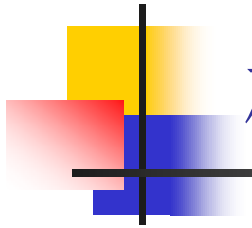
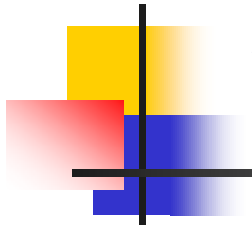
中文垃圾邮件过滤系统的实现和评估

田莹
北京 清华大学 网络中心
Email:tianying00@mails.tsinghua.edu.cn



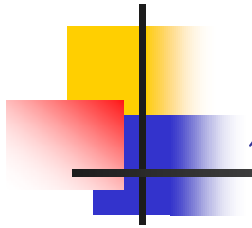
概要

- 引言
- 研究背景
- 中文垃圾邮件过滤系统的实现
- 中文垃圾邮件过滤系统的评估
- 最新研究进展及结论



引言

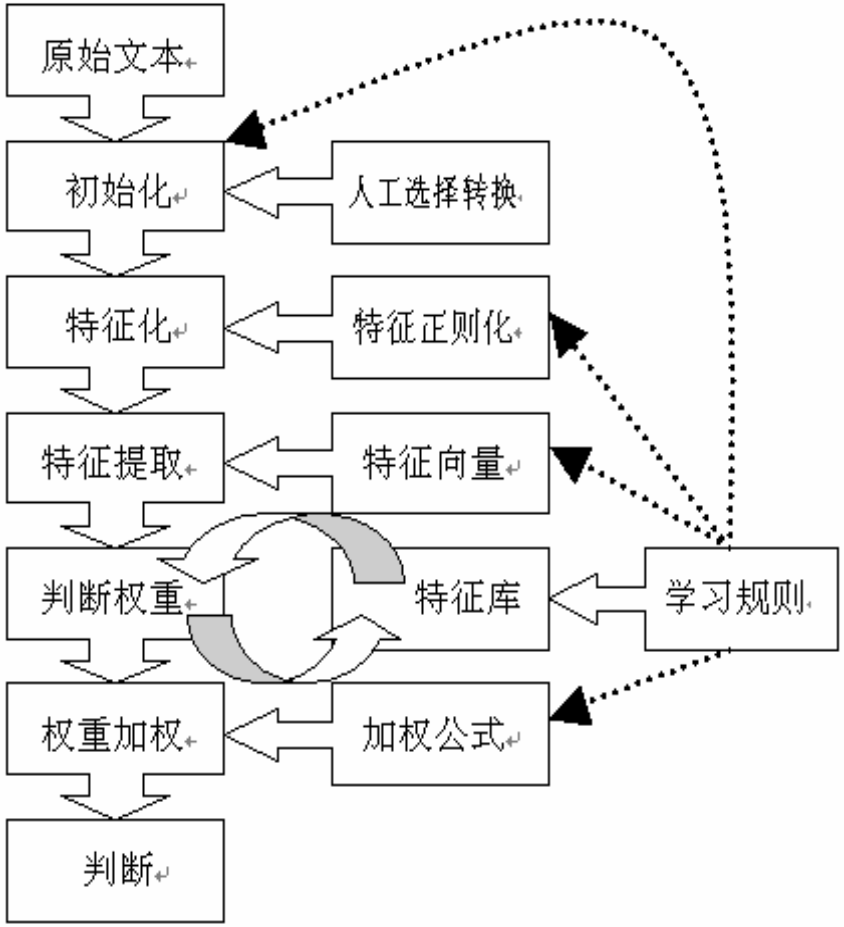
- 垃圾邮件的定义
- 垃圾邮件的危害
- 反垃圾邮件的意义
  - Email，短信，VoIP电话……
- 垃圾邮件的特性



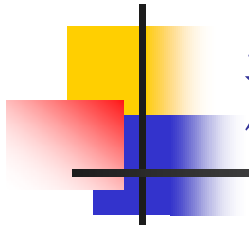
反垃圾邮件的方法

- 黑白名单
- 关键字匹配
- 贝叶斯
- SVM
- Etc.

