规则的数量。在 PNARC 模型中给出了正、负关联规则的定义。

定义 设 I 是数据库 D 的项集，A、$B \subseteq I$ 且 $A \cap B = \Phi$，$0 < supp(A)$、$supp(\neg A)$、$supp(B)$、$supp(\neg B) < 1$，$minsupp$、$minconf > 0$；若 $corrA$，$B=1$，A，B 相互独立；否则，A，B 相关，且：

(1)如果 $corrA$，$B>1$，$supp(A \cup B) \geqslant minsupp$ 且 $conf(A \Rightarrow B) \geqslant minconf$，那么 $A \Rightarrow B$ 是一条正关联规则；

(2)如果 $corrA$，$\neg B>1$，$supp(A \cup B) \geqslant minsupp$ 且 $conf(A \Rightarrow \neg B) \geqslant minconf$，那么 $A \Rightarrow \neg B$ 是一条负关联规则；

(3)如果 $corr\neg A$，$B>1$，$supp(A \cup B) \geqslant minsupp$ 且 $conf(\neg A \Rightarrow B) \geqslant minconf$，那么 $\neg A \Rightarrow B$ 是一条负关联规则；

(4)如果 $corr\neg A$，$\neg B>1$，$supp(A \cup B) \geqslant minsupp$ 且 $conf(\neg A \Rightarrow \neg B) \geqslant minconf$，那么 $\neg A \Rightarrow \neg B$ 是一条负关联规则。

4 负关联在 Web 文档分类中的发展趋势

负关联规则的研究为传统关联规则开辟了一个崭新的领域。研究表明，负关联规则同样起着非常重要的作用，一方面可以进一步完善项集间的关联规则分析，另一方面可以为决策支持提供更多有用的信息。

在进行 Web 文档分类的时候，文档集产生的关联规则数往往是巨大的，所以就需要进行剪枝处理，目前在剪枝的时候直接把负相关的规则剪掉，但是在负相关中也包含着有用的信息，所以本文提出了基于负关联规则的 Web 文档分类技术。

基于负关联规则的 Web 文档分类策略，是在基于关联规则分类的基础上，为了更有效地体现现实事件直接的关联而采取的分类策略，这种分类方法和正关联规则分类相结合能够更加有效和准确地进行分类，可以更加全面地分析各种因素之间隐藏的内在联系。运用负关联的研究方法来完善文档集产生关联规则的正确度，从而提高 Web 文档分类的精确度。

现有的 WEB 文档的分类方法中，基于负关联的 Web 文档分类研究的比较少，在以后的研究中应该结合负关联规则将支持度和置信度结合起来考虑，这将是一个新的研究趋势。

参考文献

[1] QUEK C Y. Classification of World Wide Web documents Senior Honors dissertation. School of Computer Science Camegie Mellon University, 1997.

[2] MADRIA S K. Research Issues in Web Data Miming[C]. Proc. of Data Ware housing and Knowledge Discovery, First Inte1.Conf.DaWak 99.1999:303-312.

[3] 符发.中文文本分类中特征选择方法的比较[J].现代计算机，2008(6):43-45.

[4] JOACHIMS T. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of 10th European Conference on Machine Learning(ECML-98), Chemnitz, DE.1998:137-142.

[5] 饶文碧，柯慧燕，张丽.一种扩展的基于 VSM 的 Web 文本分类算法[J].计算机应用与软件，2006，23(10).

[6] 雷景生.基于模糊相关的 Web 文档分类方法[J].计算机工程，2005，31(24).

[7] 张志强，郑家恒.基于加权类轴的 Web 文本分类方法研究[J].计算机应用，2004，24(2).

[8] 胡和平，易高翔.一种基于容错粗糙集的 Web 文档分类方法[J].小型微型计算机系统，2006，27(2).

[9] 马光志，张生庭.基于关联规则的 Web 文档分类[J].计算机工程与设计，2005(9).

[10] AOKI P M. Generalizing search in generalized search trees. In Proc.1998 Int .Conf. Data Engineering (ICDE'98), April 1998.

[11] AGRAWAL R, IMIELINSKI T,SWAMI A. Mining association rules between sets of items in large database[A].Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data [C]. New York: ACM Press, 1993:207-216.

[12] WU Xin Dong, ZHANG Cheng Qi, ZHANG Shi Chao. Mining both positive and negative association rules[A]. Proceedings of the 19th International Conference on Machine Learning (ICML-2002)[C]. San Francisco: Morgan Kaufmann Publishers, 2002: 658-665.

[13] ZHANG W X, ZHANG C, ZHANG S. Efficient Mining of both Positive and Negative Association Rules, ACM Transactions on Information Systems[J]. 2004,22:381-405.

[14] ZHU Y, SUN L, YANG H. Algorithm for Mining Negative Association Rules Based on Frequent Pattern Tree[J]. Computer Engineering,Vol.32, No.22.

[15] WANG D X, SONG S H. Study of Negative Association Rules [J]. Beijing Institute of Technology Journal,2004,24 (11): 978-981.

[16] DONG X, SUN F, HAN X, et al. Study of Positive and Negative Association Rules Based on Multi-confidence and Chi-Squared Test [J]. LNAI 4093, Springer-Verlag Berlin Heidelberg, 2006:100-109.

(收稿日期:2009-04-03)