第 37 卷 第 17 期
Vol.37 No.17

计 算 机 工 程
Computer Engineering

2011 年 9 月
September 2011

·开发研究与设计技术· 文章编号：1000—3428(2011)17—0282—04 文献标识码：A 中图分类号：TN917

一种改进的安全传真服务器设计方法

陈鸿昶，于洪涛，冯晓磊

(国家数字交换系统工程技术研究中心，郑州 450002)

摘 要：传真服务器存在对垃圾传真防范能力不足的问题。为此，在现有安全传真服务器的设计基础上，提出一种改进方案。在接收传真之前增加图灵检测功能用于筛选自动传真，采用近似串匹配技术对垃圾传真进行确认，进而通过黑名单功能实现垃圾传真的过滤，并将聚类功能作为发现新垃圾传真、丰富关键词库的辅助手段。仿真实验表明，该方法在精度、对垃圾传真反应速度以及智能性等方面均优于原有设计。

关键词：安全传真服务器；垃圾传真；图灵测试；聚类；近似串匹配；黑名单

Improved Design Method of Safety Fax Server

CHEN Hong-chang, YU Hong-tao, FENG Xiao-lei

(National Digital Switching System Engineering & Technological Research & Development Center, Zhengzhou 450002, China)

【Abstract】Due to the problem of inefficient junk fax prevention, this paper proposes an improved prevention scheme based on the design methods of existing safety fax server. The basic principle is that screen automatic fax by adding Turing test function before receiving faxes. Through approximate string matching technology, junk faxes are detected, and then filtrated by blacklist function. Furthermore, the proposed system regards clustering function as an assistant method to discover new junk fax and rich keywords library. Experimental results show compared with the original design, this method has higher accuracy, faster detection speed, less manual intervention and stronger usability.

【Key words】safety fax server; junk fax; Turing test; clustering; approximate string matching; blacklist

DOI: 10.3969/j.issn.1000-3428.2011.17.095

1 概述

近年来，随着传真卡、传真服务器[1]等新型传真设备的出现，传真业务的用户群体及应用范围得到了极大的拓宽。但是，同时也带来了一种新的社会公害——垃圾传真[2]，严重影响了人们的日常学习和工作。为此，垃圾传真检测技术正成为近年来业界研究的热点。

基于用户定制的黑白名单[3]控制方法根据收到的传真内容来判断其是否为垃圾传真，并根据判断结果设置黑白名单，进而实现呼叫控制。但该方法存在 2 个缺点：黑白名单的设定必须准确，否则很容易导致误判；黑白名单的设定存在被动性和滞后性，需要不断地更新和维护，并且通常无法涵盖所有的情况。因此，黑白名单的可用性不高。预览法将接收的传真保存为电子文档图像，并进行人工预览，如是正常传真则保存并打印出来。但是该方法目前都采用人工方式，效率低下。文献[3]提出安全传真服务器的概念，采用聚类技术发现传真服务器中接收的广播式垃圾传真，并在分发前将其滤除，从而实现其分发传真的安全性，该方法以聚类方式实现，需要一定数据的积累，在垃圾传真出现的前期或者垃圾传真稀疏分布时，有效性受到限制。

由对垃圾传真的统计分析可知：绝大多数垃圾传真通过自动传真设备发送，因此垃圾传真的过滤重点应是对此类传真。本文根据此指导思想，针对现有方法存在的缺陷，设计了一种改进的安全传真服务器实现方法，通过对自动发送传真的内容识别确定垃圾传真，通过垃圾传真发送号码的识别实现垃圾传真的过滤。

