基于文本重要内容的鲁棒水印算法

姜传贤 [1] 陈孝威 [1] 李 智 [1]

摘 要 提出一种基于文本重要内容的鲁棒水印算法, 对文本的特征进行分析, 确定文本的重要内容. 根据水印序列和同义词替换评价模型, 将水印不可感知地嵌入到文本的重要内容中, 提高水印的鲁棒性. 投票原则的使用又进一步提高水印的鲁棒性并降低了误检率. 理论和实验分析表明, 采用文中算法嵌入的水印具有较好的鲁棒性、安全性和不可见性.

关键词 文本重要内容, 文本水印, 同义词, 主题句, 鲁棒水印

DOI 10.3724/SP.J.1004.2010.01250

Robust Text Watermarking Based on Significant Components

JIANG Chuan-Xian[1] CHEN Xiao-Wei[1] LI Zhi[1]

Abstract A robust text-watermarking scheme based on significant components is presented. The text significant components are acquired by analyzing text feature. According to watermarking sequence and the evaluation model of synonymy substitution, the watermark is imperceptibly embedded into the significant components of the text, which improves the robustness of watermarking. Furthermore, majority voting is used to improve the robustness of watermarking and reduce the probability of false positives. Theoretical analysis and experimental results have indicated that the proposed scheme can improve the robustness, security, and invisibility of watermarking.

Key words Text significant feature, text watermark, synonymy, topic sentence, robust watermarking

近年来, 随着网络通信和多媒体技术的发展, 数字产品的版权保护和内容认证越来越受到人们的重视. 数字水印作为保护数字产品版权和完整性的有效手段, 如何确保保护数字产品的版权信息具有鲁棒性和抗检测性是至关重要的.

数字图像作为信息交换的主流媒体, 自然成为信息隐藏的优秀载体. 视觉冗余的存在为数字图像提供了相对宽裕的携带信息空间, 因此基于数字图像信息隐藏[1−2] 的研究成为当今信息隐藏技术的主流. 然而文本存储简单、结构紧凑等独特优势使得文本在数据传真、文字识别、条码识别和数字签名中得到了广泛的应用. 但是文本有其自身的特点, 使得对文本的版权保护成为一个具有挑战和热点的研究课题. 目前文本的信息隐藏方法[3−8] 主要有: 1) 不可见字符和字体格式的信息隐藏[3−4]; 2) 文本语法与语义的信息隐藏[5−7], 如 Gupta 等[7] 提出了利用句子结构变换 (通过改变句子 “长短”, 即最低有效位 (Least significant bit, LSB)) 隐藏方法, 其主要思想是: 先对句子进行分组, 通过调整组内句子的长短 “LSB” 来实现水印嵌入, 然后根据海明距离, 通过投票原则来提取水印; 3) 文本图像的信息隐藏[8].

从以上方法可知, 文本语法与语义处理的信息隐藏方法嵌入的信息难以被检测, 同义词替换信息隐藏方法不仅有此优点, 而且实现简单, 并对自然语言处理技术要求较低 (对于其他语系, 如英语的文本只需准备一个同义词库, 对于处理中文文本, 需要准备同义词库和一个分词系统), 因此该方法具有很强的扩展能力和较好的抗检测能力的特点; 另一方面, 以上方法都没有将水印与文本载体的重要内容绑定[9], 这实际上对算法的安全性和鲁棒性是不利的, 当带水印的文本载体遭受到有意的 (如恶意的破坏或删除水印) 或无意的攻击行为 (如扫描与复印、噪声污染、尺寸变化等) 时, 攻击者往往可以很容易地在不破坏载体基本质量的情况下而去掉水印. 因此水印与文本载体重要内容捆绑在一起, 当文本载体的重要内容被破坏时, 水印才被破坏. 这样的水印技术是人们希望的, 并且能够达到保护版权的目的.

根据以上对文本水印方法的分析, 本文提出一种基于文本重要内容的鲁棒水印算法. 分析了文本的特征, 提出了文本主题词集概念并得出以下性质: 一个重要的句子是包含重要词的句子; 一个重要的词就是经常出现在重要句子中的词. 并给出了同义词替换评价模型. 通过分类投票表决原则和同义词替换将水印嵌入到文本的重要内容中, 从而使该算法既具有一定的扩展性、良好的不可见性和抗检测性的特点, 又能提高水印的鲁棒性和安全性.

收稿日期 2009-05-19 录用日期 2010-05-20
Manuscript received May 19, 2009; accepted May 20, 2010
贵州省自然科学基金 (QKH20052109, QKH20102257) 资助
Supported by the Science Research Foundation of Guizhou Province (QKH20052109, QKH20102257)

1. 贵州大学计算机科学与信息学院 贵阳 550025
1. School of Computer Science and Information, Guizhou University, Guiyang 550025

1 相关知识

一个中文文本可看成由句子和标点符号构成的集合, 而根据频率标准, 中文文本又可看成由高频词、中频词和低频词组成的集合. 主题的所谓有效词 (或称实词) 往往是中频词的特征[10], 而没有实际意义的词 (如: “的”、“得” 等) 往往是高频词的特征. 假设句子是不可分割的最小单元. 为叙述方便给出如下定义:

定义 1 (类停用词 SU, 实词 E). 汉语的词语按不同要求可以分成以下两类: 类停用词是一类可以穷尽地枚举的词语, 如: 叹词、助词、量词、介词、代词、连词和助动词; 实词是除去类停用词以外的词, 如名词、动词、形容词和副词, 而类停用词往往是高频词.

