自然语言水印鲁棒性分析与评估

何 路[1),2),3)] 桂小林[1),2)] 田 丰[1),2)] 武睿峰[3)] 房鼎益[3)]

[1)](西安交通大学计算机科学与技术系 西安 710049)

[2)](陕西省网络重点实验室 西安 710049)

[3)](西北大学信息学院 西安 710127)

摘 要 自然语言与图像、音频信号的性质截然不同,图像水印等的鲁棒性分析方法不适用于自然语言水印,但是直到目前还没有专门针对自然语言水印鲁棒性的研究和评估工作.文中针对自然语言的特点,提出适合自然语言水印的敌手模型.然后将现有的自然语言水印分类,并总结各类的一般算法模型.利用文本提出的敌手模型分析自然语言水印编码算法的鲁棒性,并通过实验验证鲁棒性的理论模型.本项工作为对比、评估自然语言水印算法的鲁棒性提供了理论依据.

关键词 自然语言水印;鲁棒性;主动攻击;水印攻击;自然语言信息隐藏

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2012.01971

Analyzing and Evaluating the Robustness of Natural Language Watermarking

HE Lu[1),2),3)] GUI Xiao-Lin[1),2)] TIAN Feng[1),2)] WU Rei-Feng[3)] FANG Ding-Yi[3)]

[1)](*Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an* 710049)

[2)](*Key Network Laboratory of Shaanxi Province, Xi'an* 710049)

[3)](*School of Information Science and Technology, Northwest University, Xi'an* 710127)

Abstract The nature of natural language are quite different from the signal of images and sound, and the methods of robust analysis on image watermarks are not to be able to apply to NLW. However, the study of robustness of NLW is still absent. In this paper, based on the nature of natural language, we propose an adversary model that suit for NLW. Then, we classify the existing NLW methods and propose common algorithms. Third, we analyze the robustness of NLW algorithms using the adversary model that we proposed, and verify the theoretical model by experimental results. Our theoretical robust models are useful for comparison and evaluation robustness of different NLW algorithms.

Keywords natural language watermarking; robustness; active attack; attacking watermarking; natural language information hiding

1 引 言

自从 2001 年 Atallah 等人[1]提出自然语言水印的概念后,几乎每年的国际信息隐藏大会都会有关于自然语言信息隐藏的论文发表.其它各类期刊、会议发表的论文更是繁多.自然语言文本水印在载体文本里,利用自然语言处理(NLP)技术进行等价

收稿日期:2012-05-15;最终修改稿收到日期:2012-07-05.本课题得到国家科技计划重大专项(2012ZXD3002001)、国家自然科学基金(61172090)、陕西省科技攻关项目(2012K06-30)和陕西省教育厅科学研究计划(12JK0742)资助.**何 路**,男,1977 年生,博士研究生,讲师,主要研究方向为信息隐藏、文本水印、文本密写分析、网络与信息安全.E-mail: helu@nwu.edu.cn.**桂小林**(通信作者),男,1966 年生,博士,教授,博士生导师,主要研究领域为信息安全、服务计算.E-mail: xlgui@mail.xjtu.edu.cn.**田 丰**,男,1985 年生,博士研究生,主要研究方向为信息安全.**武睿峰**,女,1989 年生,硕士研究生,主要研究方向为文本水印、文本密写分析.**房鼎益**,男,1958 年生,博士,教授,博士生导师,主要研究领域为网络信息安全.

信息替换、语态转换等办法把水印信息嵌入文本中，并且不改变文本的原意[2]. 鲁棒性是水印的一个重要性能指标，它是指水印嵌入算法应该能够抵抗标准的或恶意的数据处理所引入的失真[3]. 由于水印信息是嵌入在文本的语法、语义结构之中，不像基于文本排版的文本水印[4]、基于字符图像的文本水印[5]和基于不可见字符的文本水印①，水印信息不会因为格式排版、文件格式转换、字符集转换或光学字符识别(OCR)等常见的编辑处理而被破坏，因此适用范围可以更加广泛. 但恶意的修改载体文本的内容只能够借助水印编码算法来对抗[6].

虽然目前已经有了很多自然语言水印的研究工作，但是关于恶意攻击的研究却还不多. 本文首先分析了自然语言水印面临的各种攻击，提出自然语言水印的恶意攻击的敌手模型，然后将已有的自然语言水印进行分类归纳以便进一步的分析. 由于自然语言与图形、图像、音频信号在本质上截然不同，我们结合本文提出的敌手模型，利用组合数学的方法来分析各类自然语言水印并建立鲁棒性模型，这些模型可以用来评估各类自然语言水印编码算法的鲁棒性.

本文第 2 节分析自然语言水印面临的攻击，并提出敌手模型；第 3 节是定义与符号；第 4 节描述评测工具的具体实现；第 5 节通过对现有算法的回顾，把自然语言水印进行分类和归纳；第 6 节针对各类算法分析鲁棒模型，并通过实验验证；最后是总结与展望.

2 敌手模型

目前对自然语言水印攻击方法的研究基本都遵循 Atallah 等人[1]提出的敌手模型：攻击者必须在不大量改变句子原意的条件下破坏文本中包含的水印信息，可采用以下措施来对含有水印信息的载体文本进行攻击：

(1) 对文本中的句子进行保留意义的转换.

(2) 如果将句子的意义进行了改变，那么有可能会破坏文本中的秘密信息，但修改的句子不能过多，否则影响文本的原意.

(3) 在文本中插入一个新的句子，也有可能会破坏文本中的秘密信息.

(4) 将文本中一块连续的部分从一个地方移动到另外一个地方，有可能会破坏文本中的秘密信息.

Gupta 等人[7]又补充了一条规则：

(5) 在保留文章意义的前提下可在文本中删除一些句子.

以上的敌手模型只是针对句子级别的水印嵌入方法，但其原则可以适用于各种改写方法. 因此可以重新整理为

(1) 根据 Kerckhoffs 原则[8]，攻击者可以使用与水印算法相同的 NLP 工具对载体文本中的同义词或句子等进行语义不变的变换，以期擦除水印[9].

(2) 改变语义的同义词替换、句子变换等，以期破坏水印同步. 但攻击的地方不能过多，否则影响文本的原意.

(3) 插入或删除少量或不重要的句子以期破坏水印同步.

(3.1) 插入新的句子. 插入的句子不可能是任意的，因为要保证语义的流畅，所以只能是在语义上是重复性的或透明性的句子[1].

(3.2) 删除不重要的句子. 人类语言具有一定的冗余，比如重要的信息往往会在文本中重复出现；需要强调的信息也会通过不同表达方式重复叙述. 从载体文本中摘除少量重复性的句子不会对载体文本的使用价值造成明显影响.

(4) 调整句子、段落的顺序. 相当于打乱水印比特的顺序. 此类攻击称为排序攻击.

其中第(2)条相当于删除原有语义后再插入新语义，效果相当于(3.1)和(3.2)同时使用，但造成语义改变较(3.1)和(3.2)大，而且 NLP 工具难以自动实现.

第(3)条中，考虑到要保证语义的流畅，(3.1)也难以实现：因为要找到合适的位置可以插入重复性或透明性的句子，根据目前的 NLP 技术还难以做到. 而(3.2)可以利用自动摘要软件实现. (3.1)和(3.2)的效果分别来看是相同的，两种攻击从效果上看都是相当于同步攻击，即使得嵌入时和提取时水印比特序列错位造成秘密信息无法正确提取[3]. 因此，只须考虑(3.2)即可.