2 系统总体结构

本文方法的系统结构如图 1 所示。

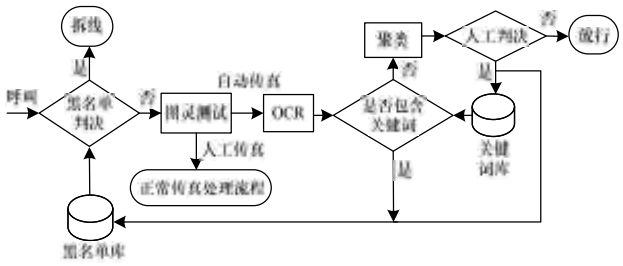

图 1 安全传真服务器结构

3 基于信令的垃圾传真过滤

ITU-T 在 T.30 中详细规定了电话网传真通信的过程和信号方式，其中，规定传真通信过程分为 A、B、C、D、E 5 个阶段，如图 2 所示。

图 2 传真通信的 5 个阶段

在图 2 中，阶段 A 为呼叫建立阶段，在此阶段，双方经传输线连通，传真机接入线路；阶段 B 为报文前阶段，在此阶段进行报文传输前的准备，包括传真机功能的辨别和选择、传送模式的确立、相位核对、发送线路质量测试等；阶段 C

基金项目：国家“863”计划基金资助项目(2008AA011001)

作者简介：陈鸿昶(1964－)，男，教授，主研方向：电信网安全，信息系统；于洪涛，高级工程师；冯晓磊，硕士研究生

收稿日期：2011-05-24 **E-mail：**15937101921@139.com

为报文传输过程，在此阶段完成传真报文的发送或接收；阶段 D 为报文后阶段，在此阶段，收发双方传真机均要对报文传输是否结束、接收的情况如何、是否还有报文，以及传真过程是否结束等作为出判断，以便确定下一步是转向 E 阶段，还是返回 B 阶段；阶段 E 为呼叫释放阶段。

传真服务器通常为自动接收模式，在接收到呼叫信令后自动应答，随后转入传真规程中的 A 阶段，本方法在转入 A 阶段之前增加信令判决功能，将接收到信令中的主叫号码与黑名单库进行比对，如果是黑名单，则送出拆线信令结束通信过程，否则进入后续流程。黑名单库通过后面的方法进行维护。

4　基于图灵测试的自动传真筛选

图灵测试(又称“图灵判断”)是图灵提出的一个关于机器人的著名判断原则，其方法是通过随意提问以测试机器是否具备人类的智能。图灵测试方法的有效性目前已被多数人承认，本文用该方法对自动传真设备发送的传真进行筛选。

具体方法是在传真服务器建立话路后，转入 A 阶段之前，不立刻发送传真信号，而是向主叫播放提示语音，如“您好，这里是××公司，发送传真请拨 1，否则请挂机”，然后等待接收主叫选择。人工发送传真时会按照提示拨号，接收端收到预期的号码后认为本次传真由人工发送，后续处理按照正常传真过程进行；而自动发送设备由于不具备人的智能，无法对语音提示作出判断及执行相应的动作，因此接收端如果在设定时限内没有接收到“1”的双音信号，则认为是自动传真，虽然照常发送传真信号，但是后续处理进入特殊流程，即内容识别、聚类分析等。

5　基于关键词匹配的垃圾传真内容识别

由于本文方法只关注自动传真，而自动传真通常版面规整、可辨认度高，因此本文采用文种识别基础之上的光学字符识别(Optical Character Recognition, OCR)技术，将传真图像准确地转换为文本文档，在此基础上，设计一种基于 Trie 规则的关键词匹配方法实现对垃圾传真的精确判定。具体过程如图 3 所示。

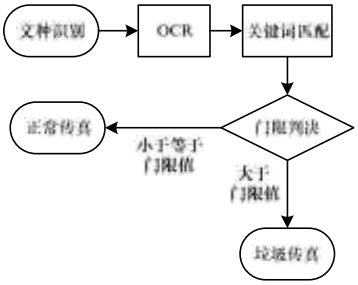

图 3　垃圾传真内容识别过程

5.1　文种识别

一般来说，预先指定文种的 OCR 识别效果比未知文种的识别效果要好很多，因此，本文在对传真图片进行 OCR 识别之前，对其进行文种识别。对于企业传真来说，大部分的传真采用中文和英文，其他文种是极少数，本文不予考虑。