定义 2 (概念项 C). 同义词和近义词往往被映射为同一个概念, 比如: “爸爸” = “父亲”. 对于那些意义相近但是仅仅用词不同的文本, 使用概念项作为特征能够更好地表示文本的内容.

定义 3 (内容词集 M). $Textd$ 表示文本词的集合, 有集合 $E = Textd - SU$, 然后对 E 进行概念项处理形成文本的内容词概念集, 简称为内容词集 M.

定义 4 (主题词集 $Comp$). $Comp = \{p_i \mid p_i \in M \wedge Fre(p_i) > thr\}$, 其中 “$\wedge$” 是逻辑与, $Fre(w)$ 为词 w 的频率且 thr 是内容词的频率阈值, $w \in M$.

定义 5 (文本词频矩阵 WF). 设 $wf_{ij} = Fre(w_{ij})$, $(S_1, S_2, \cdots, S_i, \cdots, S_m)$ 表示文本的句子序列, $(w_{i1}, w_{i2}, \cdots, w_{ij}, \cdots, w_{in})$ 和 $(wf_{i1}, wf_{i2}, \cdots, wf_{ij}, \cdots, wf_{in})$ 分别表示句子 S_i 的词序列和词频序列, $w_{ij} \in M$, 则文本词频矩阵为

$$WF = \begin{bmatrix} wf_{11} & \cdots & wf_{1j} & \cdots & wf_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ wf_{i1} & \cdots & wf_{ij} & \cdots & wf_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ wf_{m1} & \cdots & wf_{mj} & \cdots & wf_{mn} \end{bmatrix}$$

定义 6. 每一个句子和每一个词特征项都定义一个权值. 句子 S_i 的权值 WS_i 反映句子对整篇文本语义的贡献程度, 权值越大的句子越能反映文本语义的中心, 概括程度越强; 词 w_i 的权重 Ww_i 表示该词对于整个句子语义的贡献程度, 权值越大的词越能反映句子的中心, 当然是重要的词, 从而得出句子和词的重要性之间存在如下性质:

1) 一个重要的句子是包含重要词的句子;

2) 一个重要的词就是经常出现在重要句子中的词.

以上性质可以理解为对重要性的一个循环定义, 无法各自独立地定义句子和词的重要性. “投票原则 (Sergey Brin 和 Larry Page 提出 PageRank 算法)” 给了我们很大的启发. 对于这种循环, 首先用北京大学计算语言学研究所开发的汉语词语切分与词性标记软件产生分词序列, 再用定义 5 来产生出文本的词频矩阵, 用迭代方法[11] 计算文本句子和词的权重来证实此性质的正确性.

1.1 同义词替换评价模型

对于普通的自然语言文本, 其中的词可以用对应的同义词替换掉, 但这种替换可能影响文本含义, 因此需引入自然语言处理技术成果尽量降低这种替换所带来的影响.

定义 7. 设集合满足: $Syn_g = \{w \mid \mathrm{Mean}(w) \approx g, w \in E\}$, 则称 Syn_g 是词义为 g 的同义词组, 即同义词组为意义相同或相近词的集合, $|Syn_g|$ 表示词的数目, 设 $\mathrm{Mean}(w)$ 为 w 的词义, 其中 $w \in E$.

以梅家驹先生的《同义词词林》[12] 为基础, 从中提取出完全可替换和不完全可替换的词组, 建立成一个同义词库. 词林中存在很多生僻词, 有的同义词组内词之间使用频率相差很大, 这样会降低算法的不可见性. 因此根据现代汉语词语的词频统计表, 调整同义词库, 使每组同义词的词频相近. 同时为使每组同义词不相交 (编解码不出错), 删除了重复出现的同义词, 保证每一个同义词只出现一次, 形成同义词库. 在此同义词库基础上, 利用《知网》的词的义原[13] 对同义词进一步聚类[14], 形成较合理的集合 SD, 即 $SD = \{Syn_{g_1}, Syn_{g_2}, \cdots, Syn_{g_m}\}$, 对每一同义词组中的词进行编码[15]. 这样 SD 就成为一个同义词库.

定义 8 (同义词特征 SF). 根据同义词库 SD 和主题词集, 对句子 S 进行主题词和同义词检测, 如果能够检测出主题词和同义词, 则得出句中主题词 $S.p$, 取出其最高频率词记 $\mathrm{Max}f(S.p)$, 除 $\mathrm{Max}f(S.p)$ 以外同义词组 $S.x = \{x_i | i \in [1, \mathrm{num}(S.x)]\}$ 及它的个数 $\mathrm{num}(S.x)$, 同义词元素 $S.x_i$ 所在句子 S 中的位置 $\mathrm{locate}(S.x_i)$ 和同义词库中的词的编码值 $\mathrm{value}(S.x_i)$, 其中 $S.p \subset M, S.x \subset M$; 否则忽略.

