应用 TeX 控制功能和图像检测的文本密写

陈　超，王朔中，张新鹏

（上海大学 通信与信息工程学院，上海 200072）

摘　要：提出一种利用排版工具 TeX 在文本文件中实现密写的方法。该方法属于“文本－图像”方式，在 TeX 源文件中修改 TeX 命令，对一部分单词间空格进行微小的改变，从而将数据嵌入。经编译和转换，生成含隐蔽信息的 pdf 文件在互联网上传输。对隐蔽信息的提取在图像域中进行，根据水平和垂直投影分别检测文字的行和单词之间的空格，恢复出原嵌入数据。密写的安全性取决于从图像中准确检测空格微小差异的能力、文本文件相对于嵌入数据量的大小、采用的加密和编码技术等因素。

关键词：信息隐藏，文本密写，TeX，文档图像，安全性

中图分类号：TP391

Data Hiding in Text File Using TeX and Extraction of Hidden Data from Document Image

CHEN Chao, WANG Shuozhong, ZHANG Xinpeng

(*School of Communication and Information Engineering, Shanghai University, Shanghai 200072*)

Abstract: A steganographic technique using pdf files to carry secret data is proposed. Data are embedded into the control sequences in a TeX source file to slightly modify inter-word spaces in the text. Appropriate coding schemes may be applied. The coded TeX file is then compiled and converted to generate a stego-pdf file, which is convenient to be disseminated over the Internet. In reception, a jpg image is obtained from the pdf document, and the hidden data are extracted by detecting the spaces between words. Data security is determined by the ability of reliably detecting small difference in space widths, the size of the document file in terms of the amount of data to be embedded, and the encryption and coding techniques used.

Keywords: information hiding, text steganography, TeX, document image, security

密写是将秘密消息嵌入表面上正常的数字载体中，以不被他人察觉的方式通过公开渠道如互联网进行传输，接收者使用专门的工具和密钥从含密载体中提取消息，实现隐蔽通信[1]。密写将“正在通信”这一事实隐藏起来，含秘密消息的数字载体通常又与大量正常媒体混在一起，因此具有高度的隐蔽性。嵌入的数据事先可用密码技术加密，进一步提高了消息的安全性。

用于密写的载体以数字图像最为普遍，其次是数字音频，特别是他们的压缩形式[2,3]。数字视频也可用于密写，但由于视频文件尺寸过大，在网络上传输远不如图像和音频广泛，所以一般说来不是隐蔽通信的优选载体。

文本是网上传输数量最多的信息载体，其形式多样，如网页、电子邮件、包括 pdf，doc，ps 在内各种格式的文本文件等，用作密写载体具有广阔的应用前景。但与图像、音频等多媒体信号相比，在文本中嵌入额外信息而又不引起可察觉的变化相当困难，这是因为文本中可用于插入额外数据的冗余空间很小，因此利用文本作为密写载体的技术发展大大滞后。

Bender等人[4]将文本中嵌入信息的方法归纳为三类：1) 利用空格，例如单词之间空 1 格代表 0，空 2 格表示 1；2) 基于句法，包括句型和标点符号，例如句型a, b, and c表示 0，a, b and c表示 1；3) 利用语义，定义一个同义词表，例如big表示 0，large表示 1。其中第一类是基于文本格式的，后两类则是基于文字内容的。

在基于文字内容的嵌入技术中，利用句法的方法一般嵌入量很小，因为特定句型在文章中出现的次数有限。语义编码的问题在于会产生文本内容的改变，或者使语言变得不自然。例如，尽管 pacific 与 peaceful 同义，将 Pacific Ocean 变成 Peaceful Sea 却很容易引起警觉。

