# 基于同义词扩展和标签传递机制的文本无载体信息隐藏方法

张祯，倪嘉铭，姚晔，龚礼春，王玉娟，吴国华
（杭州电子科技大学网络空间安全学院，浙江 杭州 310018）

**摘　要**：为了解决传统文本信息隐藏方法隐藏容量低、抗检测性弱等问题，提出一种基于同义词扩展和标签传递机制的文本无载体信息隐藏方法。该方法将秘密信息切分成若干个关键词，分别嵌入不同的文本载体中。考虑到汉语中同义词的可替换性，使用基于知网的词语相似度计算方法筛选符合上下文语义的同义词，尽可能将不同的关键词或其同义词嵌入同一文本中，以此来提升隐藏容量。同时，记录每篇文本中所有被隐藏关键词的标签信息，将其按固定格式转换为二进制序列，作为秘密信息生成载体文本。最后，将自然载体文本和生成载体文本一起发送给接收方完成秘密通信。实验结果表明，所提方法的隐藏容量与传统方法相比有了较大的提升，具有较强的抗检测能力，使用小型载体文本数据库即可实现完整的信息隐藏和提取，减少了建立和维护索引数据库的开销。

**关键词**：隐蔽通信；无载体信息隐藏；关键词扩展；标签传递机制
**中图分类号**：TP309
**文献标识码**：A
**DOI:** 10.11959/j.issn.1000−436x.2021139

# Text coverless information hiding method based on synonyms expansion and label delivery mechanism

ZHANG Zhen, NI Jiaming, YAO Ye, GONG Lichun, WANG Yujuan, WU Guohua
School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, China

**Abstract:** To solve the problems of low hiding capacity and weak detection resistance, a method of coverless information hiding based on synonyms expansion and label delivery mechanism was proposed. To begin the proposed method, the secret information were divided into several keywords, then those keywords were embedded in different carrier texts respectively. Considering the interchangeability of synonyms in Chinese, using the word similarity calculation method based on *HowNet*, it was possible to filter synonyms which met the contextual semantics. For the purpose of increasing hiding capacity, various keywords or their synonyms were embedded in the same carrier text as much as possible. At the same moment, when recording the positional parameters of the secret information in each carrier text, those parameters were converted into a binary sequence in a fixed format, which was used to generate carrier text. Finally, the carrier texts and generated text were sent to the receiver. Comparison of the results with those of other studies confirm that the proposed method not only makes great improvement in hiding capacity and detection resistance, but also reduces the overhead of establishing and maintaining databases by using small text database.

**Keywords:** covert communication, coverless information hiding, synonyms expansion, label delivery mechanism

## 1　引言

信息技术和移动通信的应用普及，使人们越来越依赖各种数字媒体来完成日常的通信和交流任务。然而，信息的数字化也使其面临恶意攻击、非授权访问、窃听和伪造等风险。保障信息安全的主要技术有信息隐藏和信息加密。区别于信息加密技术，信息隐藏技术将秘密信息嵌入公开载体

**收稿日期**：2021−04−22；**修回日期**：2021−06−15
**通信作者**：姚晔，yaoye@hdu.edu.cn
**基金项目**：国家自然科学基金资助项目（No.62071267）
**Foundation Item:** The National Natural Science Foundation of China (No.62071267)

中，以实现秘密信息的安全通信，保证了传输过程的隐蔽性。

传统的信息隐藏技术大多通过对载体（数字文本、数字图像、音视频）内容进行细微改动，嵌入秘密信息并生成含密载体[1]。其中，文本作为日常生活中使用最频繁的载体之一，已成为信息隐藏的重要载体之一。但是，与其他载体相比，文本的数据量较小，包含的冗余信息量也较少，导致秘密信息嵌入相对困难[2]。同时，针对文本信息隐藏的检测算法[3]也日渐成熟。因此，如何在实现高容量的文本信息隐藏的同时提高秘密信息的抗检测性已成为当前的研究热点。

