基于神经网络的反垃圾邮件系统的设计与实现

马刚 北京邮电大学教育信息化办公室

罗琴 西南石油大学计算机科学学院

2015/12/7

目 录

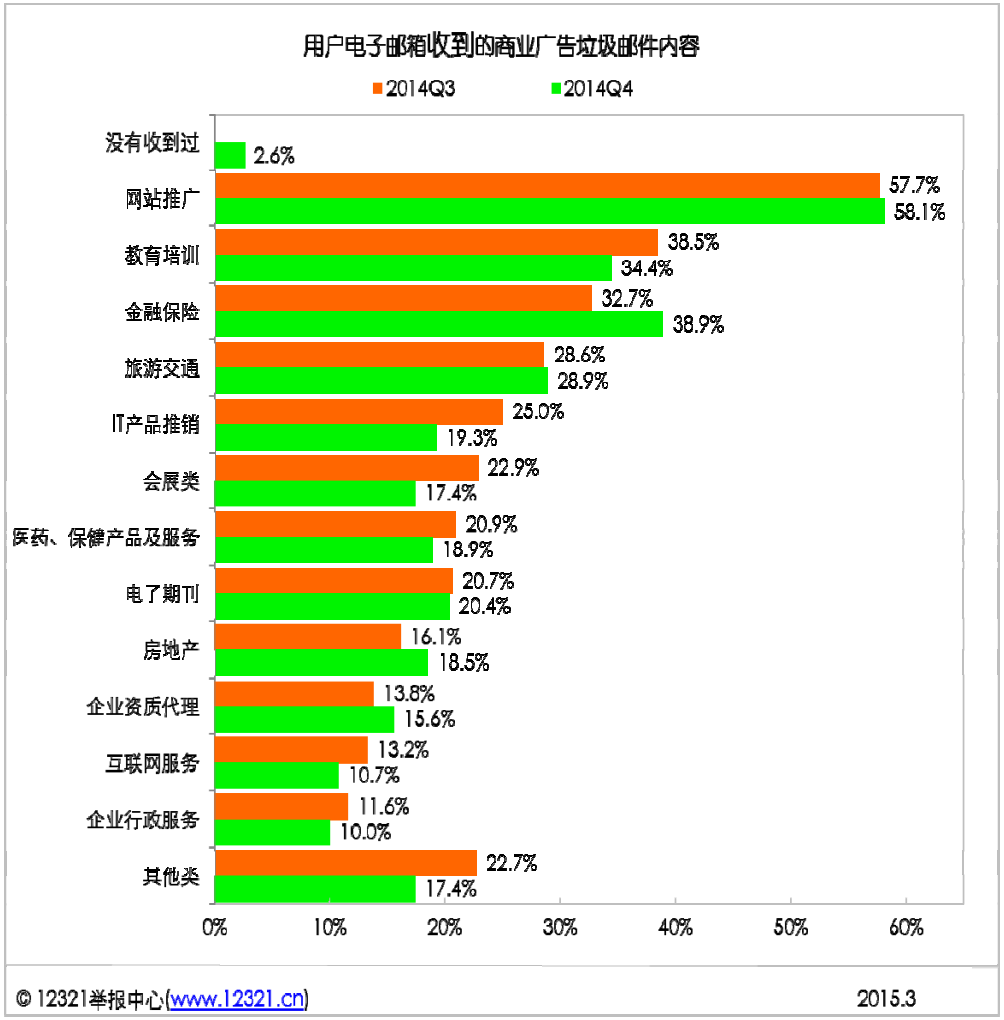
1.用户每周收到垃圾邮件数量

- 2014年第四季度，中国电子邮箱用户平均每周收到垃圾邮件数量为**14.3**封，环比上升**1.5**封，同比上升**1.7**封；
- 中国电子邮箱用户平均每周收到的邮件中，垃圾邮件所占比例为**41.0%**，环比上升了**7.9**个百分点，同比上升了**8.7**个百分点。
- 从年度走势来看，本季度中国电子邮箱用户平均每周收到**垃圾邮件数量**与垃圾邮件**所占比例**均有所**上升**。