第(4)条，一方面除非载体文本是说明书、手册之类的文体，其中包含大量的并行结构，不然调整句子、段落的顺序会显著影响语义逻辑. 另一方面，如第 4 节所述，目前自然语言水印的编码算法都不按载体单元在载体文本中出现的顺序编码，所以不受文本中物理位置变化的影响. 因此本文不考虑此类攻击.

① FFEncode. http://www.securityfocus.com/tools/1188

由此可见，评估自然语言水印算法时，实际上需要考虑的敌手模型只有第(1)条和(3.2)条.

3　定义与符号

文本涉及的术语和符号较多，为了方便描述，本节集中定义这些术语和符号.

定义 1.　载体单元. 文本中可被一种特定 NLP 技术处理并生成语义不变变换的最小语言片段.

定义 2.　可行变换. 在给定载体单元所处的上下文中，与载体单元语义相同的不同表达形式称为可行变换.

定义 3.　可行变换集合. 一个载体单元和它的所有可行变换构成的集合，称为可行变换集合.

例如，对基于同义词替换的水印算法，载体文本中具有同义词的单词是载体单元，它的同义词都是可行变换，该单词以及它的同义词构成可行变换集合. 而对于基于句式变换的水印算法，载体文本中可以做句式变换的句子就是载体单元，它的各种变换句式是可行变换，这些句子构成可行变换集合. 对于一个给定的载体文本以及该文本中的一个载体单元，可行变换集合中的每个元素，在该语境中均可以互相替换，而不影响语义.

设载体文本 D，D 中共含有 n 个载体单元，记作 $S=\{s_1,s_2,\cdots,s_n\}$. 每个载体单元的 s_i 可行变换集合记作 $T_i=\{t_{i_1},t_{i_2},\cdots t_{i_m}\}$，$l$ 代表水印长度.

定义 4.　嵌入率(e). $e=\dfrac{l}{n}$.

目前除文献[10]是采用向量作为水印外，其它自然语言水印算法均采用比特串作为水印，所以本文采用误比特率作为衡量水印算法鲁棒性的指标.

定义 5.　误比特率(BER). 载体文本经历攻击后，提取的水印比特中出错的比特个数与水印比特的总数之比.

在给定攻击方法和攻击力度的情况下，BER 越高，算法的鲁棒性越差.

定义 6.　替换攻击. 根据 Kerckhoffs 原则[8]，假设攻击者知道并且可以使用与水印算法相同的 NLP 工具对载体文本中的载体单元进行语义不变的变换，这种攻击方法称为替换攻击. 即敌手模型第(1)条.

自动摘要技术一般通过基于统计的算法，同时利用语言学技术来识别文档中的重要段落或语句，然后将这些文本片段提取出来粘帖在一起形成不具冗余的摘要，其长度虽然短于原始文档，但几乎没有信息损失[11].

定义 7.　压缩比(c). 通过摘要软件提取出来的词数占原始文档词数的百分比称为压缩比，即 $c\%$ 的摘要是从原始文档中提取 $c\%$ 的文字形成的①.

定义 8.　摘要攻击. 通过自动摘要软件，从载体文本中删除少量语义上不重要的句子，这种攻击方法称为摘要攻击. 即敌手模型第(3.2)条.

定义 9.　攻击力度(a). $a=k/n$，k 是 D 中被攻击的单元数量.

例如，替换攻击的攻击力度就是被替换的载体单元数量；摘要攻击的攻击力度就是摘要删除的载体单元数量. 假设载体单元在载体中是均匀分布的，摘要攻击力度反映了对 D 的摘除比例，即 $a=1-c$. 显然，攻击力度越高，文本摘除比例越大，因此语义损失也就越大，所以 a 应当远远小于 1.

4　鲁棒性测评工具的实现

由第 2 节的分析可见，自然语言水印的鲁棒性评估只需考虑两种攻击方法，即替换攻击和摘要攻击. 我们设计了相应的自动攻击工具②.

替换攻击使用的技术与各个自然语言水印算法完全相同，所以不再赘述. 需要注意的是同义词替换、句式变换等技术还不完美，为了保证不破坏载体的使用价值，攻击者只能实施少量的替换攻击，而不能通过替换所有可能的载体单元来擦除水印信息. 此外，Topkara 等人[11]针对词义消歧的困难性，选择语义扭曲度在可接受范围内的具有最大词义扭曲度的同义词进行替换. 如果攻击者再次进行同义词替换，消歧错误的数量会被放大，使替换攻击更加难以有效实施. 由此可见，替换攻击有相当大的局限性.

摘要攻击利用 Office Word 2007 中的自动摘要功能实现从文本中删除少量语义上不重要的句子. 根据摘要的实施策略不同，进一步把摘要攻击分为三种：对整篇文本直接实施摘要攻击称为 SAI. SAI 的缺点是删除的句子不是均匀分布在文本中的. 因为对于两个语义上相似的句子，摘要软件总是倾向

① SUMMARIST: Automated Text Summarization. http://www.isi.edu/natural-language/projects/SUMMARIST.html

② 我们已经开发出了自动攻击工具，支持替换攻击和摘要. 并提供绝对同义词替换方法，支持中英文两种语言. 集成空域一般模型和 Hash 域扩频两种水印编码，支持水印算法插件扩展. 可以作为测试使用，下载地址：http://dc-security.org/download/AutoAttacker.html

于保留出现在前面的句子.为此我们定义了 SAII:给定 D,设 D 共包含 n 个载体单元.设 p_i 是 D 的第 i 个段落,根据 Kerckhoffs 原则我们可以识别出 p_i 中的所有载体单元.首先,按照背包算法,设 p_i 是物品,物品的价值是 p_i 中包含的载体单元数量,记作 v_i,对各个 p_i 分别做摘要,被删除的载体单元数量记作重量 w_i.背包的容量为 $W=a\times n=(1-c)\times\sum_i w_i$.调节 c 可以改变删除的句子在载体文本中的分布,因为当 c 越接近于 1,一个段落中删除的句子数量越少,因此就会在越多的段落上进行摘要攻击.给定 a 和 c,SAII 会尽可能选择包含载体单元多的不重要句子进行删除,而且摘除句子的分布可控,因此理论上攻击效果应该最好.

SAII 依赖 Kerckhoffs 原则,如果攻击者不了解水印所采用的具体技术,则难以实施攻击,这时可以设 v_i 恒等于 1,我们称之为 SAIII.SAIII 是 SAII 的化简版本,但第 6 节的实验表明 SAIII 和 SAII 的性能相差不多,更适合作为通用的评估工具.

5 自然语言水印的编码技术

自然语言与图像、音频、视频等载体有本质上的区别:(1)自然语言缺乏变换域,不能将信号处理的技术运用在自然语言之上;(2)自然语言冗余空间较少,即人类对于文字的改变相当敏感,除了要考虑语法、语义之外,还要考虑语用习惯.目前虽然已经提出不少针对自然语言水印的编码算法,但自然语言水印的研究者大多来自于自然语言处理领域,对编码技术知之甚少,所以到目前为止自然语言水印的编码技术大多比较简单.以下我们回顾自然语言信息隐藏中几种主要的编码技术.

5.1 空域技术

T-Lex① 给 D 中每一组同义词集合中的单词从 0 开始编号.每一个拥有同义词的单词都对应一个进制不同的一位数字.设 D 中包含 N 个同义词,那么这 N 个单词联合在一起就可以视作一个 N 位的混合进制数.秘密信息可以看作一个二进制数,利用同义词替换,使 D 表示的混合进制数等于秘密信息表示的二进制数就完成了嵌入过程.提取过程只是简单地从文本中把这个混合进制数读取出来再转换回二进制数.

