文章编号:1008-0570(2011)03-0167-02

贝叶斯垃圾邮件过滤算法的改进与实现

Improvement and Implementation of Spam Filtering Based on Bayesian

(中国矿业大学 徐州)别玉玉 刘 飞 张书伟
BIE Yu-yu LIU Fei ZHANG Shu-wei

摘要:本文分析了目前在垃圾邮件过滤中广泛应用的朴素贝叶斯过滤算法及其优缺点,并且根据模式匹配和模糊匹配算法提出改进型的贝叶斯邮件过滤模型。首先在邮件预处理过程中进行特征项的提取----模式匹配,从训练集合中识别出正常邮件和垃圾邮件的模式集合,然后用模式集合识别垃圾邮件,再对提取出的特征项进行模糊匹配并根据匹配结果判断邮件是否为垃圾邮件。实验结果表明:应用改进后的算法有效地提高了垃圾邮件过滤的准确率。

关键词: 贝叶斯; 垃圾邮件过滤; 模式匹配; 模糊匹配
中图分类号: TP393.098 文献标识码: A

Abstract: This paper analyzes the disadvantages of native Bayesian algorithm which has been widely used in anti-spam filtering. Moreover, the paper shows a new improved Bayesian filtering model in terms of pattern matching, fuzzy matching. The first preprocess of e-mails filtering is catching feature items----pattern matching which could distinguish junk emails from normal mails with a number of training e-mails. Furthermore, extract the feature items out of the e-mail with fuzzy matching. The experimental result demonstrates that the new algorithm applied has effectively improved the accuracy of spam filtering.

Key words: Bayesian; spam filtering; pattern matching; fuzzy matching

技术创新

1 引言

随着因特网的快速发展,电子邮件逐渐成为人们日常生活中快捷、经济的通信方式。但是电子邮件在提供方便的同时也带来了一些问题,垃圾邮件严重地阻碍了互联网的发展。因此,研究自动过滤垃圾邮件技术具有十分重要的现实意义。垃圾邮件自动过滤技术的研究主要有基于规则过滤和基于概率过滤两种,而后者现已成为主要的研究趋势。由于贝叶斯算法在分类方面具有较好的性能,它在邮件过滤系统中也得到了十分广泛的应用。

Biju Issac 等人的研究中提出了使用关键词上下文的改进贝叶斯过滤算法, 其中涉及到对已有词库中关键词的上下文匹配,但是该算法较为复杂,而且求得的判断邮件是否为垃圾邮件的评价值受到给定权值的影响。本文把研究重点放在贝叶斯分类上,并针对朴素贝叶斯过滤存在的缺陷,在理论分析的基础上提出了一种新的改进型的贝叶斯邮件过滤算法。该算法在邮件预处理中先进行模式匹配, 从训练集合中识别出正常邮件和垃圾邮件的模式集合,然后对提取出的特征项进行模糊匹配。实验结果表明,该算法应用于垃圾邮件过滤确实具有很好的效果。

2 基于朴素贝叶斯算法的垃圾邮件过滤技术及改进方法

2.1 朴素贝叶斯过滤算法

贝叶斯算法是 Thomas Bayes 创建的基于概率的一种算法。该算法通过自我学习,能够及时适应垃圾邮件制造者的新把戏,同时为合法电子邮件提供保护。贝叶斯算法在分类方面具有较好的性能,因此在邮件过滤系统中也得到广泛的应用。

贝叶斯算法的基本思想是收集大量的正常邮件和垃圾邮件作为样本, 然后使用贝叶斯分类器对收集到的样本进行有指导的学习, 最后使用训练好的贝叶斯分类器对新到达的邮件进行分类。通过对邮件样本的训练和学习,贝叶斯分类器可以自动获得垃圾邮件的特征,并根据垃圾邮件特征的变化,准确地对垃圾邮件进行过滤。

朴素贝叶斯过滤算法的应用评价:

优点:

(1)综合性能好,运用了风险最小风险的评价规则,使综合性能好;

(2)空间复杂度较好,因其只存储特征词及字频和概率,因此占用空间较少;

(3)可以动态的改进特征词库,在过滤的过程中,通过学习改进性能。

缺点及改进:

(1)特征词匹配的准确性有待提高;

(2)对混淆词的匹配难以进行;

鉴于以上缺点本文提出了特征词的模糊匹配、模式匹配,使之可以更好的过滤垃圾邮件。

2.2 贝叶斯算法的综合改进

在贝叶斯过滤算法中需要提取邮件主题和邮件体中独立的字符串,进而建立相应的哈希表,特征项的提取离不开分词技术[3]。本文采用改进算法从训练集中识别出模式,并将其作为表示邮件的特征,再进行模糊匹配,在此基础上执行改进的贝叶斯分类算法,最终改进贝叶斯垃圾邮件过滤系统。

贝叶斯垃圾邮件过滤中邮件预处理过程:

(1)从训练集中得到所有符合条件的特征项;

(2)根据模式匹配算法,选择一定数量的特征项构成特征集;

(3)根据邮件对特征集中所有特征的匹配情况,计算邮件对各特征项的权值。

2.2.1 特征项提取的改进---模式匹配

运用贝叶斯算法处理垃圾邮件, 首先要从训练集中提取符

别玉玉: 学生

合条件的特征项。传统的贝叶斯垃圾邮件过滤中,把词作为特征项,词语的抽取非常容易,但提取出来的特征项较多,邮件向量空间的维数较大,邮件的分类效率较低。本文改进贝叶斯算法,以模式(即正则表达式)作为特征项,从邮件中获取出现次数超过一定阈值的基本模式,然后构造出所有的最大模式,提取出来作为特征项。

若 U 表示字母表的集合,模式被定义为符合下列形式的任意字符串:$U \| \left(\sum \{'*'\}\right) \| U$。

其中‖表示连接,* 表示任意字符,$\sum$表示中间字符的集合。即任何以字母开始和结束的字符串,并且包含一个可选的由“*”组成的字符集。

关于模式的定义:

(1)对任意模式 M,它的每个本身也构成模式的子字符串叫做 M 的子模式。

(2)将 M 中一个或多个任选字符(“*”)更换成确定的字符,或者添加一个由任意字符组成的字符串到 M 的左边和(或)右边,这样得到的模式 M ’就认为比模式 M 更具体。

(3)对于一个输入字符串集合 S,如果不存在另一个模式 M ’,使得 M ’比 M 更具体,那么模式 M 被称为对 S 是最大的。

使用上面的定义,特征项提取的问题就转换为下面的问题:

对于给定一个输入字符串结合 $S = \{s_1, s_2 \cdots \cdots s_n\}$,找到所有最大模式, 而且这些模式至少出现在属于 S 的若干个独立字符串中。

这样,在改进的贝叶斯垃圾邮件过滤算法中,解决从训练集中提取特征项的问题即是解决字符串的模式匹配问题, 将出现频率大于某特定阈值的基本模式识别出来, 然后将这些基本模式将组合成越来越大模式, 直到所有存在的最大模式都被构造出来,作为特征项。提取出来的特征项用于模式匹配,以便构成特征集。

2.2.2 特征项匹配的改进---模糊匹配

特征项的匹配是垃圾邮件过滤过程中十分重要的步骤,本文将关键词由精确的单词发展为模糊的单词来扩大命中率即模糊匹配。模糊匹配是依据两个词的“距离”来判定的是否匹配的,这个距离是可以变化的,当这个“距离”小于某个阀值可以认定匹配,否则认定不匹配。