2.用户每周收到的垃圾邮件内容

受访用户选择占前五名分别为：

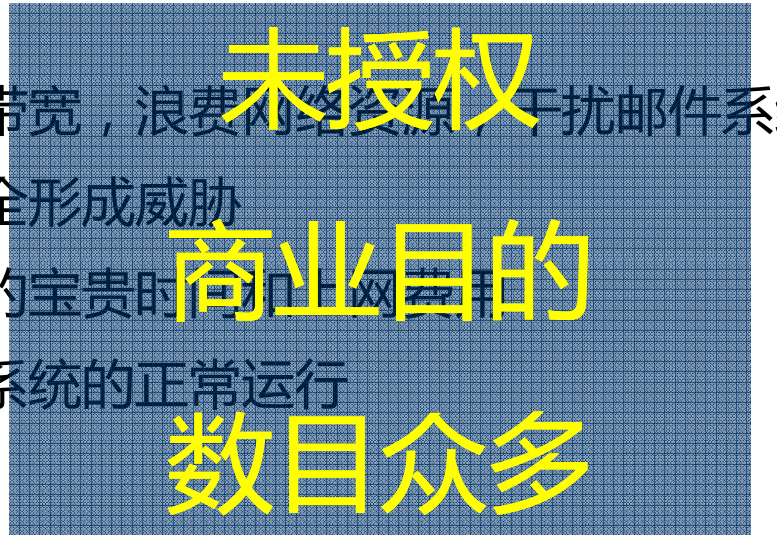
- 网站推广类（**58.1%**），上升**0.5**个百分点；
- 金融保险类（**38.9%**），上升**6.2**个百分点；
- 培训类（**34.4%**），下降**4**个百分点；
- 旅游交通类（**28.9%**），上升**0.3**个百分点；
- 电子期刊（**20.4%**），下降**0.3**个百分点



垃圾邮件的定义及危害

- 在《中国互联网协会反垃圾邮件规范》中**垃圾邮件**被界定为
  - 收件人**事先**没有提出要求或者**不同意**接收的广告、电子刊物以及各种形式的宣传邮件
  - 收件人**无法**拒收的电子邮件
  - **隐藏**发件人身份、地址、标题等信息的电子邮件
  - 含有**虚假**的信息源、发件人、路由等信息的电子邮件

- **危害：**
  - 占用网络带宽，浪费网络资源，干扰邮件系统的正常运行
  - 对网络安全形成威胁
  - 浪费用户的宝贵时间和上网费用
  - 影响邮件系统的正常运行



解决垃圾邮件的方法和手段

- **反垃圾邮件立法**
  - 中国互联网协会反垃圾邮件协调小组2004年2月发出关于加快“反垃圾邮件立法”进程的倡议
  - 但立法面临着一系列的问题：
    - 垃圾邮件很难界定
    - 法律的执行问题
- **利用垃圾邮件过滤技术**
  - 基于IP、域名和路由的过滤
  - **基于内容的过滤**
  - 基于行为的过滤

基于IP、域名和路由的过滤

- **白名单、黑名单**
  - 白名单：用户设置和维护白名单，从白名单来的邮件都被认为是合法邮件。
  - 黑名单：与白名单相反，信源IP在黑名单上的邮件将会被拒收。目前比较流行的是实时黑名单（RBL）技术。
- **在服务器端进行配置**
  - 如对Access Control List、主机路由表等服务器端进行配置。可用DNS MX记录查找、反向DNS查找、新反向查找等方法。
- **安全认证方法**
  - 典型的是Yahoo提出DomainKey技术和Microsoft提出的SenderID技术。

基于内容的过滤

- 通过分析邮件的**内容**，来过滤垃圾邮件的一种技术。它的准确率较高，是当前解决垃圾邮件的**主流**技术之一，也是研究的**重点** 。
- 本质上，垃圾邮件的过滤属于**文本分类**的一个**二值**问题：将邮件分为垃圾邮件和正常邮件。因此很多文本分类的方法都可以应用于垃圾邮件的过滤。
  - 基于**概率统计**的内容过滤方法
  - **基于规则的内容过滤方法**