但该算法没有使用密钥,之后很多基于词法的自然语言信息隐藏算法在其基础之上进行了改进.这些方法的一般过程为:在密钥的控制下对嵌入信息的词进行秘密排序,依次找到需要嵌入的比特对应位置的单词,根据编码选择与需要嵌入的比特相同的词进行替换[7,12-13].

Atallah 等人[1]分析载体文本中句子得到句法树结构,对文本中句子的句法树中每个节点按先序遍历的顺序编号,然后对每个节点的编号 j 计算 $j+H(p)$,如果是 p 的二次剩余,那么节点的标为 1;否则为 0,其中 p 是大质数,$H(\cdot)$ 是 Hash 函数.然后后序遍历节点将节点的比特连起来得到一个二进制数 B_i,使用 $d_i=H(B_i)$ XOR $H(p)$ 对句子进行排序.选择 d_i 最小句子作为"标志句",标志句在文本中的后继句为"水印句".通过句法变换在"水印句"中嵌入信息.一个水印句可以承载多个比特,但如果从鲁棒性的角度出发只嵌 1 比特更好.Atallah 等人也指出算法中水印句要由它的标识句来指示,如果进行一次排序攻击,有可能正好把一对标识句和水印句分开.由此造成水印被破坏的概率 $\leq 3\alpha/b$,其中 b 代表文本中句子的总数,α 代表个水印句个数,并且每个水印句嵌入 1 比特水印.但是 Atallah 等人没有给出一般的误比特率模型.文献[14]中不再使用标志句,水印比特逐句嵌入.如果文本足够长则重复嵌入水印,提取时通过多数投票机制对抗篡改攻击,但仍然不能抵抗排序攻击.文献[15-16]通过引入重排序机制,抵抗对于句子顺序改变的攻击.

为了提高鲁棒性,将 S 在密钥控制下重新排序而不是在载体文本中出现的位置排序,可以抵抗排序攻击,目前几乎所有的算法都采用了这种策略.还有些研究者采用随机间隔嵌入[1,17]、重复嵌入[7,16],或者纠错机制[18].这相当于减少了一个载体单元承载秘密比特的数量.承载的比特数越少,遭受攻击后的 BER 也应该越小.值得注意的是重复嵌入可以看作是空域上的直接序列扩频.

空域算法繁多,虽然对 S 排序的方法各不相同,但都是利用密钥和 s_i 进行运算,得出一个秘密数字,根据这个秘密数字的大小进行排序(由此也可以看出虽然 S 排序的结果不再与文本中出现的顺序有关,但不过是在另一个域上的顺序排序,所以本质上仍是空域性质的).为了分析方便,我们根据空域算法的一般过程总结出空域算法的一般模型:

首先找出载体文本里的所有载体单元 $S=$

① Winstein K. Tyrannosaurus-Lex[EB/OL]. http://alumni.imsa.edu/~keithw/tlex/

$\{s_1, s_2, \cdots, s_n\}$，每个 s_i 的可行变换集合 $T_i = \{t_{i_1}, t_{i_2}, \cdots, t_{i_m}\}$，根据密钥从中选择一个可行变换 t_{i_j} 作为代表元(选择代表元是为了防止嵌入时选择不同的可行变换可能会造成嵌入前排序的结果与提取时排序的结果不一致). 然后根据密钥和 t_{i_j} 计算出一个秘密数字，按照这个秘密数字的大小把 S 的元素排序. 接下来根据密钥从 S 中随机选出部分元素准备嵌入，记作 $S' = \{s'_1, s'_2, \cdots, s'_l\}$. 把 S' 中的元素 s'_i 的可行变换集合 T'_i 用密钥分成两个子集，分别代表 0 和 1. 嵌入时对比水印比特与 s'_i 表示的比特是否相同，如果不同则使用相应的可行变换对其进行替换以生成含密文章.

该一般模型不考虑应用纠错码的情况，也不考虑扩频的情况. 5.3 节将给出扩频的一般模型.

5.2　变换域技术

自然语言不像图像，缺乏变换域. 现有的变换域算法一般都是以载体单元的某些特征出现的频次作为变换域来嵌入信息的. 由于水印不是嵌入在空域，所以排序攻击均告无效.

戴祖旭等人[19]随机选择载体文本中部分标记串，连同其频数构造一个完备概率空间，通过修改文本改变标记串的概率分布使其信息熵与水印一致. 显然，该算法能抵抗排序攻击. 该算法相当于在变换域上随机选取子带进行嵌入.

Yang 等人[10]把水印表示成 l 维整数向量. 首先把文本中的同义词分成 l 组，并且每组中同义词对应的同义集合分为 A、B 两个子集. 然后用同义词替换使第 i 组中属于 A 子集的同义词个数等于水印向量第 i 维分量. 检测时采用线性相关，当大于给定的阈值时就认为水印存在. 该方法本质上相当于图像水印的归一化相关检测方法.

5.3　扩频技术

Vybornova 和 Macq[20]把文本中的句子根据密钥秘密排序并按水印长度分组，根据每组句子中包含的前提数量的奇偶性表达水印信息，如包含奇数个前提表示秘密比特 1；包含偶数个前提表示秘密比特 0. 显然，这种编码方案也可以抵抗排序攻击. 同时，这种编码的好处在于变换数量不多于水印长度，变换的数量越少隐蔽性自然越好. 但是由于句子的排序不过是在另一个域上的顺序，其本质上属于空域扩频技术. 与其类似的还有文献[16].

何路等人[21]提出一个更通用的算法. 根据密钥和载体单元的代表元计算 Hash 值，把载体单元是否包含特定的特征，按照文献[20]的方法进行编码. 但并不排序，分组是按 Hash 值的定义区间(即 $0 \sim 2^{128}$)均匀划分为 l 个子区间. 这样载体单元之间不再有序列关系，而成为变换域(如 Hash 域)上的扩频.

类似的还有姜传贤等人[22]基于文本的重要内容提出的一种水印算法. 该算法首先挑选出文本中所有包含主题词的句子，组成主题句集合 CS；然后，根据同义词词典对 CS 的句子进行筛选：选取包含同义词的子主题句集合 $subCS$；接着，根据给定的密钥对每个句子中主题词的最高频率求 Hash，以求得的 Hash 值做模 l 运算(l 是水印长度)，具有相同运算值的句子将被分入一组；最后，依次将水印的每一位嵌入到一组句子中.

以上两种扩频技术都属于直接序列扩频. 空域和变换域的扩频技术区别在于对载体单元排序和分组是否与顺序有关. 下面给出扩频编码算法的一般模型：

扩频编码模型：首先根据密钥和载体单元的代表元计算秘密值. 然后根据秘密值把 S 划分成 l 个组. 每组包含的载体单元数量记作 x. 最后利用组内包含规定特征的载体单元的数量进行编码. 当组代表的比特与目标水印比特不同时，对组内的载体单元进行替换使规定的特征出现的次数正好表达水印的比特.

6　自然语言信息隐藏的鲁棒性

引理 1.　在同义攻击中，s_i 被成功攻击的概率至少为 0.5.

证明.　按照编码算法的不同，T_i 可以被划分成两个子集 T'_i 和 T''_i. 设 T'_i 表示“0”；T''_i 表示“1”. 若集合 T_i 中的元素个数为 m，则集合 T'_i 和 T''_i 中的元素个数有表 1 所示的几种情况.