在基于格式的信息隐藏方面，某些利用增减空格的简单方法过于明显，或者难以抵抗文本编辑的攻击。例如在每一行末尾插入的一个或两个不可见空格，只要改变页面的设置就会被完全破坏。Brassil等人[5,6]提出了移词编码、移行编码、特征编码三种方法。由于单词间距并不固定，移词编码通常需要通过与原始文本的比较才能提取出嵌入的信息，用作水印尚可，用于密写则不可行。特征编码的一个例子是用字母b、d和h顶部的短横线长度来表示嵌入信息。虽然这些方法都会使原始文本发生轻微变化，只要采取适当的措施仍能达到外人很难辨认的程度。若正常文本的行距是均匀的，则移行编码嵌入数据的提取不依赖原始文本，因此既可用于保护版权，亦可成为一种隐蔽通信手段。此外移行编码通常比移词编码更为稳健，故一直受到研究者的关注[7,8]。Villan等人最近构建了一个新的理论框架，将文本中的信息隐藏看成Gel’fand-Pinsker问题的特例[9]。他们在这一框架下提出两种视觉隐蔽性优良且嵌入量大的基于特征文本信息隐藏方法：半色调量化和颜色量化，该方法从电子版本和打印硬拷贝中均能自动提取隐蔽信息。

在某些特定格式的文本文件例如XML和排版工具TeX的源文件中也能嵌入额外信息。此类文件中除了文字内容外，还包含大量控制符号或命令。改变控制符号的某些属性并不影响他们的功能，例如可在XML文件中嵌入附加信息而保持浏览器显示结果不变[10]，或者在TeX源文件中嵌入信息而保持生成的dvi和pdf文件不变[11]。

本文提出一种基于 TeX 的密写方法，通过修改 TeX 源文件中的控制命令格式，将隐蔽信息嵌入 pdf 文件，通过图像处理从文本图像中提取秘密信息。

1．基于 TeX 的信息隐藏

Knuth创造的TeX是一种应用很广的计算机排版语言[12]，特别是用于编排科学论文和著作时可得到优良的版面，是许多国际刊物和学术会议指定的排版工具。常用的文字处理软件如Word是在文件中直接设定字符和段落等的格式，所谓“所见即所得”就是在屏幕上立即反映所设置的格式。排版工具TeX则不同，使用者创建纯文本形式的源文件，其中包括文字内容和各种控制序列，由控制命令规定排版效果和具体格式。源文件经TeX程序编译，生成与设备无关的dvi文件，将dvi文件送到各

种输出设备如显示器和打印机，通过驱动程序给出最后的排版结果。也可以将dvi文件转换成pdf或PostScript格式，便于在互联网上传输或提供下载。随后Leslie Lamport又在TeX基础上发展了LaTeX[13]，提供大量的新功能，使用者不必再记忆繁琐的TeX命令，因此为更多的用户接受。本文将TeX及其一系列衍生软件统称为TeX。

利用 TeX 的特点可以将源文件本身作为密写载体。例如在 TeX 控制命令序列中插入空格不会影响编译生成的 dvi 文件，因而可在\usepackage[dvips]{graphics}中的[dvips]之前以插入或不插入空格分别表示 0 和 1，接收方直接从源文件中提取嵌入信息。我们称这种方式为“文本－文本”方式，例如[11]就采用了这种方式。

另一种方案是在 TeX 源文件中插入额外的命令，使编译结果发生微小的改变，例如对行距、词距、字符、空格大小等进行调制。将含隐蔽信息的源文件编译成最终文件形式（如 pdf）传递，代表嵌入信息的变化反映在 pdf 的行、单词、字符形式、空格等格式属性中。只要这种修改幅度很小就不会引人注意。接收方用 pdf 阅读器将文件转换为图像，即可从图像中提取隐蔽信息，因此称为“文本－图像”方式，本文采用这一方式，用 pdf 英文文档为密写载体，在文字图像中检测嵌入信息。

2．在 TeX 源文件中嵌入密写信息

我们通过微调文本中单词之间的空格嵌入隐蔽信息。为了保证可靠检测并使密写不被察觉，需要应用适当的编码技术。一种防止误码扩散的有效方法是利用句末空格作为参考点。在用 TeX 排版所生成的 pdf 文本文件中，句号（惊叹号、问号）后面的空格比同一行其他单词之间的空格大一些。用“\hspace”命令调节句子内的词间空格可嵌入二进制序列：嵌入 1 就将空格略增大一些，嵌入 0 则不变。检测时只要能在文本图像中准确定位句末就能正确提取原来嵌入的信息。对句末的定位可用相关法检测空格前的特定标点符号来实现。只要能可靠检测，词间空格增大的程度愈小愈好，以能够可靠检测为限。