基于内容的过滤

基于概率统计的内容过滤方法

- 通过分析邮件的内容，计算其属于垃圾邮件的**概率**，从而进行分类的方法。训练过程是一个**统计学习**过程，得到相应的**分类器**，如贝叶斯分类器。
- 贝叶斯分类器具有较强的**分类**能力，但其过滤准确性依赖大量的**历史数据**。在某个训练集上分类效果比较好，可能在另一个训练集上分类效果不是很好。

基于规则的内容过滤方法

- 通过**训练集**得到一些过滤**规则**，然后基于这些规则对邮件进行评分，来决定邮件是否为垃圾邮件。
- 该方法需要**更多**的过滤规则并且这些规则需要经常的**更新**来更准确的判断出新类型的垃圾邮件。

基于行为的过滤

- **过滤群发软件所发送的邮件**
  - 跟据邮件头中的信息，判断是否群发软件，若为群发软件所发送的邮件，则判定为垃圾邮件。
  - 优点在于快速简单，缺陷在于一则需要不断跟踪新的群发软件；二则某些正当的邮件可能也通过群发软件来发送，这样造成了误判。
- **流量控制（速率控制）**
  - 在MTA端对邮件来源进行监控，若某个IP的流量在短时间内超过一定范围，则认为该IP对应的服务器很可能在发送垃圾邮件。
  - 该技术效果明显，但同样可能存在误判。

目 录

本文工作

- 设计并且实现了一种用**神经网络**方法来**优化**过滤规则的反垃圾邮件系统。
- 系统是针对基于**规则**的方法进行**优化**改进，针对传统过滤规则的静态性，使用**BP单神经元**神经网络，**自动提取**和**学习**垃圾邮件的改变特性，然后对过滤规则进行**优化**。
- 通过对比实验，显示该反垃圾邮件系统在英文语料集上有非常好的分类效果。

目 录

神经网络算法

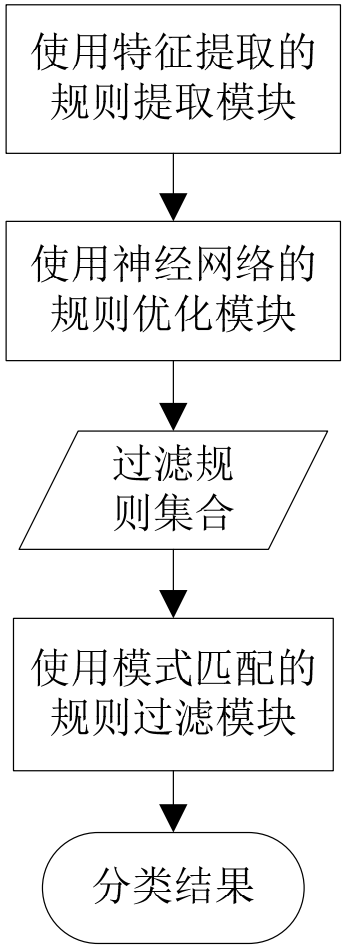
- 神经网络的应用非常**广泛**，主要的研究工作集中在网络模型、算法研究和生物原型研究等方面。
- 神经网络的最大优点是具有很强的自学习功能和自适应能力，以及很强的**分类**能力。
- 本文使用这项技术来对反垃圾邮件系统中的**规则集合**进行**优化**，使之更符合用户的实际情况，更加准确的识别垃圾邮件。

系统设计

- 系统的任务是使用优秀的**特征选取**算法从很多的邮件样本中提取出特定的垃圾邮件特征，然后用**神经网络**算法来**优化**这些过滤规则，最后建立有效的垃圾邮件**规则集**。
- 这个规则集就可以用来对要分类的邮件进行判断了。
- 在分类之前，要设定一个适当的分类**阈值**，根据规则集对邮件评分，如果大于这个阈值，该邮件就被识别为垃圾邮件，否则，被标识为非垃圾邮件。