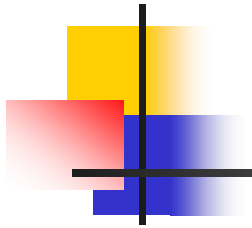
基于内容的过滤器的流程图



英文垃圾邮件的贝叶斯过滤流程

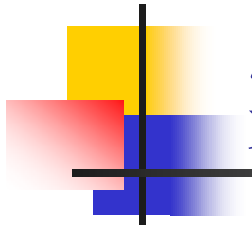


- 收集两个数据库
  - 垃圾邮件数据库
  - 正常邮件数据库
- 在每一个数据库中，学习并定义出一些关键词，计算这些关键词的概率
- 新邮件到来时，计算出新到来的邮件中包含的关键词的联合概率
- 通过联合概率判断新到来的邮件是否是垃圾邮件



中文邮件的预处理

- 中文分词的概念
- 分词算法
  - 基于字符串匹配
  - 基于理解
  - 基于统计
- 中文分词的词典
  - 基于整词二分
  - 基于TRIE索引树
  - 基于逐字二分



实验数据来源

- CCERT提供
  - http://www.ccert.edu.cn/spam/index.htm
- 训练用邮件数
  - 5000
- 测试用邮件数
  - 500

评估指标

- 定义L为正常邮件，S为垃圾邮件。S→L表示将垃圾邮件判定为正常邮件，同理，L→S表示将正常邮件判定为垃圾邮件。
- 在文本分类问题中，有两个评估指标被经常使用。

$$Acc = \frac{n_{L\to L} + n_{S\to S}}{N_L + N_S}, Err = \frac{n_{L\to S} + n_{S\to L}}{N_L + N_S}$$

- Acc称为正确率。Err称为错误率。这里 $n_{L\to L}$ 表示将正常邮件判断为正常邮件的个数。
- $n_{L\to S}, n_{S\to S}, n_{S\to L}$ 的含义可以类推。N_L 和 N_S 分别表示待判定的正常邮件和垃圾邮件的总个数。

评估指标（续1）

- 考虑到L→S和 S→L分别会有不同的代价，并设L→S的代价是S→L的代价的 λ 倍，我们定义两个新的评估指标，分别是WAcc（加权的正确率）和WErr（加权的错误率）

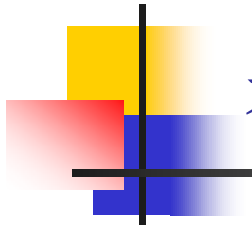
$$WAcc = \frac{l\,n_{L\to L} + n_{S\to S}}{l\,N_L + N_S}, WErr = \frac{l\,n_{L\to S} + n_{S\to L}}{l\,N_L + N_S}$$

