大数据挖掘分类算法在垃圾邮件过滤中的应用

张 赟 贾宇波 (浙江理工大学,浙江 杭州 310018)

摘要:大数据挖掘突破传统数据分析,在大数据领域占据重要位置。随着信息交流日益多样化,电子邮件成为日常交流重要工具之一,但是垃圾邮件的产生成为一大难题亟待解决。为给互联网用户提供一个绿色的信息交流环境,利用大数据挖掘中朴素贝叶斯算法、Winnow 算法、PageRank 算法等对电子邮件中垃圾邮件进行过滤处理,从而彰显电子信息交流中数据的价值。

关键词:大数据挖掘,朴素贝叶斯算法,Winnow 算法,PageRank 算法

据不完全统计,全世界每天发送近 3000 亿封电子邮件,但其中广告性质的邮件至少占 30%。本文结合大数据挖掘的相关知识,对电子邮件中海量的数据进行分析,主要运用朴素贝叶斯过滤技术解决垃圾邮件这一实际问题。

1 大数据挖掘概述

何为大数据?多么大的数据才能称为大数据?不同的年代会有不相同的答案[2]。大数据挖掘是指通过一定方法在大数据上进行相应的数据挖掘,从海量数据中提取特征数据并对数据再分析挖掘出那些隐含的不为人所知的信息。大数据挖掘的过程是从设计到创造相应大数据挖掘模型的过程[10]。

大数据挖掘相关领域是相当广泛的,现在有很多的方法和应用技术已经占据主导地位[9]。美国数据存储公司 EMC 首席市场官 Jeremy Burton 曾说:"大量杂而无章的信息无休止地增加,带来了无穷无尽的机会,将促使社会、技术、科学和经济发生根本性改变。信息是企业最重要的资产,大数据正在促使企业改变信息管理方式,并从信息中挖掘出更大的价值[1]。"

2 垃圾邮件过滤

2.1 基于朴素贝叶斯算法的分类技术

朴素贝叶斯(Naive Bayes,NB)分类算法适用于大型数据库且分类准确度高,应用该算法理论对电子邮件中已经确定的垃圾邮件学习,根据邮件相比较后出现的相似度来推断是否为垃圾邮件。贝叶斯定理公式如下:

$$P(A|B)=\frac{P(B|A)}{P(B)}\times P(A) \tag{1}$$

公式(1)中的 P(A|B)表示在 B 事件发生的前提下 A 事件发生的概率,即后验概率,P(A)是先验概率,即在整个事件中 A 事件发生的概率。现假定有 n 个电子邮件,每个电子邮件有 n 个特征($\alpha_1,\alpha_2,\cdots,\alpha_n$)给定 k 个类用 c_k(k=1,2,…,m)表示,则某一电子邮件含特征集 b 时,被判定为是 c_k 类的概率为:

$$P(c_k|b)=\frac{P(b|c_k)}{P(b)}\times P(c_k)\ \ (k=1,2,\cdots,m) \tag{2}$$

其中:$P(b|c_k)=P(\alpha_1,\alpha_2,\cdots,\alpha_n|c_k)$,此时需要确定邮件所属类别的概率大小,即 $P(c_k|b)$的大小。在上式中 P(b)与电子邮件所属类别无关,要比较电子邮件属于哪一类,只需要计算出 $P(c_k)$和 $P(b|c_k)$。$P(c_k)$是先验概率较为简单,关键在于 $P(b|c_k)$。朴素贝叶斯算法的特点在于为了使计算简便假设特征项之间互相独立.在此假设下,通过以下公式(3)可以方便求出 $P(b|c_k)$:

$$P(b|c_k)=P(\alpha_1,\alpha_2,\cdots,\alpha_n|c_k)=\prod_{i=1}^{n}P(i|c_k) \tag{3}$$

2.2 基于 Winnow 算法的分类技术

Winnow 算法是除朴素贝叶斯算法之外另一种较为常见的垃圾邮件分类过滤处理方法[11]。根据相关数据经验为该算法设定阈值 θ 后按以下方法操作:

1)记空间大小为 n,i 为 1 至 n 间任意数,初始化向量 a_i;

2)调整向量,若分类正确则不做改动;若电子邮件非垃圾邮件但 $\Sigma a_i x_i>\theta$,可以重新确定当 $x_i\neq 0$ 时所对应的 $a_i=\alpha a_i(0<\alpha<1)$,不断调整使其满足 $\Sigma a_i x_i<\theta$;若电子邮件为垃圾邮件但 $\Sigma a_i x_i<\theta$,可以重新确定当 $x_i\neq 0$ 时对应的 $a_i=\beta a_i(\beta>1)$不断调整使其满足 $\Sigma a_i x_i>\theta$;

3)对数据集进行检测,将 $x_i=(x_1,x_2,\cdots,x_n)$与 $a_i(i=1,2,\cdots,n)$加权求和,若加权和 $\Sigma a_i x_i>\theta$ 则 xi 是垃圾邮件。

Balanced Winnow 算法是由 Winnow 算法衍生出的算法,可以通过它使得电子邮件的分类效率进一步提高。

2.3 基于搜索引擎的分类技术

在垃圾邮件问题处理上,基于搜索引擎的 PageRank 算法起到了相当重要的作用。如果某页面频繁被其他不同页面引用就体现出此页面的重要性,如果页面没有被不同页面引用但是被某一重要页面所引用,则此页面也可能为重要页面。页面体现出的重要性平均传递到引用它的页面中,计算得出的分值为 PageRank 值[6]。

2.4 基于黑白名单的分类技术

黑名单指垃圾邮件的发送者信息集合,白名单指通过验证且值得信任的发送者信息集合。中国互联网协会反垃圾邮件综合处理平台通过拒绝恶意邮件站点链接有效拒绝垃圾邮件,使用 ISCBL 对 IP 实时监控[13]。目前较流行的是通过 DNS(查询和区域传输)方式实现实时黑名单技术(RBL)。该技术运用 DNS 查询 IP 地址是否存在于实时黑名单中,如果存在于黑名单中,则得到一个肯定的答案,否则查询会得到一个否定的答案[5]。虽然黑白名单技术尚且只能应对少部分的垃圾邮件,但一定程度上为遏制垃圾邮件做出贡献[12]。

2.5 图片识别过滤技术(OCR)

打击图片垃圾邮件的主导技术有图片垃圾指纹识别技术、OCR 识别技术以及之后的第三代图像防御技术[3]。其中新型 OCR 引擎可以对图片进行深入分析,处理过程更加规范化,对垃圾邮件过滤的准确率高达 95%。该项技术的针对性较强,对图像分析处理的严密性使得垃圾邮件过滤效果明显改善。

3 应用案例分析

3.1 基于 NB 算法的案例分析

为了能更好地了解大数据挖掘在垃圾邮件处理中体现出的优势,本文将结合朴素贝叶斯算法对一篇英文垃圾邮件进行过滤。首先在庞大的数据集中选取部分数据进行数据集成,然后按

以下步骤过滤处理。

(1)电子邮件的预处理

将电子邮件中一些标点和无意义词语消除，然后从剩余内容中寻找特征词汇。英文垃圾邮件内容:We have revised our Hatch Cover drawing to reflect changes that should ensure the passing of the next samples that will be provided.The changes on the hatch cover are highlighted on the drawing so it will be easy to see what needs to changed.I have attached a word file of the revised hatch cover drawing reflecting the necessary changes and I will mail three hard copies to the Shanghai Office per your request。

预处理后内容为:revised hatch cover drawing reflect changes ensure passing samples provided changes hatch cover highlighted drawing easy see needs changed attached word file revised hatch cover drawing reflecting necessary changes mail three hard copies shanghai office per request。

(2)构建词频矩阵

预处理之后需构建相应的词频矩阵。词频矩阵由电子邮件中词语出现频率大小即词语出现次数组成。矩阵中行代表电子邮件文档编号,列表示电子邮件中特征词汇,内容为词汇在对应电子邮件文中出现的次数。表 1 是该简单数据集中显示的垃圾邮件部分词频矩阵。