表 1　T'_i 和 T''_i 中元素的划分情况列举

T'_i 中的元素个数	T''_i 中的元素个数	概　率
1	$n-1$	$(n-1)/(n-1)$
2	$n-2$	$(n-2)/(n-1)$
3	$n-3$	$(n-3)/(n-1)$
…	…	…
$n-1$	1	$1/(n-1)$

设任意一个载体单元 s_i 被编码为“0”，则在上述情况中 s_i 被成功攻击的概率如表 1 中第 3 列.

因为集合 T_i 中的载体单元编码的方式是随机的，所以每一个载体单元被编码为“1”或者“0”的概

率是相同的，即上表中每一行情况出现的概率是相同的. 因此 s_i 被成功攻击的平均概率为

$$I=\frac{\frac{n-1}{n-1}+\frac{n-2}{n-1}+\frac{n-3}{n-1}+\cdots+\frac{1}{n-1}}{n-1}$$

$$=\frac{\frac{n\times(n-1)}{2}}{(n-1)^2}=\frac{n}{2(n-1)},$$

其中 $n\geq 2$，n 为整数.

令函数 $f(x)=\frac{x}{2(x-1)}$，$x\geq 2$. 对 $f(x)$ 求导，得

$$f(x)'=\frac{2(x-1)-2x}{4(x-1)^2}=\frac{-2}{4(x-1)^2}<0,\ x\geq 2.$$

因此，$f(x)$ 在定义域内是单调递减函数. 对 $f(x)$ 求极限：

$$\lim_{x\to\infty}\frac{x}{2(x-1)}=\frac{1}{2}.$$

$f(x)$ 在定义域内有下界 1/2.

因此可知，在同义攻击中，每一个载体单元被成功攻击的概率至少为 0.5. 证毕.

6.1 空域算法的鲁棒性

6.1.1 替换攻击的误比特率分析

因为对一个载体单元 s_i 进行替换攻击，根据密钥挑选的代表元并不会改变，所以 S' 中的元素序列也不会改变. 显然，只有 s_i 被成功攻击时才会产生 1 比特误码. 嵌入秘密信息时的嵌入率为 e，攻击力度为 a，根据引理 1，空域算法在替换攻击下的理论 BER 为

$$BER=ae/2 \tag{1}$$

由此可见空域算法对抵御替换攻击的效果较好，通过降低嵌入率可以进一步降低替换攻击造成的误比特率.

6.1.2 摘要攻击的误比特率分析

分两种情况讨论空域算法在摘要攻击下的理论误比特率：

(1) 当 $e=1$ 时

首先考虑攻击只造成一个载体单元被删除的情况：设删除第 i 个载体单元(为了计算方便 i 是在 S' 中从后往前的计数，即 s'_l 为 1，s'_1 为 l，下同)，在提取过程中，第 $i-1$ 个载体单元会被错误地识别成第 i 个载体单元，而第 $i-2$ 个载体单元会被错误地识别成第 $i-1$ 个载体单元，以此类推，排在第 i 个载体后面的 $i-1$ 个载体单元将全部错位，因此 $BER=\frac{i}{l}$.

现在考虑攻击 k 个载体单元被删除的情况：由第一种情况可知，删除某一个载体单元，则其后所有的载体单元将错位，因此当多个载体单元删除时，造成的误比特率仅由秘密排序中最靠前的载体单元的位置决定(也就是最大的 i). 设攻击力度为 a，则需要攻击 an 个载体单元. 设最靠前的载体单元的位置为 i，此时产生的误比特率为 $\frac{i}{l}$. 而剩下的 $an-1$ 个载体单元的攻击位置都在 i 之后，总共有 $p=\binom{i-1}{an-1}\Big/\binom{n}{an}$ 可能. 所有可能的误比特率构成了一个离散型随机变量 X，X 的所有的可能取值为 $X_i(an\leq i\leq n)$. 式(2)给出在攻击力度为 a 的摘要攻击下 BER 的期望.

$$BER=\sum_{i=an}^{n}\frac{i}{L}\times\frac{\binom{i-1}{an-1}}{\binom{n}{an}} \tag{2}$$

(2) 当 $e<1$ 时

可以将 n 个载体单元分成 l 组，每一组有 $x=\frac{n}{l}$ 个载体单元，但只有 1 个载体单元承载水印比特. 设被摘要攻击删除的最靠前的载体单元的位置是 i(从后向前计数)，那么平均来看，这个载体单元处于从后往前计数的第 $\left\lceil\left(\frac{i}{\frac{1}{e}}\right)\right\rceil$ 组. 当摘要攻击删除的载体单元处于本组中承载水印比特的载体单元之前时，那么本组中的承载水印比特的载体单元就会发出错位，反之则不会. 如图 1 所示，黑色圆圈代表被摘要攻击删除的载体单元，虚线圆圈代表承载水印比特的载体单元.

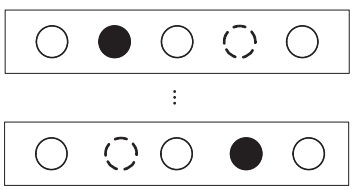

图 1 攻击位置与误比特关系

组内承载水印比特的载体单元处于 i 之后的概率是 $\left(\frac{i \bmod \frac{1}{e}}{\frac{1}{e}}\right)$，此时被攻击到的比特个数的期望是 $\lceil i\times e\rceil$，反之，组内的承载水印的词处于 i 之前的概率是 $\left(1-\frac{i \bmod \frac{1}{e}}{\frac{1}{e}}\right)$，此时被攻击到的比特个数

的期望是($\lceil i\times e\rceil-1$). 又每个 i 的发生概率是 $\binom{i-1}{an-1}\Big/\binom{n}{an}$. 所有可能的误比特率构成了一个离散型随机变量 X, X 的所有的可能取值为 X_i($an\leq i\leq n$),式(3)给出在攻击力度为 a 的摘要攻击下 BER 的期望.

$$BER=\sum_{i=an}^{n}\left[\left(\frac{i \bmod \frac{1}{e}}{\frac{1}{e}}\right)\frac{\lceil i\times e\rceil}{L}+\left(1-\frac{i \bmod \frac{1}{e}}{\frac{1}{e}}\right)\frac{\lceil i\times e\rceil-1}{L}\right]\times\frac{\binom{i-1}{an-1}}{\binom{n}{an}} \tag{3}$$

$$BER=\sum_{i=an}^{n}\frac{\lceil i\times e\rceil}{L}\times\frac{\binom{i-1}{an-1}}{\binom{n}{an}},\ i=\frac{k}{e},k\text{ 为正整数} \tag{3'}$$

如果 i 恰好为 $\frac{l}{e}$ 的倍数,那么攻击到的载体单元就是第 $\left|\left(\frac{i}{\frac{1}{e}}\right)\right|$ 组最靠前的载体单元,因此本组内承载比特的载体单元一定会发生错位. 此时 BER 的期望退化为式(3′)的形式.

基于以上分析,可以获得空域算法误比特率的理论曲线,如图 2 所示. 横坐标是攻击力度,纵坐标是 BER 的期望. 由图 2 可以看出降低嵌入率对抵抗摘要攻击几乎没有帮助. 另外,空域算法抵御摘要攻击的能力非常差,删除一个载体单元平均就会造成 50%的误比特率. 而且由于摘要攻击的效果相当于删除比特后造成的同步丢失,纠错码并不能找回丢失的比特,所以直接对水印比特应用纠错编码反而雪上加霜.