- 在没有过滤的情况下（无论是正常邮件还是垃圾邮件一律通过），我们得到基准WAcc和基准WErr分别为：

$$WAcc^b = \frac{l\,N_L}{l\,N_L + N_S}, WErr^b = \frac{N_S}{l\,N_L + N_S}$$

评估指标（续2）

- 为了方便比较，定义比率R为 $R=\frac{WErr^{b}}{WErr}$
- 不难看出R越大，过滤的效果越好。R如果小于1，意味着过滤比不过滤效果还差



参数说明

- 我们的算法中有两个重要的参数
  - 用于训练的样本个数n
  - 在过滤中计算最终概率的特征数目m
- 实验中，主要研究R和n以及R和m之间的相互关系。

实验结果

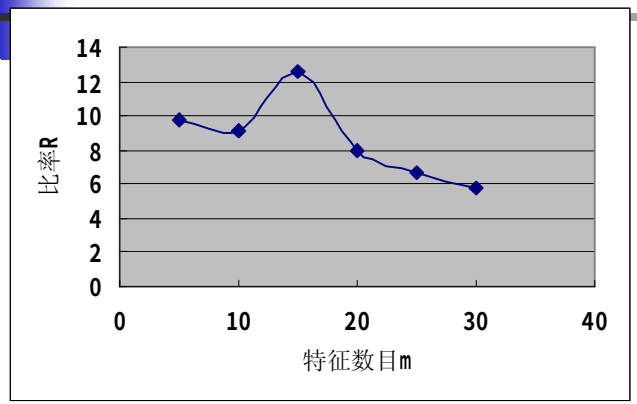


图1 R-m关系图 $l=1$

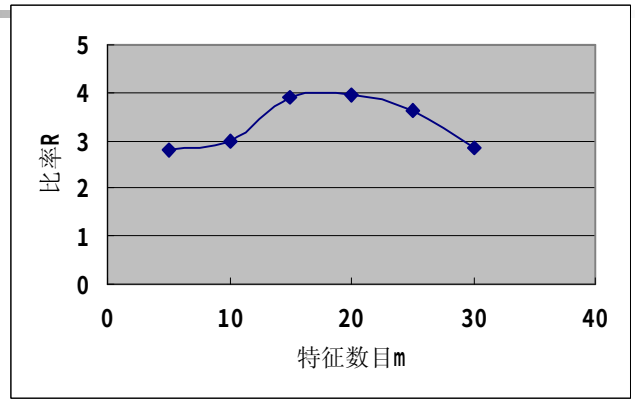


图2 R-m关系图 $l=9$

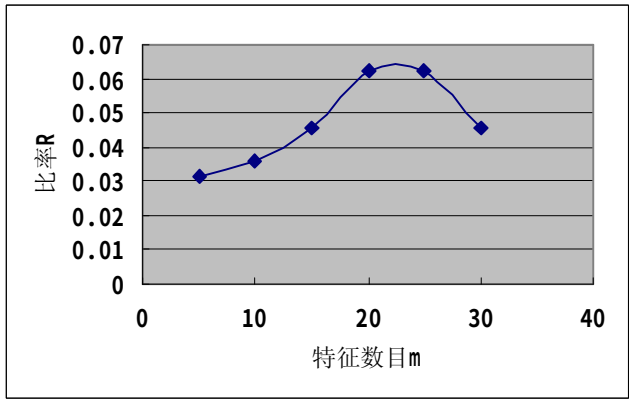


图3 R-m关系图 $l=999$

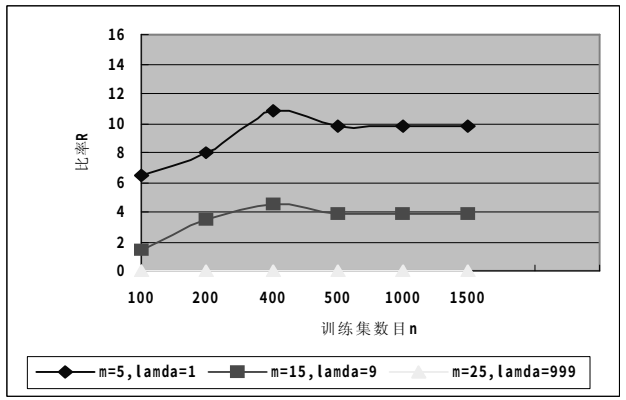
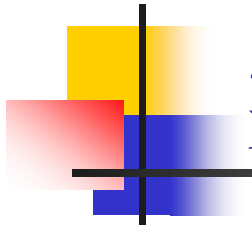
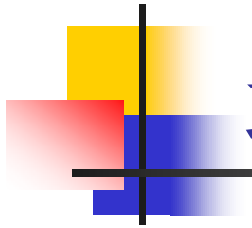


图4 R-n关系图 $l=1, m=5$
$l=9, m=15$
$l=999, m=25$



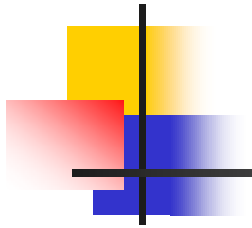
实验结果说明

- 在过滤中计算最终概率的特征数目m以及用于训练的样本个数n都存在某个最优值
- 当用于训练的样本个数逐渐超过这个最优值时，过滤效果会略微下降并趋于一致。



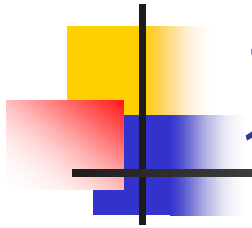
最新研究进展

- 相关会议
  - MIT spam conference
  - CEAS（电子邮件和反垃圾邮件会议）



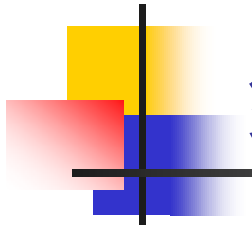
贝叶斯过滤发展方向

- 从单一关键词到关键词链
- 从线性到非线性
- 从单一用户到综合多用户
- 从客户端到服务器
- 利用电子邮件网络
- Etc.



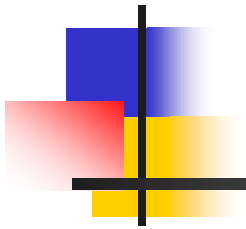
系统和产品

- 微软公司： SmartProof
- IBM公司： SpamGuru
- Etc.



结论

- 反垃圾邮件的挑战



谢谢大家

Q & A