表 1 词频矩阵(部分)

	1	2	3	4	5	6
change	5	0	0	0	0	0
sample	1	0	0	0	0	0
easy	1	0	0	0	0	0
airport	0	1	0	0	0	0
Chama do	0	0	0	0	1	1
job	0	0	1	0	0	0
ensure	1	0	0	0	0	0

(3)统计特征词出现概率

假设特征词间是相互独立的,利用以下公式求出 $P(blc_k)$的值。

$$P(blc_k)=P(\alpha_1,\alpha_2,\cdots,\alpha_n|c_k)=\prod_{i=1}^{n}P(\alpha_i|c_k) \tag{4}$$

公式(4)中可知,只需将电子邮件中特征词在各电子邮件中出现概率相乘,通过大数据挖掘中的布尔矩阵得出所属类别,与词频矩阵相似唯一不同的是矩阵中的值仅为 0 和 1，若包含特征词数值为 1,否则数值为 0。

(4)垃圾邮件最后分类

电子邮件分类的最后一步是计算该电子邮件中特征词在各类电子邮件中出现概率并相乘，根据朴素贝叶斯定理比较该电子邮件分别是垃圾邮件和正常邮件概率大小，从而判别是否为垃圾邮件。

3.2 基于 Balanced Winnow 算法的案例分析

本案例将采用 Balanced Winnow 算法对垃圾邮件进行过滤。首先,在庞大的数据集中选取部分数据进行数据集成,对数据集进行特征选取形成特征集。数据集中记空间大小为 n,i 为 1 至 n 间任意数，电子邮件编号用 x_i 表示,x_i 可以取两个值 0 和 1,0 表示未能体现该特征,1 表示能体现该特征,此算法含有 a_i^+ 和 a_i^- 两个权重向量。

然后进行错误反馈学习,若 $\Sigma(a_i^+-a_i^-)x_i>\theta$ 且电子邮件显示为垃圾邮件,则无需改动。若该电子邮件非垃圾邮件但 $\Sigma(a_i^+-a_i^-)x_i>\theta$,通过 $a_i^+:=\lambda a_i^+$,$a_i^-:=\mu a_i^-$($0<\lambda<1,\mu>1$)使权重的值能够降低;若该电子邮件属于垃圾邮件,但 $\Sigma(a_i^+-a_i^-)x_i<\theta$,通过 $a_i^+:=\mu a_i^+$,$a_i^-:=\lambda a_i^-$($0<\lambda<1,\mu>1$)使权重的值上升,最后对整个电子邮件数据集进行垃圾邮件检测,将 $x_i=(x_1,x_2,\cdots,x_n)$ 与 $a_i^+-a_i^-$(i=1,2,…,n)加权求和,若 $\Sigma(a_i^+-a_i^-)x_i>\theta$ 那么 x_i 是垃圾邮件[4]。

表 2 特征词集(部分)

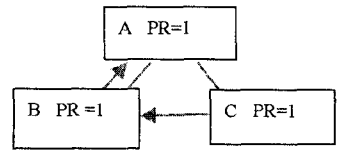
	x1	x2	x3	x4	x5	x6
change	1	0	0	0	0	0
sample	1	0	0	0	0	0
easy	1	0	0	0	0	0
airport	0	1	0	0	0	0
Chama do	0	0	0	0	1	1
job	0	0	1	0	0	0
ensure	1	0	0	0	0	0

3.3 基于 PageRank 算法的案例分析

本案例将结合 PageRank 算法具体分析垃圾邮件处理过程.页面 PageRank 值 PR 计算公式为:

$$PR(T)=\frac{1-d}{N}+d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \tag{5}$$

其中 PR(T)为页面 T 的 PageRank 值;$PR(T_i)$为页面 T_i 的 PageRank 值,T_i 链接向 T 页面;$C(T_i)$为 T_i 链接的数量;d 为阻尼系数,d 的取值范围为 0~1,一般取值为 0.85,N 为所有网页的数量[5]。从电子邮件数据集中随机找出 3 个网页分别标记为 A、B、C，各网页的 PageRank 的初始值均为 1,d 取 0.85。



