

基于词条时序的朴素贝叶斯垃圾邮件过滤方法

常 青

(辽宁医学院, 辽宁 锦州 121000)

摘 要: 朴素贝叶斯分类算法是一种有效的垃圾邮件过滤技术. 互联网上的信息随着时间推移产生概念的变迁, 最近出现的垃圾邮件词条可作为判定垃圾邮件的重要依据. 将新近的垃圾邮件词条单独记录, 在进行邮件分类时, 对于最近出现的垃圾词条, 提高其对垃圾邮件判定的先验概率. 通过实验对比, 提出的垃圾邮件过滤方法较传统的朴素贝叶斯垃圾邮件过滤具有更高的准确性、精确性和召回率.

关键词: 朴素贝叶斯; 垃圾邮件过滤; 概念漂移; 时间序列

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2010)05-0212-05

Method of Spam Filtering Based on Naive Bayesian Algorithm with Token Time-Series

CHANG Qing

(Liaoning Medical University, Jinzhou 121000, China)

Abstract: Naive Bayesian classification algorithm is an effective spam filtering method. Information on the Internet changed over time and generated new concepts. The recent spam tokens were an important basis for spam filtering. In this approach, new spam tokens were recorded separately according to their time series and those tokens got higher prior probability in the email classification. Experimental results show that the proposed algorithm has higher accuracy, precision and recall than the traditional Naive Bayesian spam filtering method.

Key words: Naive Bayesian; spam filtering; concept drift; time series

1 引言

现代化通信技术的发展极大简化了沟通手段, 降低了沟通成本. 随着 Email 应用的普及, 垃圾邮件也成为了众所周知的通信问题, 虽然垃圾邮件不像电脑病毒那样具有极强的破坏力, 但它严重影响了人在通信过程中的精力, 为网络流量和存储带来了额外的开销. 反垃圾邮件技术正是在如此背景下发展起来的信息过滤技术.

基于模式匹配的垃圾邮件过滤是反垃圾邮件技术的重要分支, 其核心思路为通过机器学习手段, 获取垃圾邮件的特征并记忆, 对于新邮件, 通过匹配邮件内容的典型特征, 判定是否为垃圾邮件. 该领域内, 最为经典的垃圾邮件过滤技术是基于朴素贝叶

斯(Naive Bayes, NB)分类器的邮件过滤^[1]. 使用朴素贝叶斯分类技术, 构建了一系列的商用垃圾邮件过滤器, 包括 Spam Trapper, GFI Mail Essentials 和 Outlook Spam Filter 等. 另外, 为提高 NB 邮件过滤的准确率, 研究者提出了多种改进算法^[2-3]. 国内的研究者, 也对此领域给予了充分的关注^[4-5].

作为典型的离线垃圾邮件过滤技术, 朴素贝叶斯邮件过滤面临新的挑战: 邮件内容随着时间的变化而出现新的趋势, 例如, 垃圾邮件发送者发现“发票”已经被过滤, 则将之改为“fa 票”或“发-票”的形式; 再例如, 垃圾邮件的发送者的“兴趣”发生了转变, 由“发票”转变为“交友”. 学术界将这种词条变化称为概念漂移(Concept Drift)^[6]. 核心词汇的变化需要快速地反映在垃圾邮件过滤系统中, 而传统的

NB 分类算法的学习过程并未考虑概念漂移. 本文的研究目标是考虑邮件内容的变化过程, 并将变化快速的反映给 NB 分类器, 使传统的 NB 垃圾邮件过滤可适应在线邮件处理.

2 基于朴素贝叶斯的垃圾邮件过滤

基于朴素贝叶斯分类器的垃圾邮件过滤技术属于典型的机器学习技术, 分为模式学习和模式匹配两个重要阶段.