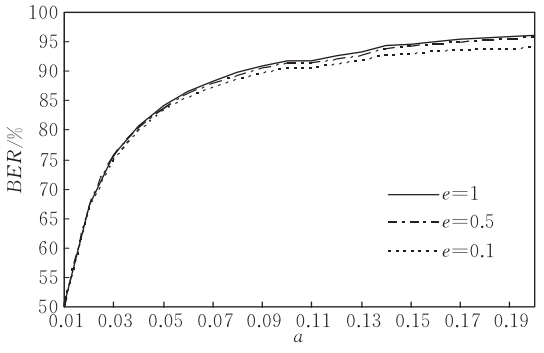

图 2　空域一般模型的理论 BER 期望

但是实际上,当第 i 个载体单元被删除,第 $i-1$ 个会被错误识别为第 i 个载体单元,只有当第 i 个载体单元和第 $i-1$ 个载体单元表示的比特不同时,才会造成误码. 如果 $e=1$,提取的水印比特将比嵌入的数量少,造成读取水印完全失败;而 $e<1$ 时,会把一部分未承载水印比特的载体单元当作嵌入水印比特的载体单元来读取. 由于秘密比特是随机的,不论是错位的比特还是误读取的比特,有一半的概率正好与嵌入的原始比特正好一致. 所以实际的 BER 曲线会比理论的 BER 曲线降低一半.

给定 $e=0.5$, $l=50$,首先使用空域的一般模型在测试语料中嵌入水印;然后令 $a=\{0.02,0.04,\cdots,0.20\}$,使用摘要攻击进行攻击,结果如图 3. 横坐标是攻击力度,纵坐标是 BER 的期望. 可以看出正如第 4 节所预期的:SAII 的攻击效果最接近于理论曲线,其次是 SAIII,SAI 的攻击效果最差. 而且 SAIII 与 SAII 的性能比较接近.

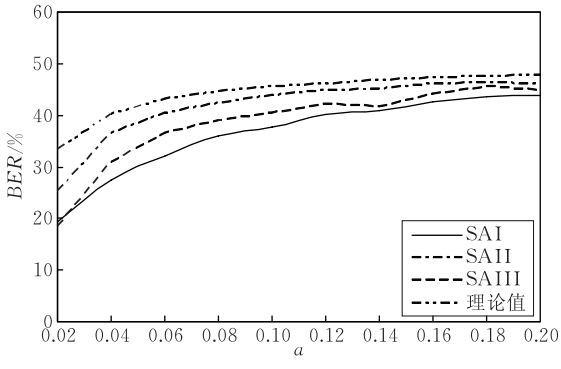

图 3　空域的一般模型的摘要攻击 BER 曲线

6.2　变换域算法的鲁棒性

由于变换域算法差异较大,难以用一个统一的模型分析,所以本节对每个算法逐个分析.

文献[19]的算法相当于在变换域上随机选取子带进行嵌入. 只有攻击到嵌入水印的子带上才有可能造成误比特. 文献[19]以长篇小说《保卫延安》为例说明了水印的嵌入提取过程.

考虑其中特别长和特别短的句子都难以进行变换;另外,根据嵌入算法原理,出现次数太少的无法嵌入信息,这就缩小了攻击的范围,据我们统计,《保卫延安》中实际可供嵌入的词性序列串模式(即句型)只有 145 种. 只有每个词性标记串都被攻击到,才有可能损坏水印比特. 替换攻击至少需要修改 145/2≈61 个句子,因为理想情况下替换攻击每次攻击一个载体单元,相当于攻击了两个不同的词性序列串模式(一个词性序列串模式的频数减 1;另一个词性序列串模式的频数加 1). 而摘要攻击至少需要删除 145 个句子,才有可能损坏水印比特. 因此替换攻击更有效率.

我们采用替换攻击，替换攻击程序采用马广平等人[9]的实现句子转换工具进行替换攻击，采用文献[19]的水印“WIT”，$l=24$. 攻击策略如下：首先，在《保卫延安》中搜寻所有可能嵌入水印的词性序列串模式. 然后，逐次从各个词性序列串模式中选择一个可以做句式变换的句子进行攻击. 记录发生变化的词性序列串模式，即已被攻击. 如果一个词性序列串模式中没有句子可以做变换，则跳过该词性序列串. 如果所有词性序列串模式均被攻击过，或剩余词性序列串模式无法攻击，则攻击完毕.

实验结果如图 4. 横坐标是被攻击的句子数量，纵坐标是 *BER* 的期望. 同样由于有一半的概率使错误的比特正好与嵌入的水印比特正好一致，所以 *BER* 趋向于 0.5 左右. *BER* 曲线呈现阶梯状，这是由于水印只嵌入在个别词性标记串模式上，替换攻击遍历各种词性标记串. 当正好攻击到嵌有水印的词性标记串时，*BER* 便会突然增加，否则 *BER* 维持不变. 全文包含 15 余万字，37031 个自然句子，搜寻并攻击几十个句子对载体的破坏十分微小. 由全文可做变换的句子有 741 个，可知攻击力度未超过 0.20. 说明文献[19]的算法鲁棒性依然不是很好.

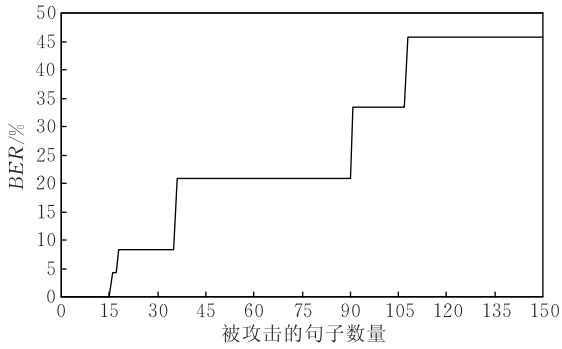

图 4 文献[19]的摘要攻击 *BER* 曲线

文献[10]的算法，在变换域上由于不需要对 S 进行排序，所以删除载体单元不会造成水印比特错位，但会造成相关性降低. 替换攻击有可能在某处减弱了水印向量，但在另一处又放大了水印向量，攻击效果有限[10]. 因此摘要攻击更有效率. Miller 和 Bloom[23]已经给出了归一化相关检测器的错误概率模型，在此不再复述. 需要注意的是：由于自然语言冗余空间非常有限，文献[10]的算法没有使用水印模板，而是直接把水印向量调制在载体上.

我们在测试语料中嵌入水印向量：{2，3，−1，0，4，−2，1，2，1，−3}；然后令 $a=\{0.02,0.04,\cdots,0.20\}$，使用摘要攻击进行攻击，攻击的结果如图 5. 横坐标是攻击力度，纵坐标是水印向量的相关度. 对比文献[10]的图 2 可见文献[10]的算法虽然能有效抵御替换攻击，但抵御摘要攻击的能力有限.

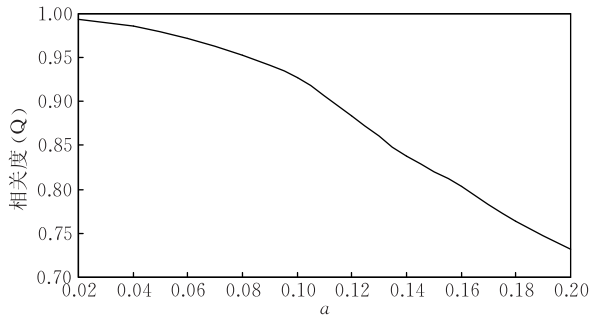

图 5 文献[10]的摘要攻击相关度曲线

6.3 扩频算法的鲁棒性

扩频算法可以根据排序方法分为两类：空域的扩频算法和变换域的扩频算法.