结合上式页面 A 传给页面 B 和页面 C 的值均为 1×0.85×1/2=0.425,B 只传给 A 的值为 1×0.85=0.85,同理 C 传给 B 的值为 1×0.85=0.85。经过一次转移后,对于 A:0.15(A 未传给其他页面)+0.85(B 转移至 A)=1,即 PR(A)=(1-d)+d×(PR(B)/

表 3 近似迭代表

	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	1.425	0.575
2	1.36125	1.06375	0.575
3	1.0541875	1.21728125	0.72853125
4	1.184689063	1.21728125	0.598029688
5	1.184689063	1.161818086	0.653492852
6	1.137545373	1.208961775	0.653492852
7	1.177617509	1.188925707	0.633456784
8	1.160586851	1.188925707	0.650487441
9	1.160586851	1.196163737	0.643249412
10	1.166739176	1.190011412	0.643249412
11	1.1615097	1.19262615	0.64586415
12	1.163732227	1.19262615	0.643641623
13	1.163732227	1.191681576	0.644586197
14	1.162929339	1.192484464	0.644586197
15	1.163611794	1.192143236	0.644244969
16	1.163321751	1.192143236	0.644535013
17	1.163321751	1.192266505	0.644411744
18	1.163426529	1.192161727	0.644411744
19	1.163337468	1.192206257	0.644456275
20	1.163375319	1.192206257	0.644418424

(下转第 131 页)

骤，与本文方法进行比较。原文方法将所有像素点均作为建立复杂网络模型的节点，节点数目固定为 10000 个。本文方法仅将提取后的轮廓点作为建立复杂网络模型时的节点，节点数目大幅下降。

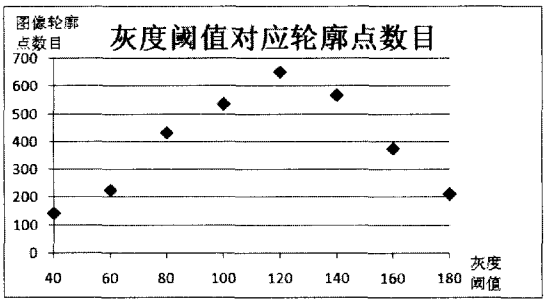


图 6 不同灰度阈值下提取图像轮廓点数目

图 6 是对 YALE 人脸数据库第 1 组第 1 幅图像应用本文方法，在不同灰度阈值下得到的轮廓点数目。最少为 140 个，最多为 653 个，平均提取轮廓点数目为 376 个。

节点数的下降使得运行程序所需存储空间相应减少。原文方法生成邻接矩阵时，对应规模为 10000×10000，即需要 1 亿个存储单位，在内存较小的实验平台下，出现 Matlab 内存溢出警示。本文方法对第 1 组第 1 幅图生成邻接矩阵时，规模最多仅为 653×653，即仅需占用约 43 万个存储单元，约为原文方法的 0.4%。

4.3.2 实验二

选取每组图像的前 5、6、7、8 幅图作为训练样本，其余图像作为测试样本进行图像识别，与算法 PCA、Fisherface 进行比较，实验结果如表 1 所示。

表 1 算法在 YALE 数据库中的识别结果

训练样本数	PCA	Fisherface	本文方法
5	83.33	92.22	86.67
6	80	92	87.27
7	85	98.33	95.15
8	91.11	97.78	97.57

从表 1 可以看出，本文方法在 YALE 人脸数据库中的实验结果较为理想。从识别率看，其效果优于经典的 PCA 方法，但与 Fisherface 方法比较仍有差距。本文只是简单选取了两种图像识别方法进行对比，在今后的研究可以对更多的识别算法进行研究对比，并考虑将其与复杂网络进行融合，改进轮廓点提取方法，更好地保留适于复杂网络方法结合的形状信息，控制复杂网络规模，进一步提高识别方法准确率。