模式学习过程中, 首先根据训练集邮件的词条(Token), 判定每个词条属于“合法”和“非法”邮件的概率, 进而获取全局性的词条分类(合法、非法)概率. 模式匹配阶段依据词条分类概率, 计算邮件的分类. 具体做法为:

(1) 建立两张哈希表, 分别代表合法邮件 Non-Spam-Hash-Table 和非法邮件 Spam-Hash-Table;

(2) 取任意一封邮件, 提取典型词条, 若此邮件为合法邮件, 则将词条和词条频度记录于 Non-Spam 表中; 否则, 写入 Spam-Hash-Table;

(3) 迭代(2), 至完成所有训练集邮件的学习;

(4) 通过 Non-Spam-Hash-Table 表和 Spam-Hash-Table 表, 获取词条对邮件分类的贡献, 垃圾邮件过滤时, 重点考察邮件属于垃圾邮件的概率:

对于词条集合 $T = \{t_1, t_2, \dots, t_m\}$, 对应 Spam-Hash-Table 的词频集合为 $S\{s_1, s_2, \dots, s_m\}$, 对应 Non-Spam-Hash-Table 的词频集合为 $N\{n_1, n_2, \dots, n_m\}$, 任意 $t_i \in T$, 其对垃圾邮件的贡献为 $P_s(t_i)$:

$$P_s(t_i) = s_i / \sum_{j=0}^m s_j \quad (1)$$

对非垃圾邮件的贡献为

$$P_n(t_i) = n_i / \sum_{j=0}^m n_j \quad (2)$$

(4) 获取学习结果, 计算各词条的垃圾邮件后验概率, 记为 Token-Prob-Hash-Table: 若 t_i 词条出现在邮件中, 则此邮件为垃圾邮件的后验概率为

$$P(\text{Spam} | t_i) = \frac{P_s(t_i)}{P_s(t_i) + P_n(t_i)} \quad (3)$$

纠错率过高是朴素贝叶斯邮件过滤的缺点, 显然, 对于“将垃圾邮件误认为合法邮件”与“将正常邮件误认为垃圾邮件”, 后者的代价更高. 因此, 对于式(3), 往往通过增大非垃圾邮件贡献 $P_n(t_i)$ 来降低纠错率, 可以将式(3) 改写为

$$P(\text{Spam} | t_i) = \frac{P_s(t_i)}{P_s(t_i) + 2 \times P_n(t_i)} \quad (4)$$

(5) 模式匹配阶段, 对于指定的测试邮件 E , 提取其典型词条 $T_E = \{t_1, t_2, \dots, t_k\}$, 查找 Token-Prob-Hash-Table, 此邮件为垃圾邮件的概率为

$$P(\text{Spam} | T_E) = \frac{\prod_{i=1}^k P(\text{Spam} | t_i)}{\prod_{i=1}^k P(\text{Spam} | t_i) + \prod_{i=1}^k (1 - P(\text{Spam} | t_i))} \quad (5)$$

3 考虑邮件时序的朴素贝叶斯算法

互联网作为动态演化的实时系统, 信息量随着时间呈递增关系, Google 服务中, 专门提供了热门搜索功能, 列出当前网络最为流行的主题与词组. 垃圾邮件的发展和互联网信息的变迁相同步. 因此, 利用朴素贝叶斯分类算法处理垃圾邮件过滤时, 需要考虑概念(词条)的漂移, 新增概念的特点主要表现为两个方面:

(1) 新增概念往往属于新造词, 或者对于原有词汇的人为变化, 因此, 通常并不会出现于 Spam-Hash-Table 中;

(2) 新增概念体现流行趋势, 并体现局部性原理: 一旦一个新的词条出现在垃圾邮件中, 则近期的邮件若含有这个词条, 往住会是垃圾邮件.