中文常被称为方块字，具有长宽比接近于 1 的特点，其书写格式及排版有相同的特性[4]。而且汉字在印刷排版时，保持“单摆浮隔”(即每个汉字所占宽度相等、相邻字符保持间隔)式排版，文本行具有明显的全局特性——字符中心的等间距性(如图 4 所示)。这一特性不受字体与风格的影响，即当同一文本行中出现不同字体、不同风格的汉字时，字符中心等间距性仍然维持，不同字号的汉字在同一文本行中出现的概率很小，本文不对这种情况进行分析。

图 4　中文字符的中心等间距性

英语作为一种拼音文字，每个单词由数量不等的字母组成，而每个字母的宽度也不尽相同，因此英语文字不具有中文字符所具有的等间距特性(如图 5 所示)。

图 5　英文单词的间距

本文正是利用中文与英文字符的间距特性，将中英文传真区分开来。由上述分析可知，同一份中文传真各行连通域的平均间距相差不大，即方差小；而对英文来说，由于其连通域的中心间距相差较大，因此其方差相应的也较大。因此，采用连通域中心间距的方差可以有效区分中英文传真。另外需要说明的是，本文所指的连通域与传统意义上的连通域不同，本文所说的连通域是指一个中文字符或英文单词。

文种识别的具体过程如下：

对选定的文本区域进行连通域搜索，得到基本的连通区域，如图 6 所示。

图 6　连通域图

在图 6 中，每一个连通域都可以用一个矩形区域表示，$rect(x_1, y_1, x_1', y_1')$ 表示第一个连通域的矩形区域，其中，(x_1, y_1) 代表矩形的左下角坐标；(x_1', y_1') 代表矩形的右上角坐标，则有矩形的高度 $height = y_1' - y_1$，矩形的宽度 $width = x_1' - x_1$，以每一个文本行左上角为原点建立直角坐标系，x 轴向右为正，y 轴向下为正，遍历每一个矩形域，连通域中心点的坐标 (x_{ic}, y_{ic}) 为：

$$x_{ic} = \frac{x_i' - x_i}{2} \tag{1}$$

$$y_{ic} = \frac{y_i' - y_i}{2} \tag{2}$$

建立一个平面坐标系，取字符中心点在行中的水平位置作为数据点的 x 坐标，取连通域中心点与前一连通域中心点的水平距离作为数据点的 y 坐标。设 b 为一个文本行的平均间距，则：

$$b = (\sum_{i=1}^{n} y_i)/n \tag{3}$$

在采集数据点时，为了尽量减少非汉字的干扰，用以下规则对数据点进行过滤：(1)滤除宽度明显大于其他连通域的英文单词；(2)滤除宽度和高度都比较小的标点符号所产生的数据点。

设第 i 行连通域的平均间距为 b_i ($i = 1, 2, \cdots, n$)，则 n 行的平均间距为：

$$\bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i \tag{4}$$

方差为：

$$\delta = \frac{1}{n}\sum_{i=1}^{n} (b_i - \bar{b})^2 \tag{5}$$

设 ε 为统计所得的平均间距的方差阈值，当 $\delta \leqslant \varepsilon$ 时，判定为中文；当 $\delta > \varepsilon$ 时，则判为英文。

5.2 基于 Trie 规则的关键词匹配

通过 OCR 将传真图片转换成文本文档后，为判断该文档是否包含关键词库中的词组或短语，需要采用字符串匹配技术搜索整个文档[5]。而在 OCR 识别中，经常出现被错误识别的字符，或者在发送方处理文字时也经常存在一些拼写错误的情况，因此本文采用近似串匹配技术确认传真属性。

目前常用的计算字符串相似度的方法包括：编辑距离，混淆字符集，混淆字符矩阵以及 n 元相似度算法，其中，编辑距离法应用最为广泛[6-7]。由于待查字符串中包含各种错误，而且这些错误可能发生在字符串的任意位置，因此不能利用现有的数据库索引技术提高查询速度。例如，在 OCR 文本识别中，常将“item”识别为“ltem”。如果使用数据库索引查询，则找到的最近位置在以 L 开头的第一个词。该位置距离数据库中 item 的位置非常远。