系统模块

- **规则提取模块**
  - 使用特征选取算法从垃圾和正常邮件集中提取出垃圾邮件的特定属性
- **规则优化模块**
  - 上个步骤得到的垃圾邮件的特征由神经网络算法来进行分析，为每条过滤规则进行评分
- **规则过滤模块**
  - 当要为某封新邮件分类时，使用模式匹配算法扫描规则集，为这封邮件评分。



基于神经网络的反垃圾邮件系统总体设计图

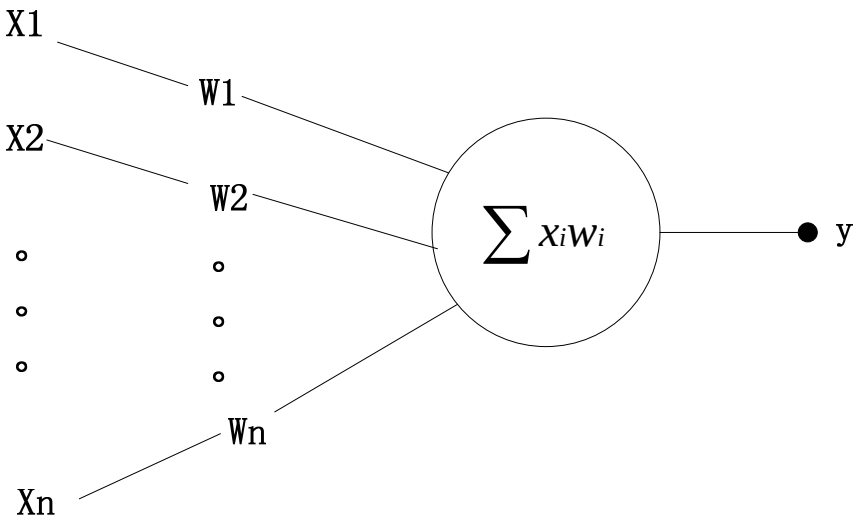
目录

规则提取模块

- 模块负责提取垃圾邮件特征，使用**信息增益**算法得到特征集中每一项的初始分数。
- 这些特征项按照分数进行**降序**排序。
- 最后，根据信息增益的变化，来决定选取多少个特征项。
- 系统主要处理**英文**邮件。主要步骤如下：
  - 通过**空格**把一封英文邮件的头部，正文和附件（文本）分隔成单独的词
  - 删掉出现频率**过低**或者**过高**的单词，如“a”，“the”，“and”，这些单词对邮件的分类几乎没有作用
  - 使用**支持向量机**模型来表示新的邮件
  - 通过**信息增益**算法，得到特征集中每一项的信息增益。结果按照**降序**排序。后面的分类中选择具有**高分数**的特征项。

规则优化模块

- 系统选用了单神经元**BP**（back-propagation）神经网络。**BP**网络，即误差反向传播算法的学习过程，是现在应用的比较广泛的神经网络算法之一，具有较强的信息处理能力。
- 其基本算法如下：W_i 是神经元的权向量，X_i 是神经元的输入向量。