6.3.1 替换攻击的误比特率分析

因为对载体单元 s_i 进行替换攻击，根据密钥挑选的代表元并不会改变，所以载体单元序列不会改变. 因此空域和变换域的扩频算法抵抗替换攻击的性能是等价的. 以下讨论不区分空域或变换域.

若要使 *BER* 为 t/l，则被成功攻击的组就要有 t 个. 设 r 为一个组被攻击成功需要攻击的载体单元数量占组包含的载体单元数量的比例，攻击力度为 a，则被攻击的载体单元个数 $k=a\times n$. 设恰有 t 个组被成功攻击，从 l 个组中选择 t 个组的组合数为 $\binom{l}{t}$. 将 l 个组按照是否被成功攻击划分为集合 W_1 和 W_2.

$$W_1=\{\alpha_1,\alpha_2,\cdots,\alpha_t\},$$

$$W_2=\{\beta_1,\beta_2,\cdots,\beta_{l-t}\},$$

其中，α_i 是被成功攻击的组；β_i 是未被成功攻击的组. 用 x_i 表示 α_i 中被攻击的载体单元数，用 y_i 表示 β_i 中被攻击的载体单元数.

根据引理 1，要成功攻击某一组，则该组内至少需要攻击 $2rx$ 个载体单元. 则有

$$x_1+x_2+\cdots x_t+y_1+y_2+\cdots+y_{l-t}=n \quad (4)$$

其中，$2rx\le x_i\le x$，$0\le y_i\le 2rx-1$.

设 $z_i=x_i-2rx$，则 $x_i=z_i+2rx$，代入式(4)，得

$$z_1+z_2+\cdots+z_t+2trx+y_1+y_2+\cdots+y_{l-t}=n,$$

即

$$z_1+z_2+\cdots+z_t+y_1+y_2+\cdots+y_{l-t}=n-2trx,$$

其中，$0\le z_1\le x-2rx$，$0\le y_i\le 2rx-1$.

设 P_1 代表该组内攻击的载体单元个数至少为 $x-2rx+1$ 个的性质；P_2 代表该组内攻击的载体单元个数至少为 $2rx$ 个的性质. $A_1,A_2,\cdots,A_t$ 是满足性质 P_1 的集合；$A_{t+1},A_{t+2},\cdots,A_l$ 是满足性质 P_2 的集合. 根据容斥原理，式(5)给出了恰有 t 组被攻击成功的组合数.

$$\binom{l}{t}\times|\overline{A_1}\cap\overline{A_2}\cap\cdots\cap\overline{A_l}|=$$
$$\binom{l}{t}\times(|S_1|-\sum|A_i|+\sum|A_i\cap A_j|-$$
$$\sum|A_i\cap A_j\cap A_k|+\cdots+$$
$$(-1)^l\sum|A_1\cap A_2\cap A_3\cap\cdots\cap A_l|) \quad (5)$$

$$|\overline{B_1}\cap\overline{B_2}\cap\cdots\cap\overline{B_l}|=$$
$$|S_2|-\sum|B_i|+\sum|B_i\cap B_j|-$$
$$\sum|B_i\cap B_j\cap B_k|+\cdots+$$
$$(-1)^l\sum|B_1\cap B_2\cap B_3\cap\cdots\cap B_l| \quad (6)$$

其中 $|S_1|=\binom{n-2trx+l-1}{n-2trx}$.

若不考虑被成功攻击的组的个数，设每一组内攻击的载体单元个数为 k_i，则有

$$k_1+k_2+\cdots+k_l=n,$$

其中，$0\leq k_i\leq x$.

设 P_3 表示该组内攻击的载体单元个数至少为 $x+1$ 个的性质. $B_1,B_2,\cdots,B_l$ 是满足性质 P_3 的集合. 根据容斥原理，式(6)给出其组合数. 其中，$|S_2|=\binom{n+l-1}{n}$.

因此可以得到恰好有 t 组攻击成功的概率 P_t：

$$P_t=\frac{\binom{l}{t}\times|\overline{A_1}\cap\overline{A_2}\cap\cdots\cap\overline{A_l}|}{|\overline{B_1}\cap\overline{B_2}\cap\cdots\cap\overline{B_l}|}$$
$$=\frac{\binom{l}{t}\times(|S_1|-\sum|A_i|+\sum|A_i\cap A_j|-\sum|A_i\cap A_j\cap A_k|+\cdots+(-1)^l\sum|A_1\cap A_2\cap A_3\cap\cdots\cap A_l|)}{|S_2|-\sum|B_i|+\sum|B_i\cap B_j|-\sum|B_i\cap B_j\cap B_k|+\cdots+(-1)^l\sum|B_1\cap B_2\cap B_3\cap\cdots\cap B_l|}.$$

在对分成 l 组的 n 个载体单元进行攻击力度为 a 的攻击时，所有可能产生的误比特率构成一个离散型随机变量 X，设 X 的所有可能取值为 x_t（$1\leq t\leq l$ 且 t 为整数），X 的分布律为 $P\{X=x_t\}=p_t$（$1\leq t\leq l$ 且 t 为整数）. 式(7)给出随机变量 X 的数学期望 $E(X)$.

$$E(X)=\sum_{t=1}^{t=\min\{n,l\}}\frac{t}{l}\times P_t=\sum_{t=1}^{t=\min\{n,l\}}\frac{t}{l}\times$$
$$\frac{\binom{l}{t}\times(|S_1|-\sum|A_i|+\sum|A_i\cap A_j|-\sum|A_i\cap A_j\cap A_k|+\cdots+(-1)^l\sum|A_1\cap A_2\cap A_3\cap\cdots\cap A_l|)}{|S_2|-\sum|B_i|+\sum|B_i\cap B_j|-\sum|B_i\cap B_j\cap B_k|+\cdots+(-1)^l\sum|B_1\cap B_2\cap B_3\cap\cdots\cap B_l|} \quad (7)$$

令 $r=0.5$，$e=0.5$，$l=50$，在测试语料中使用扩频的一般模型嵌入 50 比特水印；然后令 $a=\{0.02,0.04,\cdots,0.20\}$，使用替换攻击，画出 *BER* 曲线如图 6. 横坐标是攻击力度，纵坐标是 *BER* 的期望，实线是 *BER* 的理论曲线，虚线是 *BER* 的实验曲线. 可见扩频算法在替换攻击下的 *BER* 曲线是呈对数趋势增长，如果扩频范围增大，*BER* 曲线还会进一步下降；而空域算法在替换攻击下的 *BER* 曲线根据式(1)是呈线性增长. 所以扩频算法抵抗替换攻击的效果比非扩频的空域算法略好一些.

