新软周报
http://www.nusoft.com.tw
Internet Security Fighter
2007/9/17- 2007/9/23 第43期

## 多功能UTM / MS系列报导

### 技术浅谈与应用 - 为何初架设的新软UTM，其贝氏过滤资料库中并无数据

现今网络垃圾邮件泛滥成灾，是所有电子邮件使用者的梦魇。数年前，垃圾邮件顶多只是让收件者感到麻烦，现在却成为耗损企业生产力的严重问题；许多垃圾邮件内容不只不堪入目且无实际效益，还会严重占用企业网络带宽，同时浪费公司邮件服务器主机的储存空间与运算效能。也就是因为垃圾邮件的泛滥已经开始拖累经济发展（虚耗企业成本、占用网络带宽...），并且延伸出许多社会问题（网络钓鱼、情色广告...）。因此各国政府开始针对垃圾邮件订定各种法案来遏止其泛滥，但至今仍无太大作用。在还未有强效解决方案之前，垃圾邮件与反垃圾邮件的战争只会越演越烈。

为了协助企业防堵垃圾邮件，许多信息安全公司推出各种反垃圾邮件机制。但道高一尺，魔高一丈，在进步的不只是反垃圾邮件机制，垃圾虫（Spammer，指滥发垃圾邮件者）也在研拟更新的垃圾邮件发送方法，使得部份较早推出的反垃圾邮件机制逐渐失去舞台（例如：垃圾邮件黑名单－RBL）。但是，也有种反垃圾邮件机制在这场「垃圾邮件大战」中仍能历久弥新，并广泛被应用于各种反垃圾邮件产品中－「贝氏过滤法」（Bayesian Filtering）。

「贝氏过滤法」是利用「贝氏定理」为基础而推出的机制。「贝氏定理」是在公元1763年贝士（Thomas Bayes）的遗着中所发现。这个历史长达两世纪的古老定理是透过“事前机率”与“条件机率”，来计算出“事后机率”，常常运用在统计学当中。这定理虽为古老，但它所延伸出来的「贝氏过滤法」却对付垃圾邮件却出奇的好用。

「贝氏过滤法」的运作方式是将整封信件（含信件Header、信件内文...，附加文件除外）分割成一个个单一词句（Token），再利用特定算法分析每个词句的出现机率给予信件评分。最后，以信件分数的多寡来评断该信件是否为垃圾信件。

上述「贝氏过滤法」之运作，所依赖评分的标准就是拥有大量词句的「贝氏过滤数据库」。「贝氏过滤数据库」分为两个－“垃圾邮件数据库”与“非垃圾邮件数据库”。“垃圾邮件数据库”专门存放各种垃圾邮件用词，而“非垃圾邮件数据库”则保存了企业往来信件中的常用词。「贝氏过滤法」就是在这两个数据库中比对每个词句之出现机率，来判断该信件是否为垃圾信件。

有一点要注意的是，新软系统产品－多功能 UTM 在出厂时，「贝氏过滤数据库」上并无任何数据，因此「贝氏过滤法」在一开始时并无法正常使用。唯企业以“辨识学习”方式，提供「贝氏过滤数据库」正常信件、垃圾信件各 200 封后，方可正常运作。

为甚么新软多功能 UTM 在出厂时就不事先提供预设的「贝氏过滤数据库」呢？还要客户提供正常信件、垃圾信件学习不是很麻烦吗？其实主要原因有以下两点：

1. 信件是否属于垃圾邮件是依收件者的主观认知来判定；就同一信件而言，有可能每位使用者、企业的判断都不同，更何况是处于模糊地带的“电子报”。依目前情况，绝大部分“电子报”皆属于推销产品的垃圾邮件，但是假如收件者刚好有此种产品之购买需求，他也就不会把这封“电子报”视为垃圾邮件。
2. 多功能 UTM 在出厂时并没有企业往来邮件的正确数据。倘若「贝氏过滤法」使用预设的「贝氏过滤数据库」来过滤信件时，将有可能导致正常信件被误判为垃圾信件的情况发生。