为了提高水印的不可见性和抗检测性, 引入自然语言处理技术成果 (搭配特征和同现特征). 对于同义词 w 属于歧义词组的, 其上下文语境是指句子中在其周围出现的词的集合. 实际上, w 的语义特征的确定只跟集合中的部分词相关, 而与其他的词不相关. 因此若抽取出相关的词, 就可以缩小上下文的范围, 降低问题的复杂度. 由于词性特征之间有较确定的搭配关系, 因此可以通过对句子进行句法分析, 再根据句法搭配关系从中得到与 w 相关的词,

称为 w 的搭配词[16]. 因此可以通过依存句法分析来获取同义词的搭配词. 依存句法由法国语言学家 L. Tesniere 最先提出, 描述各个词语之间的依存关系. 例如 "学校开展丰富多彩的校园文化活动." 的依存句法分析如图 1 所示 (使用哈工大信息检索研究室语言技术平台[15]):

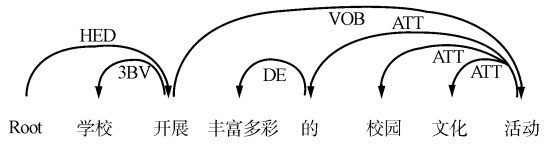


图 1 依存句法分析

Fig. 1 Sentence dependency structure

从图 1 可以看出, 词 "开展" 与 "学校", "活动", 构成搭配关系, 故 "学校", "活动" 是 "开展" 的搭配词.

假设句子 S 的词序列为 $w_1,\cdots,s_1,\cdots,w_n,s_1$ 属于歧义词组 $\{s_1,s_2\}$, s_1 的搭配词集为 $W_{de}=\{w_1^{'},\cdots,w_i^{'},w_m^{'}\}$, 这里 $w_i^{'}\in S$, $m\leq n$. 令 $P(s_2|w_i^{'})$ 是在已知搭配词 $w_i^{'}$ 的条件下 s_2 出现的概率, 则

$$P(s_2\mid w_i^{'})=\frac{P(s_2,w_i^{'})}{P(w_i^{'})} \tag{1}$$

通过对实际的汉语语料进行统计, 可以得到上式中的参数. 如果 $\mathrm{Count}(s_2,w_i^{'})$ 表示 s_2 与 $w_i^{'}$ 互为搭配词在训练语料中出现的次数, $\mathrm{Count}(w_i^{'})$ 表示 $w_i^{'}$ 在训练语料中出现的次数, 根据最大似然估计原理, 可以近似地认为

$$P(s_2\mid w_i^{'})=\frac{\mathrm{Count}(s_2,w_i^{'})}{\mathrm{Count}(w_i^{'})} \tag{2}$$

令 $\mathrm{Sum}P_{s_2}$ 为 s_2 与 W_{de} 中各个搭配词共现的概率之和, 即

$$\mathrm{Sum}P_{s_2}=\sum_{i=1}^{m}\frac{\mathrm{Count}(s_2,w_i^{'})}{\mathrm{Count}(w_i^{'})} \tag{3}$$

则计算 s_1 所在的同义词组中所有 $\mathrm{Sum}P_{s_2}$ 的值, 找出最大的 $\mathrm{Sum}P_{s_2}$, 然后将 s_2 替换 s_1.

1.1.1 同义词替换算法

先对文本载体进行分词和词性标注 (使用北京大学计算语言学研究所开发的汉语词语切分与词性标记软件), 再根据同义词库 SD 和式 (3) 进行同义词替换隐藏信息. 算法如下:

算法 1. 同义词替换

步骤 1. 对文本载体句子中的该同义词 s, 将其所在的句子进行依存句法分析, 抽取搭配词集 W_{de};

步骤 2. 根据式 (3) 和式 (2) 计算训练语料中得到的参数, 分别计算 s 所在同义词组其他词 $s_i^{'}$ 的所有值, 找出最大的 $\mathrm{Sum}P_{s_i^{'}}$ 的词 $s_i^{'}$ 替换 s.

2 水印嵌入和提取算法

设 $|\cdot|$ 是求集合元素的个数, !value 为非编号值, $KEY[\,]$ 表示数组并赋初值 $KEY[\,]\leftarrow$!value, mod 是求余, $key1$ 为给主题词伪随机编号的密钥值, $\mathrm{index}(x)=\mathrm{hash}(x,key1)$, 其中, $\mathrm{hash}(x,y)$ 是单向函数[17].

2.1 水印嵌入算法

其思想是: 首先对文本载体进行分词和词性标注, 根据主题词集定义和定义 6, 找出主题词, 然后找出包含有主题词的主题句集 (CS), 再通过同义词特征定义 8 过滤句子, 得出子主题句集 ($subCS$), 再分类得子集 (as), 通过同义词替换将水印嵌入其中. 水印嵌入流程见图 2.

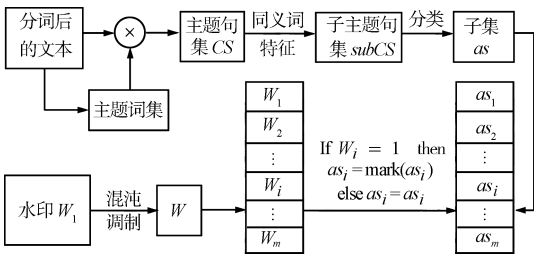


图 2 水印嵌入算法流程

Fig. 2 Watermark embedding

水印嵌入的具体步骤如下:

1) 水印信号生成

使用拥有者和用户信息产生二进制水印序列 W_1, 用密钥 $key1$ 生成二进制混沌序列 L, 它们的长度为 m, 由 L 对 W_1 进行混沌调制, 得待嵌入水印信号 W, $W_i=L_i\oplus W_{1i}$, $0\leq i<m$, 这里 "$\oplus$" 表示异或.