6.3.2　摘要攻击的误比特率分析

由于在变换域的扩频算法分组不是按 S 的序列关系，所以删除载体单元不会影响其它分组. 因此，变换域的扩频算法抵御摘要攻击的性能与抵抗替换攻击的性能是等价的. 令 $e=0.5$，$l=50$，在测试语料中使用文献[21]的算法嵌入水印；然后令 $a=\{0.02,0.04,\cdots,0.20\}$，使用摘要攻击，结果如图 7，其中 ideal 代表理论曲线. 由图 7 可以看出变换域算

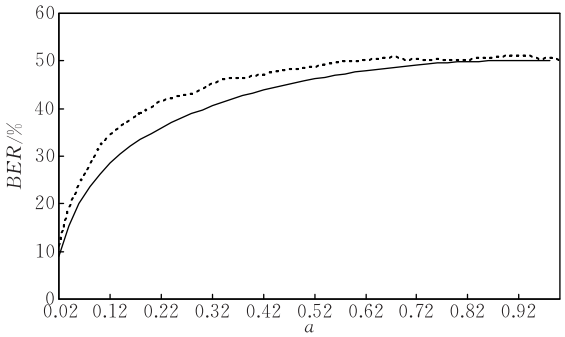

图 6　扩频一般模型的替换攻击 *BER* 曲线

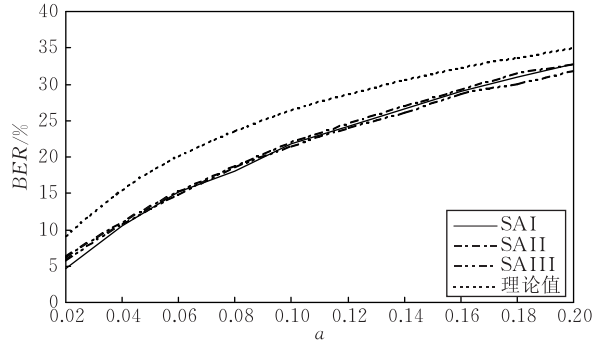

图 7　变换域扩频一般模型的摘要攻击 *BER* 曲线

法在抵御摘要攻击方面远远优于空域算法.另外SAI、SAII和SAIII的性能相差无几,且它们都与理论误比特率的曲线比较接近,而SAIII的前提更弱.结合图3可以看出SAIII作为测评工具更加通用.

下面分析空域扩频算法在摘要攻击下的鲁棒性.

显然至少要删除 $2rx$ 个载体单元才有可能造成1比特误码.所以当多个载体单元删除时,造成的误比特率仅由秘密排序中第 $2rx$ 个被攻击的载体单元所处的分组决定,设这个分组为 i(i 是从最后一个分组开始计数的).与5.1.2节的讨论类似,其后的分组错位的比特超过 $2rx$ 也都会出现误比特.设攻击力度为 a,则总共攻击 an 个载体单元.此时产生的误比特率为 $\frac{i}{l}$.如第 $2rx$ 个被攻击的载体单元处在 $i=l$ 组,则它在组的第 $2rx$ 到 x 之间;如果第 $2rx$ 个被攻击的载体单元处在其它组,则它在组的第1到 x 之间.所以它的平均位置近似为 $\left\lfloor\frac{1+x}{2}\right\rfloor$.在第 $2rx$ 个被攻击的载体单元之前有 $2rx-1$ 个载体单元被攻击,总共有

$$p_p=\binom{(l-i)x+2rx-1}{2rx-1}$$

种可能.而剩下的 $an-2rx$ 个载体单元都处在第 $2rx$ 个被攻击的载体单元之后,总共有

$$p_n=\binom{ix-1+x-\left\lfloor\frac{1+x}{2}\right\rfloor}{an-2rx}$$

种可能.由于空域扩频算法是在载体单元排序后按水印长度分组的,随着攻击力度的增加,必然造成每个分组包含的载体单元变少,所以 p_p 和 p_n 中的 x 不是固定的,但由于摘要攻击一般攻击力度不大,所以可以忽略.因此,所有可能的误比特率构成了一个离散型随机变量 X,X 的所有的可能取值为 X_i($an\leq i\leq n$).式(8)给出在攻击力度为 a 的摘要攻击下的 BER.

$$BER_{\mathrm{SAII}}=\sum_{i=\left\lfloor\frac{an-2rx}{x}\right\rfloor}^{l}\frac{i}{L}\times p_p p_n \tag{8}$$

令 $r=0.5$,$e=0.5$,$l=50$,在测试语料中使用扩频一般模型的空域方法嵌入50比特水印;然后令 $a=\{0.02,0.04,\cdots,0.20\}$,使用摘要攻击,画出 BER 曲线如图8.横坐标是攻击力度,纵坐标是 BER 的期望.

对比图3可见,空域的扩频算法只有当摘要攻击删除的载体单元较少时,产生误比特率才比较低;而当被删除的载体单元不少于 $2rx$ 后,BER 迅速上升,同样由于错位的原因 BER 与空域非扩频算法的情况几乎一样.所以空域的扩频算法不能有效抵御摘要攻击.

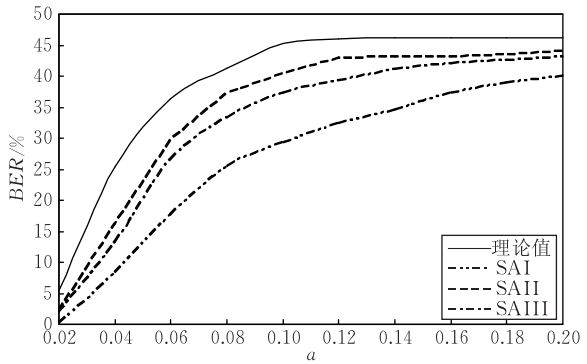

图 8

7 总结与展望

本文通过分析自然语言水印面临的恶意攻击发现,不像图像水印面临众多的攻击手段,自然语言水印实际上只需要考虑替换攻击和删除攻击两种攻击就可以代表自然语言水印所面临的所有主要威胁.这使得对自然语言水印的鲁棒性进行全面分析和评估成为可能.

本文为自然语言水印建立了鲁棒性模型,并通过实验验证了鲁棒性理论模型的正确性.该模型可以用于对比、评估自然语言水印的鲁棒性.分析和实验也发现自然语言水印编码算法中的一些问题:

(1)空域算法在摘要攻击下的鲁棒性非常差,并且随机间隔嵌入、应用纠错码和扩频编码都不会真正提高鲁棒性.

(2)即便在变换域上,如果仍然按某种序列关系对载体单元进行排序,其实质仍然是空域的,虽然可以抵抗排序攻击,但对于摘要攻击依然没有帮助.

(3)变换域上的扩频编码算法对各种攻击显示出较好的性能,应当成为研究鲁棒的编码算法的重点方向.但是寻找变换域需要慎重,比如,以词性标记串代表句型的这种变换域由于句型数量有限,不一定能够保证鲁棒性.

为了使我们的工作更实用、更具客观性,下一步的工作包括以下几点:

(1)将各种典型的语义不变变换的技术集成在一起,使我们的攻击工具可以方便地评估不同水印算法.

(2)开展攻击后语义损失评估技术的研究工作[24],以便评估攻击对载体文本造成的影响,确定

合理、可行的攻击力度范围.

