电子邮件智能分类系统的设计

熊 应[1],朱 斌[1],朱海云[2]
(1. 华南理工大学计算机系,广东广州 510640;2. 广东省电信科学技术研究院,广东广州 510630)

摘 要: 互联网时代,人类的几种主要通讯方式中,电子邮件是既快捷又经济的方式. 对当前人们往来的大量电子邮件,一般在电脑网络中,使用基于关键字比较之分类系统的过滤器对邮件进行分类. 这种传统方式存在缺乏灵活性,不能有效地处理概括性描述复杂性问题等缺点. 使用效果不甚理想. 本文在传统规则分类方法的基础上引入机器学习的知识,设计了一个新的电子邮件智能分类系统,较好地解决了概括性的分类问题,并且通过实验说明了该系统的可行性.

关键词: 电子邮件; 分类; 专家系统; 机器学习
中图分类号: TN182 **文献标识码**: A **文章编号**: 0372-2112 (2001) 12-1653-03

Design of Email Intelligent Classifier

XIONG Ying, ZHU Bin, ZHU Hai-yun
(1. *Computer Science Dept., South China Univ. Of Tech., Guangzhou, Guangdong* 510640, *china*;
2. *Guangdong Telecommunication Academy of Science and Technology, Guangzhou, Guangdong* 510630, *china*)

Abstract: In the Internet Age, Email is one of the fastest and the most economical methods of communication. the large number of emails, by which people communicate with each other nowadays, are usually classified by the filters, which use the regular classifier based on keyword-compared. However, this traditional method has some shortcomings, such as lack of flexibility, incapability of describing the compilcated problims summarily and effectively. Therefore, its effectiveness if not very satisfactory. This paper introduces the knowledge of machine learning on the base of the traditional regular classification. A new intelligent email classifier is desinged to solve the summary classified problems well and demonstrate the feasibility of the system through experiments.

Key words: email; classify; expert system; machine Learning

1 引言

为提高工作效率,人们常使用过滤器(Filter)对电子邮件进行预处理. 过滤器可以把邮件按一定的规则筛选到不同的文件夹中,这一过程又称为分类(Classify).

常见的过滤器都使用规则作为知识表示的方法:

IF X is A, THEN Y is B

Microsoft Outlook Express[7]中的收件助理就是一个使用规则的例子. 在规则方法中,规则前件的匹配通常通过关键字的匹配来完成,其结果只能有两个:要么为 1(匹配),要么为 0(不匹配). 由于不允许出现"如果正文内容与电影有关"这样的概括性描述,使得用户在某些时候难以通过一条或多条规则来描述一些实际问题. 因此,本文在传统方法的基础上引入机器学习的知识,设计了一个电子邮件智能分类系统,较好的解决了概括性的分类问题.

2 智能分类的相关知识和算法

2.1 分类概念简述

数据分类通常可分为两个步骤,首先是训练系统使之具有对新数据进行分类的能力,然后根据已定义的类别为新数据做出标记. 本系统中使用的是根据样本进行分类的方法. 这是基于过去求解类似问题的经验来获得当前问题求解结果的一种推理模式. 所需的样本包括正样本和负样本,分别表示属于/不属于这个类的实例. 我们的训练过程是通过设置和修改正负样本集合来完成的,分类过程是通过比较新元素和样本的相似度来完成的.

2.2 电子邮件的特征描述

RFC822[8]规定了电子邮件的格式标准. 电子邮件由信头(header)和可省略的信体(body)组成. 信头中包含若干个域(field),这些域说明了寄信人、收信人、发信时间等许多信息. 因此我们可以将一封电子邮件看成如下的一个向量:

$$(field_1, field_2, \cdots, field_n, body)$$

此向量的每一维都是一个文本. 由于分类算法的需要,我们采用特征字向量法[6]作为原始文本信息的目标表示方法. 使用′号来表示一个文本的特征字向量,则一个邮件可表示为:

$$(field'_1, field'_2, \cdots, field'_n, body')$$

简单地说,我们把一个邮件表示成由向量组成的向量.

收稿日期:2000-10-11;修回日期:2001-01-02

2.3 对一个类的描述

一个类的定义是通过设置它的样本集合来实现的. 如果每个类的正负样本都分别存放,那么有 n 个类时就要定义 $2n$ 个样本集合. 为简化系统,我们为所有的类设置一个共享的负样本集合,只有正样本集合是专有的,使得样本集合的数量减少到 $n+1$ 个. 这是基于一个合理的假设决定的:通常负样本集合中存放的垃圾邮件对于任何一个类来说都是负样本.

此外每个样本集合还需要一个附属的特征集合 T 来说明在哪些维度上需要比较它们的相似性. 记 $T=(t_1, t_2, \cdots, t_p)$. 其中 t_i 与 2.2 节中邮件向量的每一维相对应, t_i 具有匹配方式、权重等属性.

2.4 智能分类算法

2.4.1 特征字向量的生成 本系统中的字可以是单个的汉语或英语的词汇. 选出的特征字要能最大限度的反映文本的信息特征. 选择特征字时我们制定了以下两条原则:

(1) 设置禁用字表,排除"的"、"了"、"the"等信息含量极低的常用字;

(2) 在正样本中出现频繁而在负样本中极少出现(或反之)的字信息含量高.