Trie 是一种树形结构，用于保存大量的字符串。它的优点是：利用字符串的公共前缀来节约存储空间。在 Trie 结构中，每一个节点由一个字母和指向其子节点的指针组成，根节点不包含字母。一个公共前缀就是由根节点到树中一个节点的路径所经过的所有节点的字母的序列。例如，表示 t、to、te、tea、ten、a、ar、are 这 8 个字符串的 Trie 结构仅占用 8 Byte(不包含指针)，见图 7。为了辨认出 Trie 结构中的一个独立的字符串，需要在表示该字符串的路径末端的节点上加一个位标志，表明到该节点为止已经形成了一个完整的单词。

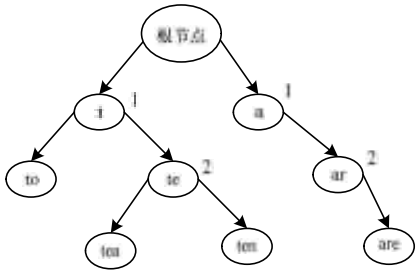

图 7 Trie 树结构示意图

本文利用 Trie 数据结构与编辑距离法结合的近似字符串匹配算法，可以比较好地解决效率问题。同时该方法也适用于带通配符的近似字符串查询，这为更精确地定位近似字符串提供了一种方便的手段。

本文使用 Trie 树状结构，并非仅因为它节约存储空间的能力，还因为它可以很好地与编辑距离的计算结合起来，有效地减少相似字符串匹配的计算量。在计算编辑距离的动态表中，列 j 单元值的计算，实际是目标字符串前 j 个字符组成的前导子字符串与待匹配字符串的编辑距离的计算。因而对有共同前缀的目标字符串，它们的动态计算表的前几列的值是相同的。例如，计算上图中“tea”和“ten”与“tra”的编辑距离的过程中，子字符串“t”、“te”与“tra”的编辑距离是相同的，见图 8。

	t	e	a
t	0	1	2
e	1	1	2
n	2	2	1

(a)ten 与 tea 的编辑距离

	t	e	a
t	0	1	2
r	1	1	2
a	2	2	0

(b)tra 与 tea 的编辑距离

图 8 基于 trie 结构的编辑距离计算示例

从图 8 中可以看出，t、e 2 列的计算是相同的。即，计算 tea 与 ten 的动态表的前 2 列可以直接拷贝到计算 tea 与 tra 的动态表的前 2 列中。对其他以“te”开头的字符串与 tra 的编辑距离计算也可以同样处理。这样，在通过遍历 Trie 树时，计算每一个节点的前缀字符串与待匹配字符串的编辑距离，并把计算的结果保存在一个统一的表中。有相同的前缀的字符串的编辑距离可以共享前面部分列的计算结果，减少了编辑距离的计算量。

此外，对于本文设定的最大可接受的编辑距离 k(即最大允许误差为 k)，如果某一列的所有单元都>k，则以该列之前的前导字符串为开头的所有字符串与待匹配字符串的编辑距离都不可能小于或等于 k。而这些字符串的后续字符串与待匹配字符串的编辑距离一定大于和等于 1，因而它们与待匹配字符串的编辑距离一定大于 k。在这种情况下，可以中止以这一前导字符串为开头的 Trie 结构中的子树的编辑距离计算。这样在很大的程度上缩小比较的范围，从而进一步提高字符串搜索的效率。

这 2 个方法构成了以 Trie 树为基础的相似字符串搜索算法。它包含 2 个部分：

(1)遍历 Trie 树上某一节点的子节点的递归过程 Trie-Search。在该过程中，可能因为其编辑距离超过最大可接受的编辑距离 k 而中止对一个子树的遍历。同时，当对一个子树的遍历到达一个有字符串终止标志的节点，而且编辑距离<k 时，该节点前面路径所代表的单词成为一个候选单词。

(2)计算一个节点的编辑距离 EditDist。根据编辑距离的算法，只需要将该节点的字符与待匹配字符串的各个字符进行比较，再综合动态表的前面列的计算结果(被统一保存在一个表中)，就可以计算当前列的所有单元的值。