实现特征项的模糊匹配可以使用备忘录算法,备忘录算法为每个子问题建立一个记录项,初始时,该记录存入一个特殊值,表示该问题尚未求解。在求解过程中,对每个待求的子问题,首先要查看其相应的记录项。如果记录项存储的是初始时存入的特殊值,则表示该问题第一次遇到,此时计算该问题并保存在其相应的记录项里面,以备以后查看。若记录项里面已不是初始设置的值了,那就说明该问题已经被解决过,其记录项应该是问题的解答。

规则:

$$d(C_i\ T_0) = i \quad d(C_0\ T_j) = j$$

$$d(C_i\ T_j) = \begin{cases} d(C_{i-1}\ T_{j-1}) & C_i = T_j; \\ 1 + \min\left(d(C_{i-1}\ T_j)\ \ d(C_i\ T_{j-1})\ \ d(C_{i-1}\ T_{j-1})\right) & C_i \neq T_j \end{cases}$$

$$d(C\ T) = d(C_m\ T_n)$$

备忘录算法实际上是递归算法,伪代码算法实现如下:

```
void BeiWangLu(int m, int n, int **d, int *C, int *T) {
    for(int  i=1;i<=m;i++)
    for(int j=1;j<=n;j++)  d[i][j]=0;
}
int LookUp(int i, int j) {
  if((C[i] != T[j]) && (d[i-1][j] != 0) && (d[i][j-1] != 0) && (d[i-1][j-1] ! = 0))
    return d[i][j] = 1 + min(d[i-1][j], d[i][j-1], d[i-1][j-1]);
   else  if((C[i] = T[j]) && (d[i-1][j-1] ! = 0))  return d[i][j]=d[i-1][j-1];
    else  if(C[i] ! = T[j]) &&((d[i-1][j] == 0) || (d[i][j-1] == 0) || (d[i-1][j-1] == 0)))
    //d[i-1][j]、d[i][j-1]和 d[i-1][j-1]三者至少有一个等于 0
    return d[i][j] = 1 + min((LookUp(i-1,j) or d[i-1][j]),
     (LookUp(i,j-1) or d[i][j-1]),(LookUp(i-1,j-1) or  d[i-1][j-1]));
    //当某个需要的子问题未求解时递归调用函数求解该子问题,若已解答则可以直接用子问题的解答
    else  if((C[i] = T[j]) && (d[i-1][j-1] = 0)) return d[i][j] = 1 + LookUp(i-1,j-1);
    //递归求解所需子问题的解答
}
```

3 实验及结果

实验中收集了各类邮件 4000 封,其中合法邮件 2000 封,垃圾邮件 2000 封。由于时间和空间的有限性,实验收集到的邮件数目和领域都有限。

为了验证算法的性能, 实验中首先把 1000 封垃圾邮件和 1000 封合法邮件作为训练集,剩下的 2000 封用来进行测试。

表 1 实验结果

	贝叶斯过滤		改进后的贝叶斯过滤	
	合法邮件	垃圾邮件	合法邮件	垃圾邮件
合法邮件	930	70	950	50
垃圾邮件	130	870	100	900

根据实验结果,垃圾邮件过滤的准确性有了提高,改进后的算法在邮件过滤中具有较高的查全率和查准率, 具有较强的邮件过滤性能,与朴素贝叶斯算法相比邮件误判率也随之降低。改进后贝叶斯过滤在只有很少的训练集时也表现出了较好的过滤性能。

4 结束语

本文针对朴素贝叶斯过滤存在的不足,在理论分析的基础上提出了一种新的改进型的贝叶斯邮件过滤算法。通过对改进前的算法和改进后的算法分别用邮件进行实验验证,实验结果表明,该算法可以有效地提高过滤器的精确率。当前垃圾邮件的发送者千方百计采取新措施以绕过分类器,从而避免被过滤。为了有效处理垃圾邮件,我们应当针对最新的垃圾邮件特点进行研究,将多种反垃圾邮件技术结合起来,这样才能更好的过滤垃圾邮件。

本文作者创新点: 提出改良的贝叶斯算法,用模式匹配算法从训练集中识别出正常邮件和垃圾邮件模式集合; 使用模糊匹配算法对特征项进行分析检验。

作者对本文版权全权负责,无抄袭。

参考文献

[1]Biju Issac, Wendy Japutra Jap, Jofry Hadi Sutanto. Improved Bayesian Anti-Spam Filter - Implementation and Analysis on Independent Spam Corpuses [R]. International 2009 IEEE DOI 10.1109/ICCET.2009.170,2009. P326-330.