依据第二条原则,我们采用了由文献[3,9]中算法修改而来的信息度函数 $E(w)$. 一个字 w 的信息度函数 $E(w)$ 定义:

$$E(w)=p(w)\,I(w)+p(\bar{w})\,I(\bar{w}) \tag{1}$$

其中:$p(w)=w$ 出现的次数/所有字的个数,表示一个字出现的概率,统计范围为该类所有的正负样本;

$p(\bar{w})=1-p(w)$,表示一个字不出现的概率;

$I(w)=-p(w_+)\log_2(p(w_+))-p(w_-)\log_2(p(w_-))$, $p(w_+)$ 和 $p(w_-)$ 分别表示 w 在正、负样本中出现的概率

选出信息度最高的 k 个字组成特征字集(k 可取 96 或 128),便可生成特征字向量. 同一个类中的样本邮件使用同一个特征字集. 每个样本的每一个维度上都有自己的特征字向量,用 1 或 0 表示相应的字在该样本的这个维度中是否出现. 由于下面的分类算法中不需要与负样本的特征向量进行比较,故可以不为负样本计算特征字向量. 当一个类的样本集合发生变化时,要重新计算特征字向量.

2.4.2 相似度的计算 根据样本进行的分类很多,常用的有简单贝叶斯方法、最近邻居法、决策树方法以及神经网络法等. 我们使用的是最近邻居法[5]的变体,先查找每个特征维度上的最近邻居,求出其维度相似度,然后再用加权求和的方法求出综合相似度 . 选取综合相似度最高的类作为分类结果.

算法中用 L 来表示邮件 M 与类 C 的综合相似度,流程如图 1:

精确匹配:

精确匹配中 Vi 的定义为:

$$V_i=\begin{cases}1, & \exists A,\text{使得 } M(t_i)\ RA(t_i)\text{ 且 } A\in C_+\\ 0, & \text{其他}\end{cases} \tag{2}$$

其中 R 是一个二元关系,可根据需要定义为相等、包含等. 该公式的文字描述是:只要正样本集合 C_+ 中有任一邮件在维度 t_i 上与 M 满足关系 R,则 $V_i=1$,否则为 0.

智能匹配:

使用两个特征字向量之间的夹角的余弦来表示它们的相似程度. 用 $P=(C_1, C_2, \cdots C_k, \cdots)$ 来表示一个特征字向量(其中 $C_{1,2}, \cdots \in\{0,1\}$),则它们的相似度 sim 为

$$\mathrm{sim}(P_i, P_j)=\cos(P_i, P_j)=\frac{\sum_k C_{ik}\cdot C_{jk}}{\sqrt{\sum_k C_{ik}^2\cdot\sum_k C_{jk}^2}} \tag{3}$$

依次比较待分类邮件与样本的 sim 值,然后取 Vi 为所有 sim 值中最大者.

2.4.3 分类动作 只有当 L 大于预先设定的阈值 L_0 时,我们才认为分类结果成立,并执行预先定义的动作. 如果没有使 $L>L_0$ 的类或用户对分类结果不满意则说明分类失败,此时需要修改当前类的样本集合或定义新类来处理这些情况.

3 系统模型

3.1 设计目标

智能邮件分类系统的应用场合与传统的规则型过滤器相同,可以应用在邮件服务器端或客户端. 该系统应满足以下要求:

(1) 模块的独立性. 可以对具体的分类算法进行修改而不对软件的整体结构产生大的影响.

(2) 可移植性. 只需要很少的改动就可以使该系统灵活地和一些已有的邮件软件集成起来.

(3) 易操作性. 普通的用户也能很快掌握智能分类的简单知识及其使用方法.

(4) 高效率. 分类所需的时间越短越好,若确实不能立即完成也应随时提示当前进度和状态.

(5) 交互性. 用户通过了解分类的过程和依据,能更有效的设置样本集合,提高系统效率.

3.2 模块划分

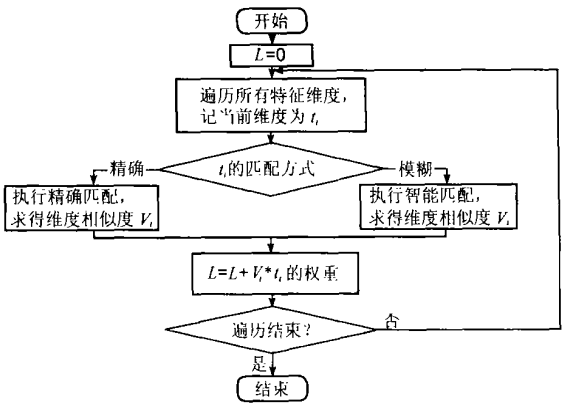

图 1 算法流程图

系统模块结构如图 2 所示. 邮件接收模块:从其它主机上接收邮件并转换成系统内自定义的格式以供进一步处理. 邮件保存库:使用文件夹方式分类保存所有邮件实体. 邮件读写模块:向系统内的其他模块提供存取邮件的服务. 分类模块:相当于专家系统中的推理机. 它从规则和样本库中获取分类所需要的知识,执行编制好的算法对新邮件进行分类. 并根据分类结果来决定应对该邮件进行什么样的操作. 规则及样本