另一种方法是在TeX源文件中的句号后面加“\”来减小空格，使每一行的空格都一样大，因此文件中所有的空格都可用于承载隐蔽信息。在数据嵌入时对全文进行统一编码。为了控制误码扩散，可使用自同步码例如T码[14]。

在pdf文本中换行处的空格是不显示的，无法从文本图像中检测出来。而在TeX源文件中又不能预测换行位置，所以必然会丢失一部分代表嵌入数据的空格。要正确提取嵌入信息，无论用何种编码方法均需采取措施将丢失的空格补上。为此，我们在整个文档最后保留一定的空间作为行末丢失空格的缓存区，参看图 1。为了表述简洁，不妨将整个pdf文件拼接成一页，不会影响处理结果。文件上部的区域I为基本数据区，从文件底部向上排列的II、III、IV等为缓存区。区域II用于嵌入区域I行末的丢失数据，区域III用于嵌入区域II行末的丢失数据，依此类推，直到一个只包含 1 行的区域，

而且该行没有行末的丢失数据为止。假定欲嵌入M比特数据，在pdf文件中每行平均空格数为N。保留 20%的冗余以防止因长单词过多、空格不够而造成数据溢出。区域I至少要有L_1=1.2M/N行。其中需缓存的行末丢失数据最多有L_1比特。各区域所包含的行数由下式给出：

$$L_k = \left\lceil M\left(\frac{1.2}{N}\right)^k \right\rceil, \qquad k = 1,2,\ldots,K \tag{1}$$