单神经网络算法

规则优化模块

- 邮件被表示成为具有**n维空间**的**向量**。向量中的一个分量实际上表示的是每条规则。神经元的净输入：

$$I = a + \sum_{i=1}^{i=n} x_i w_i$$

w_i 代表的是规则的分数，邮件中如果出现相应的这条规则，χ_i 为1，如果没有出现，χ_i 为0。a为偏置。

- 将logistic函数作用于净输入之上，即神经元的输出为：

$$y = \log sig(I) = \frac{1}{1 + e^{-I}}$$

规则优化模块

- BP网络的权值优化算法为：

$$Wi = Wi + (y - (1 - y) * (\exp ected\_y - y)) * Xi * \partial$$

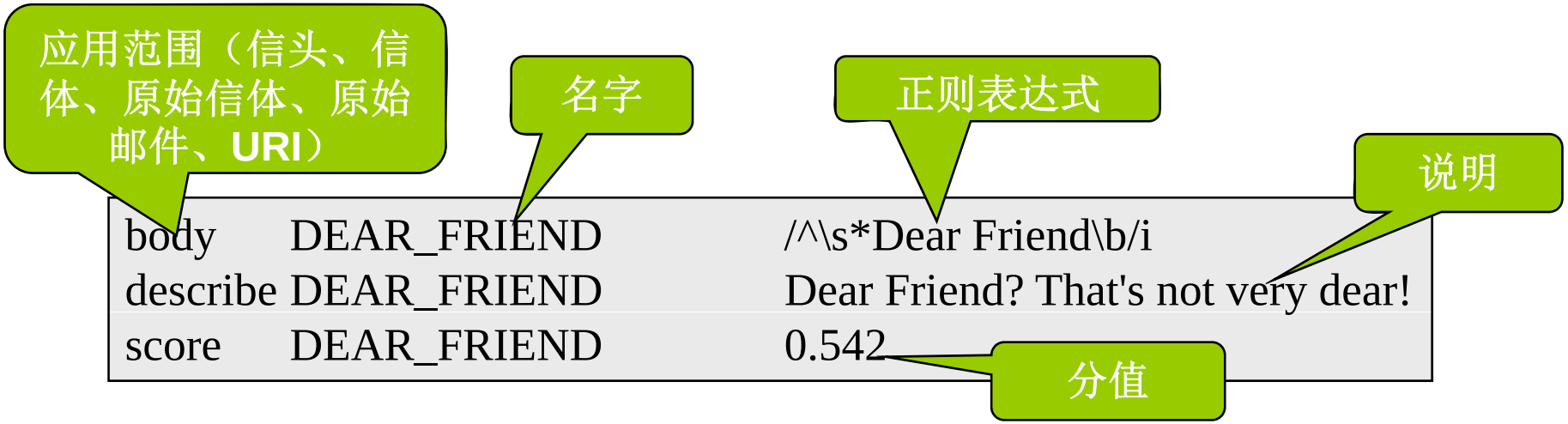
其中正常邮件我们取 $\exp ected\_y = 0$, 垃圾邮件取 $\exp ected\_y = 1$;

∂ 表示学习速率， ∂ 过大或者过小对系统都不好。∂ 过小网络收敛慢，过大则会不稳定。我们把学习速率设定为 2/(n+1)，n为在先前的训练集中正确分类的邮件数。

- 训练集里的邮件被分成 **10** 份，**9** 份作为BP网络里的**训练集**，剩下的作为**测试集**。
- 经过训练后，就更新优化了规则集里的每条规则，就得到了优化后的过滤规则集，就可以使用它们来进行分类。
- 然后，使用剩下的**1**份（包括垃圾和正常邮件）作为**测试**数据来验证垃圾邮件过滤系统的分类性能。
- 这样的过程重复**10**次，每一份邮件都作为了测试数据进行验证。

规则过滤模块

- 在系统中，有些规则是针对**邮件头**的，有些规则是针对**邮件主体**的。这些规则在系统运行过程中，还会**实时**的被优化。规则按照类别的不同，放在**不同**的文件中。
- 分类时，系统会用**模式匹配**算法去匹配相应的规则，把匹配规则的分值进行**相加**，就得到了这封邮件的一个**总的分数。**
- 系统预先设定了一个分类的阈值，该总分数和阈值进行比较，**大于**阈值，则被分类为**垃圾邮件。**



目 录

评价指标及语料集

- N_{ham}表示被系统分类为**正常邮件**的数目，N_{spam}表示被系统分类为**垃圾邮件**的数目。$n_{Y \to Z}$表示属于类别Y的邮件**被系统判定为**属于类别Z的邮件数目