3.1 新增词条获取

新增词条最准确的来源方式是反垃圾邮件服务提供商直接发布, 并更新至邮件客户端. 另外, 垃圾邮件过滤算法可以自主学习新增词条, 对于用户指定的垃圾邮件 $E_{\text{spam-custom}}$ 和用户未处理(系统判定)的垃圾邮件 $E_{\text{spam-system}}$, 新增词条的获取方式相同, 处理方式不同. 步骤如下:

(1) 新建两散列表, 分别记录“待确定的新增词条”和“新增词条”, 分别记为 New-Token-Suspend 和 New-Token, 前者记录词条和词频, 后者记录词条和词条出现的时间(邮件发送时间);

(2) 对于垃圾邮件 E_{spam} , 首先提取其所有词条 $T\{t_1, t_2, \dots, t_m\}$;

(3) 获取新增词条集合 T_{new} :

$$T_{\text{new}} = \{t_1, t_2, \dots, t_m\}$$

$$T_{\text{new}} \subset T \text{ 且 } t_i \notin \text{Spam-Hash-Table} \quad (6)$$

(4) 对于 $t_i \in T_{\text{new}}$, 若 $E_{\text{spam}} \in E_{\text{spam-custom}}$, 则将 t_i 直接记录于 Non-Spam-Hash-Table 之中, 并写入 New-Token.

(5) 对于 $t_i \in T_{\text{new}}$, 若 $E_{\text{spam}} \in E_{\text{spam-system}}$, 则将

t_i 记录于 New-Token-Suspend 中, 如 t_i 在 New-Token-Suspend 中的词频大于 3, 表示 t_i 已经在被判定为垃圾邮件的邮件中的出现 3 次, 则移除 New-Token-Suspend 并记录于 New-Token 中;

(6) 将 New-Token 中的词条依据时间顺序逆序排列, 最新出现的词条位于散列最前.

新增词条发现是一自动化过程, 其最大特点是建立了新增词条散列表 New-Token, 显然, 过期的词汇不应该出现在 New-Token 之中, 同时考虑算法性能, 文中建议 New-Token 的最大长度为 10, 当新的词汇加入后, 过期词汇将会被逐步淘汰.

3.2 邮件时序对邮件判别的影

通过新增词条的学习, 邮件过滤系统能够清楚地掌握最新的垃圾邮件流行词汇. 进行词条的垃圾邮件先验概率判定时, 需要考虑新增词条对邮件分类的判定结果. 当邮件中出现新增垃圾邮件词条时, 则此邮件为垃圾邮件的概率明显增大.

垃圾邮件模式匹配过程中, 对于新收到邮件的词条组合 $T\{t_1, t_2, \dots, t_m\}$, 逐一计算其对垃圾邮件的贡献, 计算过程中, 采用式(1)获取词条 t_i 的基础概率, 而后, 查找散列表 New-Token, 若 $t_i \in \text{New-Token}$, 则扩大 t_i 对于垃圾邮件的贡献 $P_s(t_i)$, 记为 $P_{s-\text{new}}(t_i)$:

$$P_{s-\text{new}}(t_i) = P_s(t_i) + \lambda$$

$$\lambda = 1 - \text{Index}(t_i)/10 \quad (7)$$

调节因子 λ 与 t_i 在 New-Token 中的位置关系成正比, 时间最近的新增词条对垃圾邮件的判定影响最大. 考虑垃圾邮件判别算法的纠错率, 单一词条对垃圾邮件的贡献不应过大, 取调整后的垃圾邮件贡献为

$$P_s(t_i) = \text{MAX}[0.99, P_{s-\text{new}}(t_i)] \quad (8)$$

3.3 算法流程

基于时序的贝叶斯垃圾邮件过滤算法流程如图 1 所示, 算法可以分为 5 个关键步骤:

(1) 依据历史邮件记录, 拆分邮件内容, 构建 Non-Spam-Hash-Table 和 Spam-Hash-Table;

(2) 计算各词条对垃圾邮件的先验概率, 算法如式(4)所示;

(3) 依据新邮件的词条, 使用式(5)计算此邮件属于垃圾邮件的后验概率, 判定邮件类别, 判定过程中, 结合 New-Token 中的词条顺序, 计算方法如式(7)、式(8)所示;

(4) 对于新增的垃圾邮件, 依据 2.1 节中的所述步骤, 提取新增词条, 纪录于 New-Token 表中;

(5) 将新增词条写入 Spam-Hash-Table, 完成增量式机器学习过程.