下面给出待匹配字符串在词典的 Trie 树结构中的遍历匹配算法：

(1)输入待匹配字符串 s。

(2)从词典的 Trie 树结构中从顶层开始逐层向下搜索。计算当前节点与字符串 s 的编辑距离。对于顶层，由该层的第 1 个节点开始。如果得到的编辑距离：

1)小于或等于最大允许编辑距离，则重复本步骤以对该节点的每一个子节点进行匹配。

2)大于最大允许距离，终止以当前节点为根的子树的匹配。转而比较当前节点所在层的下一个节点。重复本步骤。

3)如果该节点为末端节点，则找到一个目标字符串。输出该末节点代表的词到候选词集中。比较当前节点所在层的下一个节点。重复本步骤。

(3)当所有子树都被遍历或终止后，全部算法结束。

通过上述方法得到文档中所包含的关键词的个数和同一关键词的词频，本文给出垃圾传真的判决目标函数为：

$$J = \sum_{i=1}^{n} p_i \tag{6}$$

其中，n 为待识别文档中包含的关键词个数；p_i 为第 i 个关键词出现的次数。通过训练给出判决门限 T，当 $J > T$ 时，则判定为垃圾传真，反之，则为正常传真。

通过上述方法，可以准确检出垃圾传真，从而实现对黑名单库的自动维护。

6 基于聚类的自动广播传真识别

垃圾传真通常是将一份传真进行广播式散发，因此在传真服务器的接收端，垃圾传真重复严重，而正常传真很少重

复。所以，一种可能的筛选方法是对接收到的传真进行聚类处理[8]，这种方法可以得到传真服务器接收到的所有广播式传真，其中既包含垃圾传真，也可能包含部分用户期望的或者有利用价值的广告传真，因此需要对筛选出来的传真进行区别对待。

对于不包含特定关键词的自动传真，进行聚类处理，从中发现广播传真，聚类可采用多种特征，采用文献[1]中的方法也可以取得较好的效果。大多数的自动广播传真是垃圾传真，但仍有少部分为订单、会议通知等，为保证不发生错误，这里采用人工方法进行判决，当发现是垃圾传真，则将其主叫号码加入黑名单库，同时从该传真中凝练提取新的关键词加入关键词库。

只对自动传真进行聚类，减少了人工传真对聚类精度的影响，同时大大降低了聚类处理的数据量，从而提高了聚类效率。

7　实验与结果分析

由于系统通过图灵测试方法可以精确检出自动传真，因此，系统对垃圾传真的过滤性能取决于对自动传真中垃圾传真判别的精度。这里随机选择 3 000 个自动传真作为测试样本，设定金融诈骗类传真为垃圾传真，因此，设关键词为“致富”、“发财”、“中奖”、“抽奖”等，通过人工检查，标注传真属性为垃圾传真或非垃圾传真，再由本系统中的内容识别方法进行自动判决，两者进行比对，得到如表 1 所示的实验结果。

表 1　实验测试结果

垃圾传真		非垃圾传真	
正确判决	错误判决	正确判决	错误判决
127	0	2 871	2

从实验结果可以看出，由于系统只对自动传真进行内容识别，而自动传真版面规整，可辨认度高，加上采用基于 Trie 规则的近似串匹配技术，因此具有较高的识别准确率。

8　结束语

本文在文献[3]的基础上，提出了一种改进的安全传真服务器设计方法，通过多种技术的有机结合，实现了对传真服务器的高效、安全防护。本文重点论述了系统整体结构及其基本实现方法，但未对其中的关键词匹配、聚类技术做深入探讨。因此，在此系统框架内提高关键词匹配、聚类 2 项技术的准确性及效率，从而提高系统的整体性能，是下一步研究的重点方向。

参考文献

[1] 赵建涛，王　颖，任　俊. 智能传真服务器系统的设计与实现[J]. 华北电力大学学报, 2006, 33(4): 59-61.