  $(\ Y, Z \in \{spam, ham\}\ )$

$$精确率 = \frac{n_{ham \to ham} + n_{spam \to spam}}{N_{ham} + N_{spam}}$$

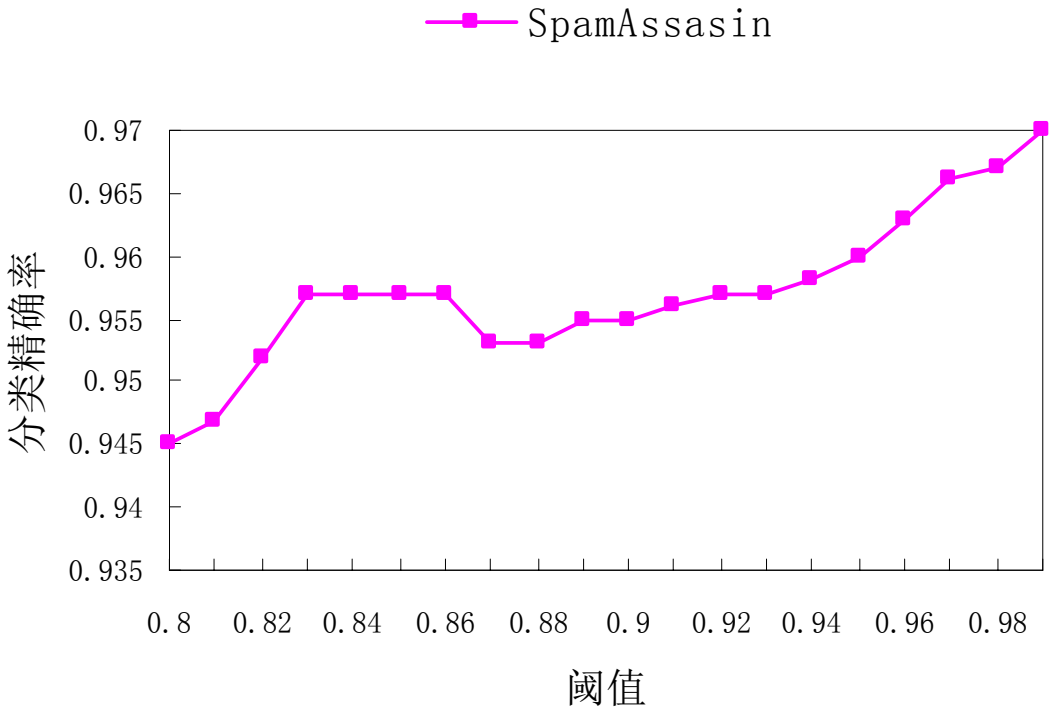
$$漏报率 = \frac{n_{spam \to ham}}{N_{spam}}$$

$$误判率 = \frac{n_{ham \to spam}}{N_{ham}}$$

- 英文语料集来自http://www.spamassassin.org，共有**6047**封邮件，包括**4150**封**垃圾邮件**和**1897**封**正常邮件**。

阈值对系统精确率的影响

- SpamAssasin语料集中分别使用**500**封**正常邮件**和**400**封**垃圾邮件**，阈值从**0.8**变化到**0.99**，步长为**0.01**来测试系统性能。



- 精确率最大时对应的阈值为**0.99**。

性能测试

- 系统和SpamAssassin(一个著名的反垃圾邮件系统)进行实验对比。
- 用SpamAssassin语料库中的900条邮件作为实验样本。
- SpamAssassin使用默认规则，我们系统使用的规则是经过神经网络优化后的。

	精确率	漏报率	误判率
SpamAssasin	97.38%	4.8%	0.33%
规则优化后系统	98.61%	2.5%	0.22%

➢ 系统精确率得到了提高，同时漏报率以及误判率得到了降低

目 录

总结及下一步工作

总结：

- 本文设计和实现了一个反垃圾邮件系统，因为考虑到过滤规则的静态性，不能检测出新类型的垃圾邮件，系统用BP神经网络算法使得过滤规则能得到优化。
- 在英文语料集上，系统和SpamAssassin进行了分类的对比。
- 实验证明，我们系统不仅有很高的分类精确率，还降低了垃圾邮件的漏报率和误判率。

下一步工作：

- 在中文语料集上，对系统进行测试。
- 对中文邮件的处理最重要的方面即需要对中文分词。接下来拟采用N-gram方法来自动切分字词组合。

谢谢各位专家、老师和同学！