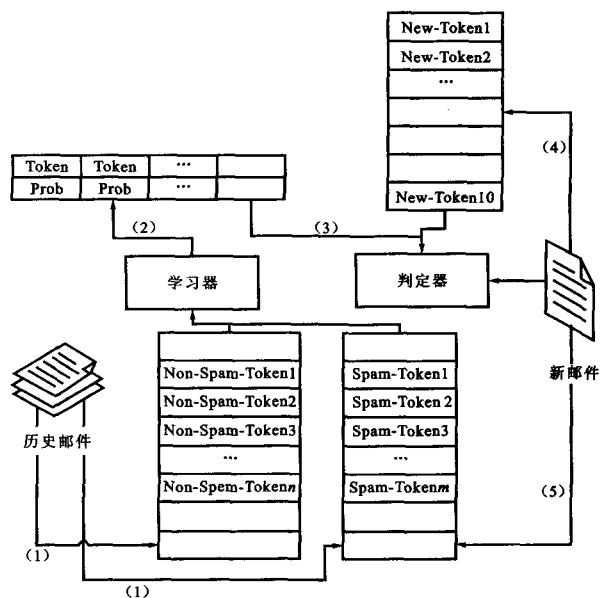


图 1 基于时序的朴素贝叶斯垃圾邮件过滤流程

实际应用中, 对于首次出现的词条, Token-Prob-Hash-Table 中不存在其先验概率纪录, 此时取固定概率 0.4; 另外, 考虑算法的处理效率, 需要严格限制 Token-Prob-Hash-Table 的规模, 为了防止其极具膨胀, 计算词条对邮件分类的贡献时, 仅考虑在 Spam-Hash-Table 和 Non-Spam-Hash-Table 中出现次数超过 3 次的词条, 这种做法同时也防止了仅出现一次的词条对于邮件误判的影响, 对于用户明确标记的垃圾词条, 则将其在 Spam-Hash-Table 中的词频自动修改为 3 次. 此外, 对于词条的选择数量, 也应该控制其规模, 并非越大规模的词条越有利于邮件判定^[7].

4 仿真实验

4.1 评价指标

垃圾邮件过滤算法的评测通常采用精确率、正确率和召回率三个技术指标, 对于 M 封测试邮件, 考虑四种情况: 将垃圾邮件判定正确, 记为 Spam-Spam; 将垃圾邮件判定错误, 记为 Spam-Non; 将非垃圾邮件判定正确, 记为 Non-Non; 最后 Non-Spam 表示对非垃圾邮件的误判, 则:

精确率(Accuracy):

$$\text{Accuracy} = (\text{Spam-Spam} + \text{Non-Non}) / M \quad (9)$$

正确率(Precision):

$$\text{Precision} = \text{Spam-Spam} / (\text{Spam-Spam} + \text{Non-Spam})$$

Spam) (10)

召回率(Recall):

$$\text{Recall} = \text{Spam-Spam} / (\text{Spam-Spam} + \text{Spam-Non}) \quad (11)$$

4.2 实验方法

单纯的研究离线朴素贝叶斯垃圾邮件分类时,可以采用 UCI 测试数据集或 PU1 数据集检验算法效能.文中提出的垃圾邮件判别算法需要考虑邮件接收时间,经典数据集所提供的数据无法满足要求.为进行实验仿真,通过互联网选择了2 400封邮件,其中包含垃圾邮件 890 封,非垃圾邮件1 510封,中文分词并非本文的关注点,故所有邮件均为英文邮件,使用标点符号和空格即可划分邮件内容词条.