其中区域总数为：

$$K = \lceil \log M / (\log N - \log 1.2) \rceil \tag{2}$$

需要的总行数：

$$L = M\sum_{k=1}^{K}\left(\frac{1.2}{N}\right)^k \tag{3}$$

例如，欲嵌入M=2000 比特，每行平均有N=12 个空格，则共分为 4 个区域，L_1＝200，L_2＝20，L_3＝2，L_4＝1，总共需要 223 行。

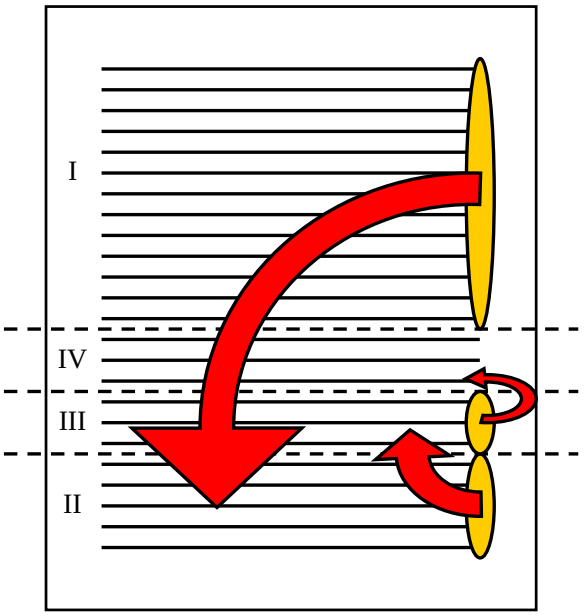

图 1 在 TeX 源文件中嵌入数据的过程

嵌入步骤如下：

1．用本节一开始所述方法将 M 比特数据嵌入 LaTeX 源文件。

2．用下节讨论的方法提取文件区域 I 中的嵌入数据。

3．将提取的数据与嵌入数据相比较，得到丢失的比特。注意段尾不含嵌入数据。

4．将丢失的比特嵌入区域 II。

5．提取数据并再与嵌入数据比较，将丢失比特嵌入下一个区域。

6．如此继续，直到丢失数据可嵌入一行中，不出现丢失比特为止。

3．从文本图像中提取密写信息

首先定位文字图像中的行。将图像的像素沿水平方向累加并归一化，得到投影函数 $P(y)$。图 2 是经过平滑的投影函数。

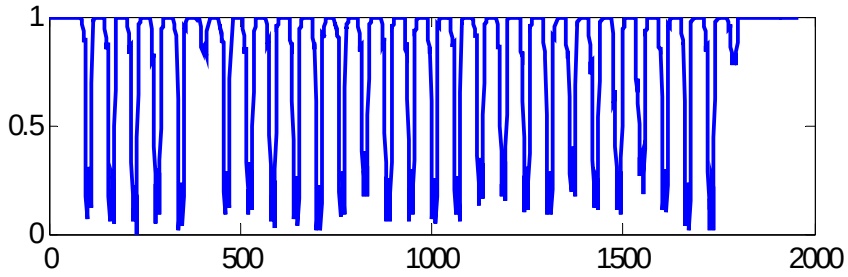

图 2 平滑的归一化水平投影：横坐标为沿文件页面垂直方向的像素数

根据投影函数中的峰值位置取出图像中的各行文字，如图 3 所示。其中左起第二个标有小方块的空格为句末，可见比其他空格略宽。图 4 为一行文字图像的垂直投影（仅画出了顶部），由于空格具有最大投影值，可从该图定位空格，如图 5 所示。

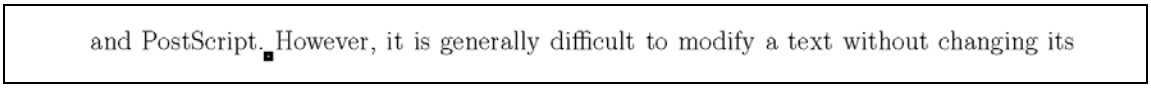

图 3 原始载体文本图像中的一行文字

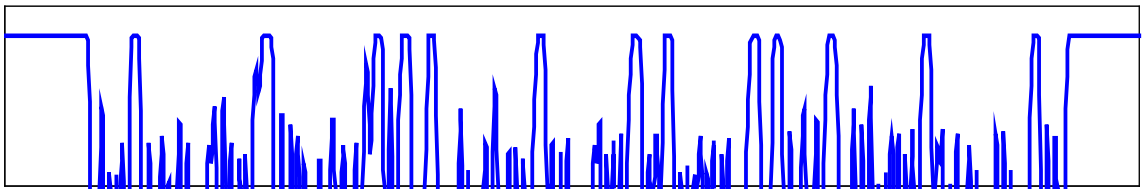

图 4 一行文字图像的垂直投影

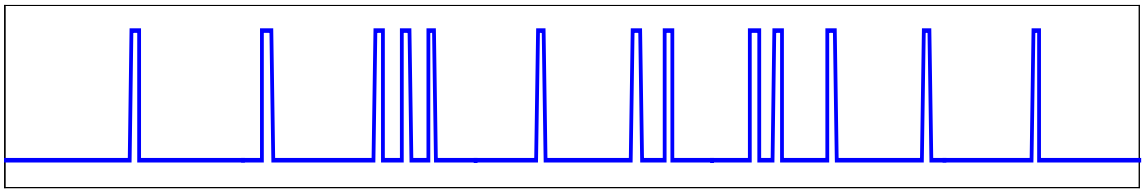

图 5 空格检测结果：句号后的空格略宽，其余基本均匀

图 6 是含有嵌入数据的同一行文字图像，其中在句末空格以后有 4 个较宽的空格，用小圆圈标出。将图 6 中的最大投影值置 1，其余置 0，可得到图 7。用一个移动积分器计算图中的脉冲宽度，以适当的阈值作判决，得到句末空格以后的嵌入数据为 10001010010，与嵌入数据相同。

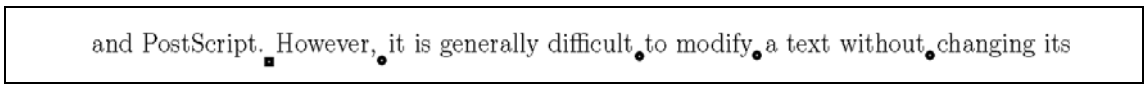

图 6 在句末空格之后有 4 个空格被加宽

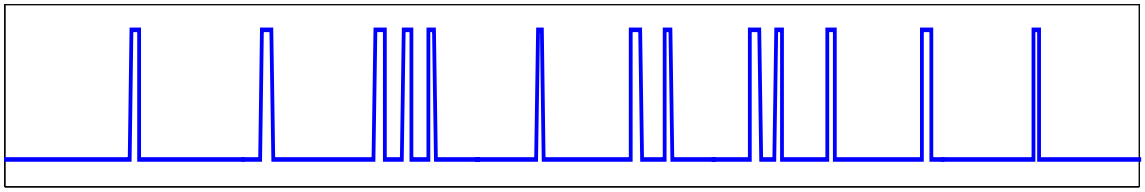

图 7 对含嵌入数据行的检测结果：准确地定位了 4 个“1”