2) 水印嵌入

步骤 1. 对文本载体进行分词和词性标注 (使用北京大学计算语言学研究所开发的汉语词语切分与词性标记软件);

步骤 2. 根据主题词集定义和定义 6, 过滤出包含主题词的所有句子集合 CS, 根据同义词特征过滤出含同义词的子主题句集 $subCS=\{S_0,S_1,\cdots,S_n\}$;

步骤 3. 根据句子的主题词集, 对 $subCS$ 进行分类, 得子集 $as=\{as_i\mid 0\leq i<m\}$;

for $(j=0;j<\mid subCS\mid;j++)$ //分类

{　$i = \text{index}(\text{Max}f(S_j.p))\text{mod}\ m; as_{ij} = S_j$;
//句子的主题词集 $S_j.p$ }

步骤 4. for $(i = 0; i < m; i{+}{+})$
{　if $(W_i == 1)$ $mark(as_i)$;} //将每位水印信息嵌入到文本载体中.
$subroutine\ mark(as_i)$
{for $(j = 0; j <\mid as_i \mid; j{+}{+})$
{1) 通过同义词特征定义 8, 得出句子 as_{ij} 的 $as_{ij}.x$, $\text{num}(as_{ij}.x)$, $\text{Max}f(as_{ij}.p)$ 和 $\text{value}(as_{ij}.x_\Upsilon)$;
2) 使用算法 1 替换第 Υ 个同义词, 得
$KEY[\text{hash}(\text{index}(\text{Max}f(as_{ij}.p)), \text{value}(as_{ij}.x_\Upsilon))]$
$= \text{hash}(\text{index}(\text{Max}f(as_{ij}.p)), \text{value}(as_{ij}.x_\Upsilon))$;
//KEY 作为提取水印的密钥.
其中 $x = \{x_\Upsilon \mid 1 \le \Upsilon \le \text{num}(as_{ij}.x)\}$,
$\Upsilon = \text{index}(as_{ij}.p)\text{mod}(\text{num}(as_{ij}.x)) + 1;\}\}$.

2.2　水印提取算法

水印提取基本上是水印嵌入的逆过程:

步骤 1. 利用第 2.1 节中 "2) 水印嵌入" 的前 3 步可以从水印文本载体中得出子主题句集 $subCS' = \{S_0', S_1', \cdots, S_n'\}$ 和子集 $as' = \{as_i' \mid 0 \le i < m\}$;

步骤 2. for $(i = 0; i <\mid as' \mid; i{+}{+})$ $W'[\,] \leftarrow \text{extract}(as_i')$; //提取水印信息.

步骤 3. 对提取的水印 W' 通过第 2.1 节中 "1) 水印信号生成" 逆混沌调制生成提取后的水印 W_1'.
subroutine extract(as_i') return integer
{ for $(j = 0; j <\mid as_i' \mid; j++)$
{1) 根据同义词特征, 得出句子 as_{ij}' 的 $as_{ij}'.x$, $\text{num}(as_{ij}'.x), \text{Max}f(as_{ij}'.p)$ 和 $\text{value}(as_{ij}'.x_\Upsilon)$;
2) if $(KEY[\text{hash}(\text{index}(\text{Max}f(as_{ij}'.p)), \text{value}(as_{ij}'.x_\Upsilon))] == \text{hash}(\text{index}(\text{Max}f(as_{ij}'.p)), \text{value}(as_{ij}'.x_\Upsilon))$ $temp[\,] \leftarrow 1$;
else $temp[\,] \leftarrow 0$;
其中, $x = \{x_\Upsilon \mid 1 \le \Upsilon \le \text{num}(as_{ij}'.x)\}$,
$\Upsilon = \text{index}(\text{Max}f(as_{ij}'.p)\text{mod}(\text{num}(as_{ij}'.x)) + 1\}$
return $majority\_voting(temp[\,])$;
//$majority\_voting()$ 是投票表决函数}.

为了能更好地理解水印嵌入和提取算法, 下面给出水印嵌入的具体例子.

同义词编码的介绍:《同义词词林》[12] 提供了三层编码, 即一层用大写英文字母表示, 二层用小写英文字母表示, 三层用二位十进制整数表示. 哈尔滨工业大学同义词词林的扩展版[15] 新增了两层编码. 我们在哈尔滨工业大学的扩展版的基础上, 做了稍微的修改, 新增了一层. 新增的三层编码与原有的三层编码构成一个 6 层的编码, 能唯一标识库存中的同义词. 编码的方法说明如下: 4 层用大写英文字母标识, 5 层用二位十进制整数标识, 5 层表示同义词组, 6 层用二位十进制整数标识同义词组中的元素. 例如, "Je05B014" 标识该同义词组 {寄予, 寄托, 依托, 委以}, 用 {Je05B0101, Je05B0102, Je05B0103, Je05B0104}分别标识组内的词, 而 "Je05B01" 表示该同义词组的编码, "4" 表示同义词的数目.