参 考 文 献

[1] Atallah M J, Raskin V, Crogan M, Hempelmann C, Kerschbaum F, Mohamed D, Naik S. Natural language watermarking: Design, analysis, and a proof- of- concept implementation//Proceedings of the 4th Information Hiding. Pittsburgh, PA, 2001: 185-200

[2] Zhang Yu, Liu Ting, Chen Yi-Heng, Zhao Shi-Qi, Li Sheng. Natural language watermarking. Journal of Chinese Information Processing, 2005, 19(1): 56-62, 70(in Chinese)
(张宇, 刘挺, 陈毅恒, 赵世奇, 李生. 自然语言文本水印. 中文信息学报, 2005, 19(1): 56-62, 70)

[3] Stefan K, Fabien A P P. Information Hiding Techniques for Steganography and Digital Watermarking. Boston, USA: Artech House, Inc., 2000

[4] Jack T B, Steven L, Maxemchuk Nicholas F, O'Gorman Lawrence. Electronic marking and identification techniques to discourage document copying. IEEE Journal on Selected Areas in Communications, 1995, 13(8): 1495-1504

[5] Qi Wen-Fa, Li Xiao-Long, Yang Bin, Cheng Dao-Fang. Document watermarking scheme for information tracking. Journal on Communications, 2008, 29(10): 183-190(in Chinese)
(亓文法, 李晓龙, 杨斌, 程道放. 用于信息追踪的文本水印算法. 通信学报, 2008, 29(10): 183-190)

[6] Moulin P, Koetter R. Data-hiding codes. Proceedings of the IEEE, 2005, 93(12): 2083-2126

[7] Gupta Gaurav, Pieprzyk Josef, Wang Hua Xiong. An attack-localizing watermarking scheme for natural language documents//Proceedings of the ACM Symposium on Information, Computer and Communications Security'06. Taipei: ACM Press, 2006: 157-165

[8] Kerckhoffs A. La cryptography militaire. Journal des Sciences Militaires, 1983, 9(1): 5-83

[9] Ma Guang-Ping, Zhuang Ya-Xuan, He Lu et al. Active attacks on syntactic natural language steganography//Proceedings of the 9th China Information Hiding. Chengdu, China, 2010(in Chinese)
(马广平, 张雅轩, 何路等. 针对语法变换信息隐藏的主动攻击算法//第 9 届全国信息隐藏暨多媒体信息安全学术大会论文集. 成都, 中国, 2010)

[10] Yang J L, Wang J M, Wang C K, Li D Y. A novel scheme for watermarking natural language text//Proceedings of the Intelligent Information Hiding and Multimedia Signal Processing. Kaohsiung, Taiwan, China, 2007: 481-484

[11] Topkara U, Topkara M, Atallah J M. The hiding virtues of ambiguity: Quantifiably resilient watermarking of natural language text through synonym substitutions//Proceedings of the Multimedia and Security Workshop'06. 2006: 164-174

[12] Chiang Y L, Chang L P, Hsieh W T, Chen W C. Natural language watermarking using semantic substitution for Chinese text//Proceedings of the International Workshop on Digital Watermarking. Seoul, Korea, 2003. LNCS 2939. Springer, Heidelberg, 2004: 129-140

[13] Wu J, Stinson D R. Authorship proof for textual document//Proceedings of the 11th Information Hiding. Darmstadt, Germany, 2009: 209-223

[14] Atallah M J, Raskin V, Hempelmann C F et al. Natural language watermarking and tamperproofing//Proceedings of the 5th International Information Hiding Workshop. Noordwijkerhout, The Netherlands, 2002: 196-212

[15] Lu P, Lu Z, Zhou Z L, Gu J Z. An optimized natural language watermarking algorithm based on TMR//Proceedings of the 9th International Conference for Young Computer Scientists. Beijing, China, 2008: 1459-1463

[16] Topkara Mercan, Topkara Umut, Atallah Mikhail J. Words are not enough: Sentence level natural language watermarking//Proceedings of the 4th ACM International Workshop on Contents Protection and Security, California, USA, 2006: 37-46

[17] Gan Can. The research on natural language information hiding based on synonym substitution [M. S. dissertation]. Hunan University, Hunan, 2008(in Chinese)
(甘灿. 基于同义词替换的自然语言信息隐藏研究[硕士学位论文]. 湖南大学, 湖南, 2008)

[18] Chen Zhi-Li, Huang Liu-Sheng, Yu Zhen-Shan, Yang Wei, Chen Guo-Liang. An information hiding algorithm based on double text segments. Journal of Electronics & Information Technology, 2009, 31(11): 2725-2730(in Chinese)
(陈志立, 黄刘生, 余振山, 杨威, 陈国良. 基于双文本段的信息隐藏算法. 电子与信息学报, 2009, 31(11): 2725-2730)

[19] Dai Zu-Xu, Hong Fan, Cui Guo-Hua, Fu Min. Watermarking text document based on statistic property of part of speech string. Journal on Communications, 2007, 28(4): 108-113(in Chinese)
(戴祖旭, 洪帆, 崔国华, 付敏. 基于词性标记串统计特性的文本数字水印算法. 通信学报, 2007, 28(4): 108-113)

[20] Vybornova O, Macq B. Natural language watermarking and robust hashing based on presuppositional analysis//Proceedings of the IEEE International Conference on Information Reuse and Integration. LasVegas, 2007: 177-182

[21] He Lu, Gui Xiao-Lin. The robust natural language hash domain spread spectrum watermark coding algorithm [Z]. Chinese Patent. No. 2011102164892(in Chinese)
(何路, 桂小林. 鲁棒的自然语言哈希域扩频水印编码算法[Z]. 专利受理号 2011102164892)

[22] Jiang Chuan-Xian, Chen Xiao-Wei, Li Zhi. Robust text watermarking based on significant components. Acta Automatica Sinica, 2010, 36(9): 1250-1256(in Chinese)
(姜传贤, 陈孝威, 李智. 基于文本重要内容的鲁棒水印算法. 自动化学报, 2010, 36(9): 1250-1256)

[23] Miller M L, Bloom J A. Computing the probability of false watermark detection//Proceedings of the 3rd International Workshop on Information Hiding. Dresden, Germany, 1999: 146-158

[24] Zhang Ya-Xuan, He Lu, Fang Ding-Yi. A method of evaluate semantics loss caused by summary attack against text watermarking. Application Research of Computers, 2012, accepted (in Chinese)
(张雅轩，何路，房鼎益. 针对文本水印摘要攻击的语义损失量评估方法. 计算机应用研究，2012，已录用)

HE Lu, born in 1977, Ph. D. candidate, lecturer. His research interests include information hiding, text watermarking, text steganalysis, network and information security.

GUI Xiao-Lin, born in 1966, Ph. D., professor, Ph. D. supervisor. His research interests include information security, cloud and grid computing and trusted systems design.

TIAN Feng, born in 1985, Ph. D. candidate. His research interests focus on information security.

WU Rei-Feng, born in 1989, M. S. candidate. Her research interests include text watermarking, text steganalysis.

FANG Ding-Ying, born in 1959, Ph. D., professor, Ph. D. supervisor. His research interests include information and network security.

Background

Natural language watermarking can be utilized for copyright protection because the watermark can survive after common editing operation. It has been the almost hot research area in digital watermarking and information hiding. The robustness of NLW is very important because any security production should be inspecting carefully before business application. In image watermarking and sound watermarking area, a lot of papers studying the robustness of watermarking have been published. However, the natural of language is totally different from image signals. The analyzing methods of image watermarking robustness cannot be applied to natural language watermarking. Until now, the study of robustness of natural language watermarking is still absent. In this paper, we propose the robustness modal according analysis on adversary modal. Different from image or sound watermarking, which robust modals are built on signal processing techniques, our NLW robust modals are built by the combinatorics. This work is partly supported by National Natural Science Foundation of China (Nos. 61172090). These projects aim to provide a secure environment of cloud computing and ensure the information is exchanged in security. Our team has working on natural language information hiding and steganalysis, and building real system. About twenty papers have been published, some of them are indexed by SCI, and several patents have been authorized.