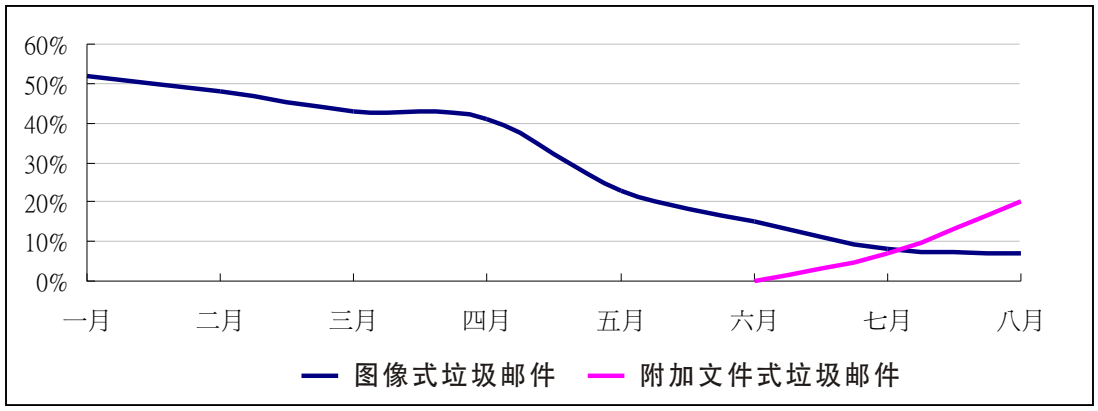
因此企业如须启用「贝氏过滤法」来滤除垃圾信件，就必须先用“辨识学习”教会「贝氏过滤法」甚么是垃圾信件，甚么是非垃圾信件。往后如有误判的情况发生，也可使用“辨识学习”矫正此错误。当然，企业提供的数据越多，「贝氏过滤法」也会越精确，甚至可高达九成九以上的辨识率！！（理想的“辨识学习”比例 － 垃圾邮件：正常信件 ＝ 2：1）

文 黄赞中 isaac@nusoft.com.tw

## 市场营销报导 - 垃圾邮件新趋势：附加文件形式垃圾邮件

在最近一两个月来，您是否会收到一些奇怪的信件？没有邮件内容，却夹带了PDF、Excel、Zip...文件，打开来看看，却发现里面都是广告！！您猜的没错，这就是目前开始流行的新型态垃圾邮件－“附加文件式垃圾邮件”。

在年初所流行以图片方式（jpg、gif...）散布垃圾邮件讯息的“图像式垃圾邮件（Image Spam）”，由于垃圾虫（Spammer，指滥发垃圾邮件者）使用了各种“防机器判读技术”替图片加料（文字变形、变色、杂点...）以躲避反垃圾邮件系统的查缉，使得“图像式垃圾邮件”一度泛滥严重，而无法阻拦。这问题在各种新式反垃圾邮件系统陆续推出之后得到解决，使得“图像式垃圾邮件”的比例持续减少，渐渐取而代之的就是“附加文件型式垃圾邮件”。




图一 图像式垃圾邮件 与 附加文件式垃圾邮件 占全体垃圾邮件比例

“附加文件式垃圾邮件”通常夹带着Word、Excel、PDF、FDF（Adobe Acrobat窗体文件）...这几种常用文件类型的广告，或将这些文件以Zip压缩后再夹带于信中。就因一般反垃圾邮件机制不会针对附加文件查验，所以这些新式的垃圾邮件能够轻易的穿过各种反垃圾邮件机制。也因为“附加文件式垃圾邮件”通常伪装成正常邮件，所以相当容易促使收件者开启，达到广告宣传的目的。

为了因应这些变化多端之垃圾邮件，新软系统的垃圾邮件过滤机制采用多层架构过滤方式（又称为鸡尾酒式），以复合方法层层滤除垃圾邮件。其中，“垃圾邮件特征”便是专门处理新式垃圾邮件之一大利器；可有效滤除各种新式垃圾邮件，当然也包含“附加文件式垃圾邮件”。

每当有新型态垃圾邮件开始出现时，新软系统的工程师们便会分析其各项特征，汇整成「垃圾邮件特征码」，供新软系统产品「多功能UTM」及「邮件服务器」下载。藉此方式有效处理来袭的各式垃圾邮件，还给企业一个干净的电子邮件环境。

文 程智伟 rayearth@nusoft.com.tw