以此句 "库克还阐述英国支持美国对伊拉克的强硬立场, 以迫使伊拉克让联合国武器核查小组不受限制地在其境内活动." 为例, 根据依存句法分析, 定义 8 和频率阈值取 $thr > 386$, 可获得该句的主题词集及频率 {"美国及 1 022, 活动及 1 025"}和同义词集 {"阐述, 迫使, 限制"}, 得出所在同义词库中的编码及同义词组: "Hi14C016 {阐述, 阐明, 阐发, 发明, 表明, 申明}"; "Hi56C0110{迫使, 强迫, 逼迫, 强逼, 强使, 驱使, 驱策, 催逼, 强求, 紧逼}"; "Je080033 {限制, 约束, 制止}". 取句子的主题词集的最高频率 (1 025) 对同义词的个数取余, 得 $\Upsilon = 3$, 确定 "限制" 作为此句可替换的对象及它的依存句法分析截图见图 3 (使用哈尔滨工业大学信息检索研究中心语言技术平台[15]).

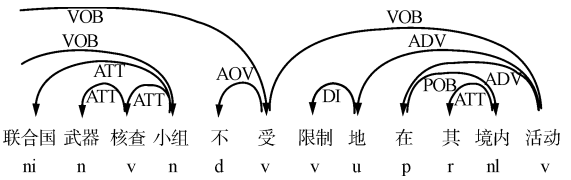


图 3　依存句法分析截图

Fig. 3　The part of sentence dependency structure

根据图 3 得到它的搭配词集为 {"活动"}. 然后根据式 (1), 计算同义词 "约束", "制止" 和搭配词 "活动" 的参数 $\text{Sum}P$ 分别为 $0.000, 965, 0.002, 903$. 根据算法 1 完成同义词替换后得 "库克还阐述英国支持美国对伊拉克的强硬立场, 以迫使伊拉克让联合国武器核查小组不受制止地在其境内活动." 和得 $KEY[\text{hash}(\text{index}(1\,025), \text{Je0800303})]$= $\text{hash}(\text{index}(1\,025), \text{Je0800303})$. 水印的提取是水印嵌入的逆过程, 限于篇幅在此不作叙述.

3　算法分析

设在自然语言文本 T 中有 n 个句子被嵌入水印, 然从自然语言文本 T^* 中检测出 i 个匹配信息, 则称 T^* 是 T 的一个可疑的副本. 若检测匹配信息过多或者过少, 则 T^* 一定是盗版.

检测水印信息有两种可能结果: "匹配" 或 "不匹配", 其概率为 1/2, 则其分布规律可以看成是 0-1 分布. 若从 k 个待检测的数据检测出了 i 个匹配信息, 实际可以看作是一个成功 i 次, 失败 $k - i$ 次的

k 重贝努利实验. 因此在投票序列中 i 个数据元素匹配的概率为

$$b\left(i,k,\frac{1}{2}\right)=\binom{i}{k}\left(\frac{1}{2}\right)^{i}\left(1-\frac{1}{2}\right)^{k-i} \tag{4}$$

如果投票序列中有至少 $k/2$ 个元素发生变化, 此时投票序列的投票结果就会发生变化, 所以水印匹配的概率为

$$B\left(i,k,\frac{1}{2}\right)=\sum_{i=0}^{\frac{k}{2}}b\left(i,k,\frac{1}{2}\right) \tag{5}$$

显然它的概率值是 1/2. 为判定 T^* 的版权, 用以下概率表达式来判定.

$$\tau=\max\{t\in[0,n/2]\mid\sum_{i=t}^{n-t}b\left(i,k,\frac{1}{2}\right)\geq 1-\delta\} \tag{6}$$

其中

$$b(i,k,p)=\binom{i}{k}(p)^{i}(1-p)^{k-i} \tag{7}$$

选取一个适当的值 $\delta\,(0<\delta<1)$ 作为置信因子.

3.1 各种攻击分析

常见的自然语言文本水印攻击[7] 有句子交换、段交换、文本再生、句子结构变换、同义词替换、插入新句子和句子删除. 设 $\lceil\cdot\rceil$ 是上整函数.

1) 句子交换、段交换攻击、文本再生攻击和句子结构变换分别对水印的影响

由本文算法可知, 句子交换和段交换攻击对水印无影响; 文本再生是改变这些格式 (如句子移动、字间距、字体格式、空格等) 来攻击文本载体, 修改这些格式对水印无影响; 句子结构变换 (如主动句变被动句, 肯定句变双重否定句等) 是通过添加 (去掉) 类停用词来实现句子变换. 而本文算法是作用在文本的实词上实现的水印算法, 因此对水印无影响.