实验过程如下:

(1)将2 400封邮件依据邮件发送时间顺序排列;

(2)选择前1 000封邮件作为初始训练集,训练朴素贝叶斯分类器;

(3)将后1 400封邮件中,随机抽取 480 封为测试集,其余邮件作为增量式学习集合;

(4)对于后1 400封邮件,依据时间顺序逐一呈交算法中,若邮件为测试邮件,则使用传统朴素 Bayes 判定其邮件类别,并使用本文提及的方法重复判断;对于非测试邮件,进行增量式学习;

(5)统计两种算法对 480 封测试邮件的正确率、准确率和召回率;

(6)重复步骤(3)~(5),共计 5 次,计算统计指标平均值.

另外,进行邮件类别判定时,对于垃圾邮件概率超过 0.5 的邮件,记为垃圾邮件.

4.3 实验结果

图2~图4给出了文中算法与经典朴素贝叶斯算法的正确率、精确率与召回率的指标对比.从图2可以看出,随着词条取值逐渐增多,算法正确率逐渐增高,词条数到达200之后,正确率维持在一个较高的水平,且变化不再明显,说明随着词条的增多,非垃圾邮件被误判为垃圾邮件的比例维持在相似水平,也就是说,垃圾邮件的特征是比较明显的,非垃圾邮件误判为垃圾邮件的可能性不高;图4中,200词条附近,召回率达到极大值,随着词条继续增多,召回率反而下降,意味着更多的词条为垃圾邮件的判断带来了噪音,使得更多的垃圾邮件被当正常邮件处理了.由于召回率的变化趋势,算法精确率表现出类似的变化,如图3所示.

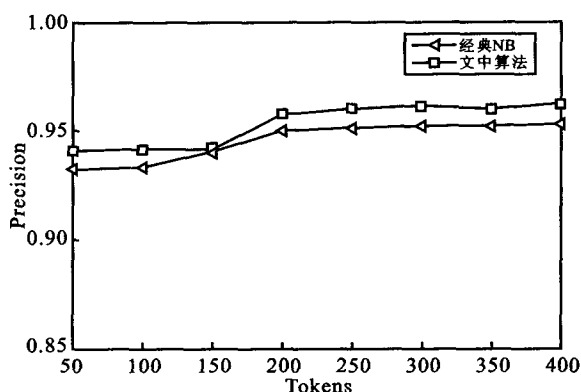


图2 算法正确率对比

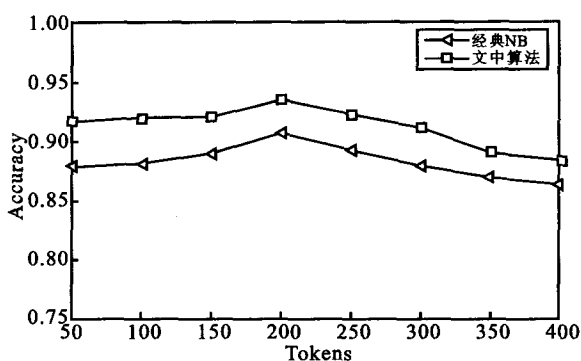


图3 算法精确率对比

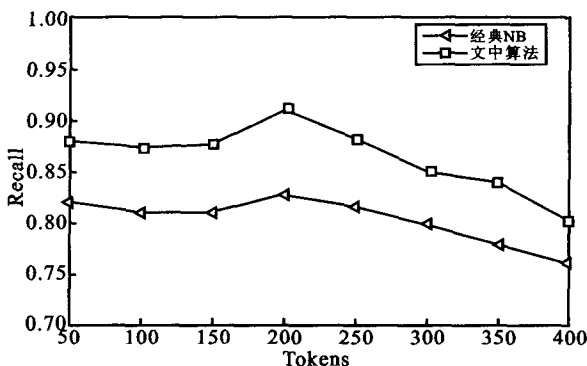


图4 算法召回率对比

需要注意,应用过程中,并非越多词条越好,在词条数目为200时,算法性能达到最优,相关指标表1如示.