[2] Quinten V M. Analysis of Spam over Internet Telephony Protection Techniques[EB/OL]. (2007-11-10). http://www.vf.utwente.nl/meentr/research/dl.php?eunice2007-quinten-meent-pras.pdf.

[3] 于洪涛，黄　海，冯晓磊. 一种安全传真服务器的设计与实现[J]. 电子技术应用, 2010, 36(12): 28-30.

[4] 王　恺，王庆人. 中英文混合文章识别问题[J]. 软件学报, 2005, 16(5): 790-791.

[5] 苏金树，张博锋，徐　昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 17(9): 1848-1850.

[6] 范立新，谢晓能，吴　飞. 基于过滤的中文多模式近似字符串匹配算法[J]. 计算机工程, 2006, 32(20): 48-50.

[7] 郭牧怡，刘　萍，谭建龙. 基于文件标题特征的网络视频去重研究[J]. 计算机工程, 2010, 36(9): 227-229.

[8] 张建辉. 基于层次划分的最佳聚类数确定方法[J]. 软件学报, 2008, 36(9): 48-52.

编辑　任吉慧

(上接第 281 页)

进行统计分析，得到两者之间的关系，实验结果如图 5 所示。

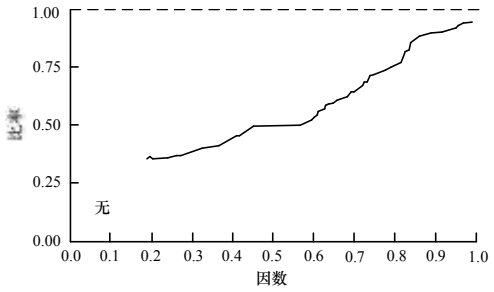

图 5　因数对 2 种算法搜索时间之比的影响

由图 5 可知，文中虚线为水平参照，曲线为限定区域搜索时间与 Dijkstra 搜索时间的比值，对比同一搜索范围内，对于不同因数限制区域搜索算法与 Dijkstra 算法的搜索效率结果不同，图标“无”表示在某一因数下没有搜索到路径。当因数小于 0.2 时就会出现找不到路径的情况，而当因数达到 1.0 时，限定区域算法可搜寻到最短路径，运算时间与 Dijkstra 算法相同。因数对运算时间和精度所起的作用是在能搜寻到最短路径的情况下，因数越小会使运算时间越短，会降低精度。

本实验在只使用限制搜索区域算法的情况下，验证不同的因数值对搜索结果的影响，在实际中，因数的选择也依赖于地图，不同的地图或者地图的不同的地区将具有不同的最小因数。

6　结束语

本文提出一种将静态存储与动态搜索相结合，以限制区域搜索算法为主，A*算法为补充的最短路径搜索算法。根据港区道路网络空间的分布特性，设计针对该特性的特殊路径搜索算法。实验结果表明，该算法的静态路径存储功能能够提高港区卡口通关的作业效率，其限制区域搜索算法能最大限度的降低时间和空间的复杂度，提高运行效率，由于该算法具有不完备搜索的特性，因此用 A*算法作为补充，以提高算法的安全性。

参考文献

[1] 傅清祥，王晓东. 算法与数据结构[M]. 北京：电子工业出版社, 1999.

[2] 王元彪. 智能交通系统中 Dijkstra 算法的高效实现[J]. 计算机工程, 2007, 33(6): 256-258.

[3] 李志发，邵伟民，卢志强. 基于 ArcGIS Engine 的城市公交换乘系统[J]. 计算机工程, 2010, 36(11): 55-57.

[4] Fu Mengyin, Li Jie, Deng Zhihong. A Practical Route Planning Algorithm for Vehicle Navigation System[C]//Proc. of the 5th World Congress on Intelligent Control and Automation. [S. l.]: IEEE Press, 2004.

[5] Goldberg A V, Harrelson C. Computing the Shortest Path: A Search Meets Graph Theory[R]. Microsoft Corporation, Tech. Rep.: MSR-TR-2004-24, 2003.

编辑　刘　冰