2) 同义词替换对水印的影响

设攻击者替换同义词的比率为 p,m 是嵌入水印后某个子集的元素数目, 则成功将一位匹配信息修改成非匹配信息的概率为 $1/2p$. 根据二项分布 $\mathrm{B}(m,1/2p)$, 得通过同义词替换去掉该子集内 i 个匹配信息的概率为

$$\mathrm{B}\left(i,m,\frac{1}{2}p\right)=\binom{i}{m}\left(\frac{1}{2}p\right)^{i}\left(1-\frac{1}{2}p\right)^{m-i} \tag{8}$$

当 i 大于 $m/2$ 时, 子集序列的投票结果就发生了变化, 因此该子集水印信息被成功攻击的概率为

$$wp=\mathrm{B}\left(i,m,\frac{1}{2}p\right)=\sum_{i=\lceil\frac{m}{2}\rceil}^{m}\mathrm{B}\left(i,m,\frac{1}{2}p\right) \tag{9}$$

注意: 当 $p=1$ 时, 该攻击就变成了随机同义词替换攻击, 此时式 (9) 等于 1/2.

3) 插入新句子攻击对水印的影响

设某个投票序列子集包含有 q 位匹配信息, 根据本文算法, 当将 0 到 $q-1$ 句子插入到该子集中, 此情况下, 子集的投票结果保持不变; 当将 q 个句子插入到该子集中, 水印改变概率是 1/2; 当插入 $q+1$ 个句子到该子集中且其不含有匹配信息, 水印转移概率是 $(1/2)^{q+1}$; 同理当 $q+i$ 个句子插入到该子集 (它包含 q 位匹配信息) 中, 水印转移概率是 $\lceil i/2\rceil(1/2)^{q+1}$, 因此总的水印转移概率, 即该子集水印被成功攻击的概率为

$$\begin{aligned}wp=&\left(\frac{1}{2}\right)^{q+1}+\left(\frac{1}{2}\right)^{q+2}+2\left(\frac{1}{2}\right)^{q+3}+2\left(\frac{1}{2}\right)^{q+4}+\cdots+\\&\lceil\frac{i}{2}\rceil\left(\frac{1}{2}\right)^{q+i}+\lceil\frac{i+1}{2}\rceil\left(\frac{1}{2}\right)^{q+i+1}+\cdots=\\&\frac{4}{3}\left(\frac{1}{2}\right)^{q},\quad i\to\infty\end{aligned} \tag{10}$$

从式 (10) 可以得出: 子集中的句子元素数目越多, 抗攻击能力越强.

4) 句子删除攻击对水印的影响

句子删除是删除文本中某些句子. 设 c 表示子集的数目, 某个子集的投票序列包含 m 位匹配信息. 删除文本 k 个句子, 当 $k<m$ 时, 虽然改变了子集的元素个数, 但不改变子集的投票结果, 所以此子集的水印值保持不变; 当删除句子 k 大于等于 m, 则该子集水印被攻击的概率为 wp. 为了分析本文水印算法的鲁棒性, 给出图 4 (设 $n=1\,000, c=4, k$ 取 $10\sim80$, $m=3,4,5$ 和三种方案: 1 位水印嵌入 3 个句子中; 1 位水印嵌入 4 个句子中; 1 位水印嵌入 5 个句子中), 可以得出类中元素数目越多, 抗攻击能力越强.

$$wp=\begin{cases}\dfrac{\binom{1}{c}\cdot\left(\binom{k-1\cdot m}{n-1\cdot m}-\sum\limits_{t=1}^{\lfloor\frac{k}{m}\rfloor-1}\cdot\binom{t}{c-1}\cdot\binom{k-(1+t)m}{n-cm}\right)}{\binom{k}{n}}, & m\leq k\leq cm\\ \dfrac{\binom{1}{c}\cdot\left(\binom{k-1\cdot m}{n-1\cdot m}+(-1)^{t}\sum\limits_{t=1}^{c-1}\cdot\binom{t}{c-1}\cdot\binom{k-(1+t)m}{n-(1+t)m}\right)}{\binom{k}{n}}, & k>cm\end{cases} \tag{11}$$

其中

$$\binom{m}{n}=\frac{n!}{m!(n-m)!} \tag{12}$$

以上攻击都假设在等概率随机攻击的条件下而进行分析所得的结果. 从攻击者希望破坏水印的同时, 又能较好地保持文本载体的正常使用价值的角度出发, 这样攻击往往不会针对载体的重要内容, 本文算法实际的性能要优于以上分析的结果. 为了更好应对具体攻击 (句子插入和句子删除), 提高算法的鲁棒性, 可以对第 2 节 (水印嵌入和提取算法) 进行改进: 在原水印算法的基础上, 对嵌入水印的子主题句集 $subCS$ 所有句子进行编号, 增加一个空间用来存放这些编号; 当水印提取时, 只需得出编号对应的句子, 进行水印提取. 此时改进的算法能够较好抗插入新句子攻击, 并且也增强了抗句子删除攻击的能力. 但需增加一个空间的开销来换取水印的鲁棒性, 因此, 根据不同的应用情况, 使用改进的水印算法. 另外, 当子集中只包括一个句子元素时, 本水印算法具有半脆弱认证的功能.