表1 算法指标对比(200 词条) (%)

算法	指标		
	Accuracy	Precision	Recall
经典 NB	92.70	95.10	83.01
本文算法	93.51	95.77	91.00

5 结束语

基于朴素贝叶斯理论的垃圾邮件过滤得到了广

泛的关注和应用,文中考虑了词条的时序特征,采用近期出现的垃圾词条辅助垃圾邮件判别过程,实验结果表明,文中方法较传统贝叶斯垃圾邮件过滤更加有效.基于机器学习的邮件过滤技术派生出了多个分支,使用决策树等其他经典机器学习技术处理垃圾邮件过滤,将是下一步研究的方向.

参考文献:

- [1] Androutsopoulos I, Koutsias J, Chandrinou G, et al. An evaluation of naive bayesian anti-spam filtering[C]// Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning. Barcelona, Spain, 2000:9-17.
- [2] Androutsopoulos I, Koutsias J, Chandrinou K, et al. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with encrypted personal email messages[C]// Procs of 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Greece, Athens, 2000:160-167.
- [3] Sakkis G, Androutsopoulos I, Paliouras G, et al. A memory-based approach to anti-spam filtering for mailing lists[J]. Information Retrieval, 2003,6(1):49-73.
- [4] 李翔鹰,叶枫.一种基于多贝叶斯算法的垃圾邮件过滤方法[J].计算机工程与应用,2006(31):114-116.
- [5] 廖明涛,张德运,李金库.基于朴素贝叶斯和层次聚类的两阶段垃圾邮件过滤方法[J].微电子学与计算机,2007,24(8):1-4.
- [6] Sarah J D, Padraig C, Alexey T, et al. A case-based technique for tracking concept drift in spam filtering[C]// Applications and Innovations in Intelligent Systems XII. Procs. of AI, England:Cambridge, 2004:3-16.
- [7] Androutsopoulos I, Koutsias J, Chandrinou K V, et al. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with encrypted personal e-mail messages[C]// Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000). Athens, Greece,2000:160-167.

作者简介:

常青女,(1980-),助理实验师.研究方向为网络信息安全、计算机应用.

(上接第211页)

工厂层在三层架构中不算实体层次,只是为了方便客户端调用而存在的一个服务对象外观类.

4 结束语

面向重用的构件化平台的开发,将从根本上提高软件开发的效率,被认为是一种有效的解决软件危机的手段,并接受新兴软件制造业带来的挑战,但是,目前该研究还处在开始阶段,很多想法还不成熟,如何用更好的设计思想来完善构件化平台,对企业级应用提供更好的技术支撑,是我们下一步研究的重点.

参考文献:

- [1] 杨芙清,梅宏.构件化软件设计与实现[M].北京:清华大学出版社,2008.
- [2] Erich Gamma, Richard Helm, Ralph Johnson, et al. 设计模式——可复用面向对象软件的基础[M].北京:机械工业出版社,2007.
- [3] Rockford Lhotka. Expert C# 2005 business objects[M].北京:电子工业出版社,2006.
- [4] Clinton Begin, Brandon Gooding, Larry Meadors. Ibaitis in Action[M]. USA, Greenwich: Manning Publications, 2007.
- [5] Juval Lowy, Clemens Vasters. Programming WCF services[M]. USA, Sebastopol, O'Reilly Media,2008.
- [6] Robert E Filman, Tzila Elrad, Siobhan Clarke, et al. Aspect-oriented software development[M]. Boston: Addison-Wesley,2005.
- [7] Christian Bauer, Gavin King. Hibernate 实战[M].北京:人民邮电出版社,2008.

作者简介:

杨晔女,(1978-),讲师.研究方向为信息管理技术、软件开发及测试技术.