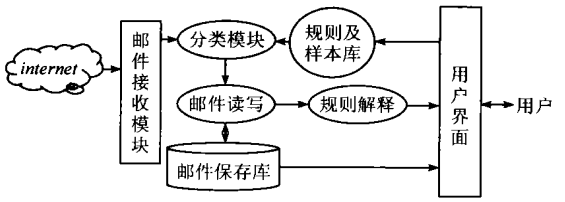

图 2　系统模块图

库:相当于专家系统中的知识库. 它保存了每一个类的全部设定,包括样本集合、特征集合以及所有的特征字向量等. 用户可根据需要对其进行修改. 规则解释模块:从分类模块那里获得分类依据和结果等信息,将其转换成格式化的自然语言提供给用户. 它使得用户知道系统正在干什么和为什么这样干. 用户界面:整个系统与用户交互的窗口. 要尽量做到美观、易用.

4　实验结果

根据上述思想,我们编写了一个原型程序对算法的效率和可行性进行了验证. 用于实验的邮件数据是从若干个人用户的日常邮件中抽样取出的,较好的反映了使用中的常见情况. 实验方法是:首先由该用户对这些邮件进行人工分类,并将这些分好类的邮件定义为样本. 然后从这些样本中随机抽取一个,取消其样本身份后对其进行机器分类. 通过比较人工分类结果和机器分类结果来验证分类的正确性.

按上述方法,我们在关键字匹配型和概括型两种分类标准下分别比较了传统方法和智能方法的差异,结果如表 1:

表 1　传统方法和智能方法比较表

	关键字匹配型问题			概括型问题		
	解决方法	准确率	速度	解决方法	准确率	速度
传统方法	查找某关键字在待分类邮件中是否出现	100 %	快	试图将其简化为关键字匹配型问题进行处理	40 ～ 70 %	快
智能方法	按需要为相应的类设计若干正样本,并使用 2.4.2 节中的精确匹配方法	100 %	快	构成相应的正负样本集合,并使用 2.4.2 节中的智能匹配方法	70～ 80 %	中等

由表 1 可知,智能分类系统在解决关键字匹配型问题时保留了规则方法的全部优点;在解决概括型问题时准确率有了明显提高,其代价是速度略有下降. 智能方法的准确率较高是因为不恰当的关键字在规则方法中常引起误判,而智能分类方法中选取了信息度最高的若干字组成特征字集,并使用相似度函数进行衡量,这些措施都较好的避免了误判的出现.

5　总结

本文是把机器学习的有关知识用来解决实际问题的一个尝试,在保留了传统方法优点的基础上,较好的解决了概括性分类问题. 在确定了算法之后,我们提出了一个改进了的系统结构,并用实验检验了它的可行性和正确性. 实验结果证明:这是一种更先进,更有发展前途的方法.

参考文献:

[1] 蔡自兴,徐光佑. 人工智能及其应用(第二版) [M]. 清华大学出版社,1996.

[2] 邹涛,黄源,张福炎. 基于 WWW 的文本信息挖掘 [J]. 情报学报 1999,18(4).

[3] M Pazzani,L Nguyen,S Mantik. Learning from hotlists and coldlists: towards a WWW information filtering and seeking agent [A]. In Proceedings of AI Tools Conference,Washington,DC,1995.

[4] Bay S D(1999). Nearest Neighbor Classification from Multiple Feature Subsets [J]. Intelligent Data Analysis. 3(3):191 - 209.

[5] Cover T M,Hart P E(1967). Nearest neighbour pattern classification [J]. IEEE Transactions on Information Theory,IT-13(1),21 - 27.

[6] Gerald Salton,A Wong,C S Yang. A Vector Space Model for Automatic Indexing [J]. Communications of ACM,1975,18(11):613 - 620.

[7] http://www.microsoft.com/office/outlook/ [OL]

[8] RFC822 (http://www.faqs.org/rfcs/rfc822.html) [DB/OL]

[9] Billsus D Pazzani M. Revising user profiles: The search for interesting Web sites [A]. Proceedings of the Third International Workshop on Multistrategy Learning (MSL'96),AAAI Press.

作者简介:

熊　应　男. 1978 年生于广西南宁. 1995 年起就读于华南理工大学计算机系,主修计算机软件. 1998 年至今为华南理工大学计算机系硕士研究生,研究方向是计算机应用.

朱　斌　男. 1940 年 3 月生于云南个旧. 教授,研究生导师,中国电子学会高级会员. 1963 年毕业于华南理工大学计算机专业. 毕业后至今长期在华南理工大学计算机系从事教学和科研工作,参加和主持过国家自然科学基金、国家九五攻关、广东省等多项科研. 主要研究方向是计算机应用,曾获国家教委、省科委科技进步奖.