[2]Jiansheng Wu,Tao Deng . Research in Anti -Spam Method Based on Bayesian Filtering[R]. IEEE Pacific-Asia Workshop on

(下转第 248 页)

线的精确拟合,为实现凸轮轴全数控加工提供了轨迹参数。

作者对本文版权全权负责,无抄袭。

参考文献

[1]Tsay Der-Min, Tseng Kuo-Shu, Chen Hsin-Pao. A Procedure for Measuriing Planar Cam Profiles and Their Follower Motions[J]. J. Manuf. Sci. Eng., 2006, 128(3): 697-704.

[2]WONG Y S, ONG C J. Optimization approach for biarc curve-fitting of B-spline curves [J]. CAD Computer Aided Design, 1996, 28(12): 951-959.

[3]Jacek M Zurada. Introduction to Artificial Neural Systems[M]. Paul: West Publishing Company, 1992.

[4]张春仙,张巍等.基于 BP 神经网络的电路最优测试集的生成设计[J].微计算机信息.2009,2-2:288-289.

[5]M. S. Dawson, A. K. Fung, and M. T. Manry. Surface Parameter Retrieval Using Fasting Learning Neural Networks [J]. Remote Sensing Reviews, 1993, 7: 1~18.

作者简介:吴占涛(1982-),男(汉族),河南人,工学硕士,研究方向:计算机测试与控制、电机控制。

Biography: WU zhan-tao (1982-), Male (clan of han), Henan Province, Engineering Master, Research Area: Computer Test and Control Technology, Motor Control Technology.

(410100 湖南 长沙 三一重工股份有限公司) 吴占涛 戴胜军

(410082 湖南 长沙 湖南大学) 吴占涛

(SANY Heavy Industry CO., LTD., Changsha 410100, Hunan, China) WU Zhan-tao DAI Sheng-jun

(Hunan University, Changsha 410082, Hunan, China) WU Zhan-tao

通讯地址:(410100 湖南省长沙市星沙镇三一重工泵送技术部) 吴占涛

(收稿日期:2010.07.22)(修稿日期:2010.10.22)

(上接第 168 页)

Computational Intelligence and Industrial Application, 2008.

[3]夏克俭,张涛等.基于贝叶斯算法的垃圾邮件过滤的研究[J].微计算机信息.2008,3-3:P179-180

[4] 张楠. 一种基于可重构计算的汉字模糊匹配算法与硬件实现[D].中国科学院研究生院硕士论文,2006.

[5]王晓东.计算机算法设计与分析[M]. 电子工业出版社,2007.

作者简介:别玉玉(1990 年—),女(汉族),山东济宁人,大学本科,研究领域为信息安全;导师简介:毕方明(1974—),男,汉族,博士,信息安全专业,中国矿业大学计算机学院副教授。

Biography: BIE Yu-yu (1990-), female (the Han nationality), Shandong Province, China University of Mining and Technology, Undergraduate, major is information security, research area is computer system security.

(221116 江苏 徐州 中国矿业大学计算机科学与技术学院信息安全系) 别玉玉 刘 飞 张书伟

(Dept. to Information Security, School of Computer Science, China University of Mining & Technology, Xuzhou Jiangsu 221116, China) BIE Yu-yu LIU Fei ZHANG Shu-wei

通讯地址:(221116 江苏省徐州市中国矿业大学南湖校区竹一 B2051) 别玉玉

(收稿日期:2010.08.16)(修稿日期:2010.11.16)

技术创新