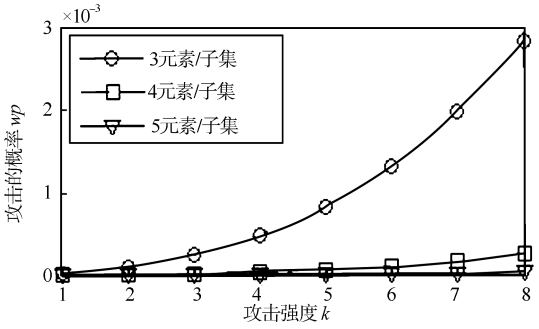


图 4　攻击强度增强水印鲁棒性反映

Fig. 4　The relationship between the attack intensity and the watermarking robustness

3.2　误检率

从没有匹配信息的文本中随机选取并检测出 n 位匹配信息, 误检概率为 2^{-n}. 然而本文算法是将 1 位水印嵌入到 m 个句子中, 并要求在 m 个句子中至少检测出 $m/2+1$ 匹配信息, 才能提取这 1 位水印信息. 所以本文算法的实际误检率为 $2^{-n}\cdot 2^{-(m/2+1)}$ 与文献 [7] 基本相同, 但是远低于文献 [6].

4　仿真实验

仿真实验数据的训练语料来自人民日报 1998 年 1 月的免费语料, 共约 956 973 个词. 对训练语料进行自动分词和依存句法分析, 实验使用的句法分析器是哈尔滨工业大学的依存句法分析器. 根据依存句法树抽取同义词库中歧义同义词的词语搭配对, 计算式 (2) 可以得到一个参数库, 采用同义词库 SD 和取 $thr > 386$, 在 Windows XP 平台上, 使用 C 语言软件进行实验. 通过反复实验, 给出部分实验结果见表 1.

从表 1 可以看出水印嵌入对载体数据改变很小. 随着水印位数据的增加, 同义词替换个数并没有多大的改变, 主要在于子集数目的改变 (也就是说水印位数的增加, 子集数目越多, 子集包含的元素越少). 另一方面由于当前自然语言处理技术 (汉语词语切分和汉语依存句法分析) 的处理精度和同义词库的构建也影响本文算法的实验结果.

表 1　水印嵌入对文本的影响

Table 1　Text responses with increasing of watermark size

水印大小 (位)	替换同义词 (个)	文本词修改比率
100	417	417/956 973
150	423	423/956 973

4.1　与现有算法的比较

本文算法抗各种攻击[7] 表现出较好的鲁棒性. 与文献 [4−6] 水印算法相比, 表现出更强的鲁棒性和安全性.

在 n 位水印序列 W ($n/2$ 个 "0" 和 "1" 组成水印序列) 和待嵌水印的某子集含 m 个句子的条件下, 给出本文算法与文献 [7] 比较: 1) 算法的鲁棒性见表 2; 2) 算法的不可见性: 嵌入水印序列 W, 文献 [7] 算法改变 $n\cdot m$ 个句子, 而本文算法改变 $n\cdot m/2$ 个句子; 3) 本文算法实现了水印与载体重要内容的捆绑, 提高了水印的安全性和鲁棒性, 而文献 [7] 没有此性质; 4) 本文算法与文献 [7] 相比有一定的扩展性, 文献 [7] 算法用于英语语言体系, 不易扩展到其他语言体系中应用.

表 2　算法鲁棒性比较

Table 2　Comparison of robustness with the existing algorithm

攻击方式	本文算法	文献 [7] 算法
句子交换	无影响	能交换 $\left(\frac{m}{2}-1\right)$ 句子
段交换	无影响	无影响
文本再生	无影响	无影响
句子结构交换	无影响	能交换 $\left(\frac{m}{2}-1\right)$ 句子
同义词替换	能替换 $\left(\frac{m}{2}-1\right)$ 句子	无影响
插入新句子	能插入 $(m-1)$ 句子	能插入 $\left(\frac{m}{2}-1\right)$ 句子
句子删除	能删除 $(m-1)$ 句子	能删除 $\left(\frac{m}{2}-1\right)$ 句子

5　结论

针对文本水印各类算法和水印与文本载体内容绑定的问题, 本文提出了基于文本载体重要内容的鲁棒水印算法, 与现有水印算法比较, 主要具有以下特点: 1) 具有较好的不可见性, 根据同义词替换评价模型和水印嵌入方式, 提高了文本水印的不可见性; 2) 具有较好的鲁棒性, 利用了水印作用在实

词上且与文本载体重要内容绑定和投票原则的使用, 增强了本文算法的鲁棒性、安全性和降低了误检率, 利用混沌序列调制水印及文本载体进行伪随机分类, 从水印和载体两方面, 进一步增强了本文算法的安全性; 3) 在提取水印时, 既不需要原始文本载体, 也不需要原始水印, 增强了本文算法的实际应用能力. 如何在自然语言文本较小带宽的条件下, 增大水印的容量是下一步的工作.