4．讨论

利用常用的排版工具 TeX 以“文本－图像”方式进行的密写是一种非对称信息隐藏技术，嵌入过程在源文本中完成，经编译生成 pdf 文件，接收方将 pdf 文件转换为图像，通过图像处理提取嵌入数据。

在能够准确检测的前提下尽量减小对空格宽度的改动，并采用较大的载体文件将数据分散嵌入，结合使用加密和编码技术，所生成的含密文本具有良好的隐蔽性。而“文本－文本”方式却存在一些缺点，首先是传输 TeX 源文本的情况并不多，含密文件不易隐藏；其次是在控制命令中插入过多的空格显得十分反常，不够安全。对于“文本－文本”方式，用 HTML 文件作为载体是一种更好的选择，因为这类文件通常与网页浏览器关联，很少有人会用文本阅读工具打开网页文件。只要隐蔽数据不产生异常的浏览器显示就是安全的。

利用 TeX 的其他许多属性也可进行数据隐藏，例如在字符级嵌入可望实现更大的嵌入量和更好的隐蔽性。除文字部分外，也可以将数据嵌入公式和表格中。

参考文献：

[1] Wang H, Wang S. Cyber Warfare: Steganography vs. Steganalysis [J]. *Communications of the ACM*, **47**(10), 2004: 76-82

[2] Fridrich J, and Goljan M, Practical Steganalysis of Digital Images — State of the Art [C]. *Security and Watermarking of Multimedia Contents IV, Proceedings of SPIE*, **4675**, 2002: 1-13

[3] Wang S, Chen C, Zhang X. Undercover Communication Using Image and Text as Disguise and Countermeasures [J]. *Journal of Shanghai University*, **10**(1), 2006

[4] Bender W, et al., Techniques for Data Hiding [J]. *IBM Systems Journal*, **35**(3,4), 1996: 313-336

[5] Brassil J T, Low S, Maxemchuk N F, O'Gorman L. Electronic Marking and Identification Techniques to Discourage Document Copying [J]. *IEEE Journal on Selected Areas in Communications*, **13**(8), 1995: 1495-1503

[6] Brassil J T, Low S and Maxemchuk N F. Copyright Protection for the Electronic Distribution of Text Documents [J]. *Proceedings of the IEEE*, **87**(7), 1999: 1181-1196

[7] Maxemchuk N F, Low S H. Performance Comparison of Two Text Marking Methods [J]. *IEEE Journal Selected Areas of Communications*, **16**(4), 1998: 561-572

[8] Takizawa O, *et al*. Method of Hiding Information in Agglutinative Language Documents Using Adjustment to New Line Positions [C]. IEEE IIH-MSP-05, 14-16 Sept. 2005 Melbourne, Australia, *LNCS/LNAI*, **3683**, 2005: 1039-1048

[9] Villan R, *et al*. A Theoretical Framework for Data-Hiding in Digital and Printed Text Documents [C]. The 9th IFIP TC-6 TC-11 Conference on Communications and Multimedia Security, Salzburg, 19-21 Sept. 2005, *Lecture Notes in Computer Science*, 3677, 2005: 280-281

[10] Inoue S, *et al*. A Proposal on Information Hiding Methods Using XML [C]. *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, Nov. 2001: 55-62

[11] Lin S L. *New Methods of Data Hiding in TeX Documents* [D]. Masters Thesis, National Kaohsiung First University of Science and Technology, Taiwan, June 2004

[12] Knuth D E. *The TeXbook* [M]. Reading, MA, Addison Wesley, 1984

[13] Lamport L. *A Document Preparation System LaTeX, User's Guide and Reference Manual* [M]. Reading, MA, Addison Wesley, 1994

[14] Manoharan S. Towards Robust Steganography Using T-Codes [C]. *Proceedings of the 4th EURASIP Conference on Video/Image Processing and Multimedia Communications*, Zagrab, Croatia, 2-5 July 2003: 707-711