参考文献

[1]刘剑.基于关键点检测的图像信息简约表达及应用研究[D].天津：天津大学，2012

[2]Turk M,Pentland A.Eigenfaces for recognition [J].Journal of Cognitive Neuroscience,1991,3(1):71-86

[3]Felzenszwalb P F,Huttenlocher D P.Pictorisl structures for object recognition [J].International Journal of Computer Vision, 2005, 61(1): 55-79

[4]潘珍.基于轮廓的形状识别方法研究[D].重庆：西南大学，2012

[5]Backes A R, Martinez A s, Bruno O M.A complex network-based approach for boundary shape analysis [J]. Pattern Recognition, 2009, 43: 54-67

[6]章毓晋.图像工程（上册）：图像处理与分析[M].北京：清华大学出版社，1999

[7]汪小帆，李翔，陈关荣.复杂网络理论及其应用[M].北京：清华大学出版社，2006

[8]任海鹏，马展峰.基于复杂网络特性的带钢表面缺陷识别[J].自动化学报，2011，11(5)：1407-1412

[9]Huang Zixuan,Ma Chao,Huang Jiangnan.Link Prediction Based on Clustering Coefficient[J].Applied Physics,2014,04(6)

[10]汤晓.基于复杂网络的图像目标识别方法研究[D].广州：广东工业大学，2013

[11]梅向明.微分几何[M].4 版.北京：高等教育出版社，2010

[收稿日期：2015.12.30]

（上接第 128 页）

C(B))；对于 B：0.15(B 未传给其他页面)+0.425(A 传给 B)+0.85(C 传给 B)=1.425，即 PR(B)=(1-d)+d×(PR(A)/C(A))+PR(C)/C(C))；对于 C：0.15(C 未传给其他页面)+0.425(A 传给 C)=0.575，即 PR(C)=(1-d)+d×(PR(A)/C(A))。

为使运算结果更为准确，本文采用 Google 公司使用的近似迭代方式，默认为 20 次，迭代效果如表 3。

由 PR 近似迭代值可知 B 的 PageRank 值为最大，则 B 在三个网页中最重要，其次为 A，最后为 C，以此类推计算 PageRank 值可实现垃圾邮件分类过滤[8]。

4 结束语

本文运用大数据挖掘分类算法中的 NB 算法、Winnow 算法等对垃圾邮件进行过滤是大数据挖掘领域的一次成功探索。此后将沿着这条路进一步研究垃圾邮件过滤技术，在前人的基础上使其更具有现实意义。

参考文献

[1]谭磊.New Internet 大数据挖掘[M].北京：电子工业出版社，2013

[2]Adam Jacobs.The Pathdogies of Big Data[J].2009,52(8).Communications of the ACM,2009

[3]贾云刚.垃圾邮件过滤技术研究[J].通信与信息技术，2009(2)：55-58

[4]李志刚，马刚.数据库与数据挖掘的原理及应用[M].北京：高等教育出版社，1999

[5]陆嘉恒.大数据挑战与 NoSQL 数据库技术[M].北京：电子工业出版社，2008

[6]王涛，徐洁.搜索引擎排序技术研究[J].电脑知识与技术，2009(5)：1250-1252

[7]宋明秋，曹晓芸.基于敏感特征的网络钓鱼网站检测方法[J].大连理工大学学报，2013(6)：903-907

[8]史磊峰，孟嗣仪，刘云.搜索引擎排序算法的探索[J].铁路计算机应用，2010(12)：21-24

[9]Jeffrey Hsu.DATA MINING TRENDS AND DEVELOPMENTS：The Key Data Mining Technologies and Applications for the 21st Century[R],2002

[10]Tan P.N.,Steinbach M.,Kumar V.Introduction to Data Mining [M].New York:Addison Wesley,2005

[11]张丽.基于 Winnow 算法的反垃圾邮件引擎的设计与实现[D].南京：东南大学，2006

[12]郭泓.电子邮件过滤技术浅析[J].信息网络安全，2002(10)：42-44

[13]简艳英，刘敏.基于 Winnow 算法的反垃圾邮件引擎的设计与实现[D].南京：东南大学，2006

[收稿日期：2015.11.30]