References

1 Wang Xiang-Yang, Hou Li-Min, Wu Jun. Feature-based digital image watermarking scheme robust to geometric attacks. *Acta Automatica Sinica*, 2008, **34**(1): 1−6
(王向阳, 侯丽敏, 邬俊. 基于图像特征点的强鲁棒数字水印嵌入方案. 自动化学报, 2008, **34**(1): 1−6)

2 Deng Cheng, Li Jie, Gao Xin-Bo. Geometric attacks resistant image watermarking in affine covariant regions. *Acta Automatica Sinica*, 2010, **36**(2): 221−228
(邓成, 李洁, 高新波. 基于仿射协变区域的抗几何攻击图像水印算法. 自动化学报, 2010, **36**(2): 221−228)

3 Zhang Yu, Liu Ting, Chen Yi-Heng, Zhao Shi-Qi, Li Sheng. Natural language watermarking. *Journal of Chinese Information Processing*, 2005, **19**(1): 56−70
(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): 56−70)

4 Yang H J, Kot A C. Pattern-based data hiding for binary image authentication by connectivity-preserving. *IEEE Transactions on Multimedia*, 2007, **9**(3): 475−486

5 Kankanhalli M S, Hau K F. Watermarking of electronic text documents. *Electronic Commerce Research*, 2002, **2**(1-2): 169−187

6 Atallah M J, Raskin V, Crogan M, Hempelmann C, Kerschbaum F, Mohamed D. Natural language watermarking: design, analysis, and a proof-of-concept implementation. In: Proceedings of the 4th International Workshop on Information Hiding. Pittsburgh, USA: Springer, 2001. 185−200

7 Gupta G, Pieprzyk J, Wang H X. An attack-localizing watermarking scheme for natural language documents. In: Proceedings of the ACM Symposium on Information, Computer and Communications Security. Taipei, China: ACM, 2006. 157−165

8 Li Zhao-Hong, Hou Jian-Jun, Song Wei. Binary document image authentication watermarking technique based on hierarchical structure. *Acta Automatica Sinica*, 2008, **34**(8): 841−848
(李赵红, 侯建军, 宋伟. 基于等级结构的二值文本图像认证水印算法. 自动化学报, 2008, **34**(8): 841−848)

9 Cox I J, Kilian J, Leighton F T, Shamoon T. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 1997, **6**(12): 1673−1687

10 Luhn H P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 1958, **2**(2): 159−165

11 Pu Dong-Bo. Clustering Classification Techniques and Its Application in the Field of Text Mining [Ph. D. dissertation], Institute of Computing Technology, Chinese Academy of Sciences, China, 2000
(卜东波. 聚类/分类理论研究及其在文本挖掘中的应用 [博士学位论文], 中国科学院计算技术研究所, 中国, 2000)

12 Mei Jia-Ju, Zhu Yi-Ming, Gao Yun-Qi, Yin Hong-Xiang. *Tongyici Cilin*. Shanghai: Shanghai Lexicographical Press, 1983
(梅家驹, 竺一鸣, 高蕴琦, 殷鸿翔. 同义词词林. 上海: 上海辞书出版社, 1983)

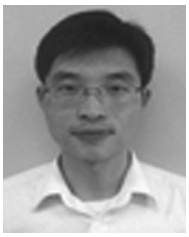
13 Dong Zhen-Dong. HowNet [Online], available: http://www.keenage.com/zhiwang/c zhiwang.html, October 26, 2009
(董振东. 知网 [Online], available: http://www.keenage.com/zhiwang/c zhiwang.html, October 26, 2009)

14 Mei Li-Jun, Zhou Qiang, Zang Lu, Chen Zu-Shun. Merge information in HowNet and Tongyici CiLin. *Journal of Chinese Information Processing*, 2005, **19**(1): 63−70
(梅立军, 周强, 臧路, 陈祖瞬. 知网与同义词词林的信息融合研究. 中文信息学报, 2005, **19**(1): 63−70)

15 Tongyici Cilin (Exended) [Online], available: http://ir.hit.edu.cn/demo/ltp/Sharing Plan.htm, October 26, 2009
(同义词词林 (扩展版) [Online], available: http://ir.hit.edu.cn/demo/ltp/Sharing Plan.htm, October 26, 2009)

16 Gan Can, Sun Xing-Ming, Liu Yu-Ling, Xiang Ling-Yun. An improved steganographic algorithm based on synonymy substitution for Chinese text. *Journal of Southeast University (Natural Science Edition)*, 2007, **37**(z): 137−140
(甘灿, 孙星明, 刘玉玲, 向凌云. 一种改进的基于同义词替换的中文文本信息隐藏方法. 东南大学学报 (自然科学版), 2007, **37**(z): 137−140)

17 Atallah M J, Wagstaff S S. Watermarking with quadratic residues. In: Proceedings of the Conference on Security and Watermarking of Multimedia Contents. San Jose, USA: SPIE, 1999. 283−288



姜传贤 贵州大学计算机科学与信息学院博士研究生. 主要研究方向为数字图像处理, 模式识别和数字水印. 本文通信作者. E-mail: emailfeibai@163.com
(**JIANG Chuan-Xian** Ph. D. candidate at the School of Computer Science and Information, Guizhou University. His research interest covers digital image processing, pattern recognition, and digital watermarking. Corresponding author of this paper.)

陈孝威 贵州大学计算机科学与信息学院教授. 主要研究方向为数字图像处理和计算机视觉.
E-mail: gzu@vip.sina.com
(**CHEN Xiao-Wei** Professor at the School of Computer Science and Information, Guizhou University. His research interest covers digital image processing and computer vision.)

李 智 贵州大学计算机科学与信息学院博士研究生. 主要研究方向为数字图像处理和数字水印.
E-mail: lizhigzu@163.com
(**LI Zhi** Ph. D. candidate at the School of Computer Science and Information, Guizhou University. Her research interest covers digital image processing and digital watermarking.)