2014 年 5 月，在全国信息隐藏暨多媒体信息安全大会上，国内的多位学者首次提出“无载体信息隐藏”的概念，开辟了一个全新的、富有挑战性的研究领域。“无载体信息隐藏”并不是不需要载体，而是采用不修改载体或直接生成含密载体的方法传递秘密信息。

Chen 等[4]提出了第一个文本无载体信息隐藏方法——基于汉字数学表达式的文本无载体隐藏方法。该方法将汉字分割为汉字部件，并以此作为隐秘标签，以一个隐秘标签和一个加密关键词的组合为单位从文本库中匹配合适的文本载体实现信息隐藏。Zhou 等[5]在此基础上按顺序对每一次按“隐秘标签+关键词”检索得到的所有文本集合求交集，以提升文本隐藏容量。由于在建立索引时，每篇文本只选取第一次隐秘标签的位置，因此交集不为空的情况很少，对隐藏容量的提升很有限。Chen 等[6]使用汉字 Unicode 编码的奇偶性设计定位标签，保证了每种标签在单篇文本中分布均匀，进而提高关键词的隐藏成功率。Long 等[7]在“标签+关键词”组合检索失败的情况下，利用 Word2Vec 匹配相似的关键词，将原有关键词替换为相似关键词后，加上标签继续进行检索，进一步提高了关键词的隐藏成功率。除上述方法外，还有相关研究者根据汉字结构[8]、汉字笔顺[9]、汉字拼音[10]、汉字声调[11]特征对文本进行编码，以寻求新的突破。但上述方法都只针对标签形式进行探索，在单篇载体文本中不同的标签只能出现一次，载体中依然存在大量冗余信息未被使用。

2017 年后，基于深度学习的生成式无载体信息隐藏方法引起了学者的关注。与上述依赖定位标签隐藏和提取秘密信息的方法相比，生成式方法直接以秘密信息生成含密载体，不需要构建、维护大型的文本数据库。生成式方法的核心是使用语言模型在文本生成的过程中对单词进行编码，以实现信息隐藏。目前，广泛应用于生成式无载体信息隐藏的语言模型包括马尔可夫链[12]、循环神经网络（RNN, recurrent neural network）[13]、长短期记忆网络（LSTM, long short term memory）[14]、生成式对抗网络（GAN, generative adversarial network）[15]、Transformer[16]等。

上述方法中，基于定位标签的搜索式隐藏方法虽然具有很好的隐蔽性，但隐藏容量较低，无法实现高效的秘密信息通信；基于深度学习语言模型的生成式方法可以实现高容量的秘密信息嵌入，但是当秘密信息的长度较大时，生成的文本大概率会出现上下文语义不连贯、语义错误等情况，因此很难抵抗隐写检测。

本文主要的创新点介绍如下。

1) 提出一种标签传递机制，实现一个标签对应多个关键词，充分利用单篇文本载体中的冗余信息。在秘密通信的过程中使用生成式方法传递定位标签和关键词位置参数，再结合搜索式方法用参数从自然文本中提取秘密信息。与传统搜索式方法不同的是，最终传递的载体文本除了多篇自然文本，还包括一篇生成文本。

2) 利用基于知网的词汇相似度计算方法对分词后的秘密关键词进行同义词扩展，选择包含多个关键词或其同义词的自然文本作为载体文本，并定义参数格式，以实现秘密信息的正确提取。

实验结果表明，本文方法融合了搜索式和生成式无载体信息隐藏方法的优点，不仅隐藏容量和隐藏成功率优于文献[5-7]提出的方法，而且在一定程度上解决了秘密信息过长导致生成文本质量下降的问题。

## 2 背景知识

### 2.1 定位标签设计及索引构建

基于 Unicode 编码的定位标签由 Chen 等[6]提出。Unicode 编码实际上是一种字符集，所有的字符都可以用 16 位二进制数来表示。汉字在 Unicode 编码中的范围为 0x4E00~0x9FA5。Chen 等根据汉字 Unicode 编码的奇偶性，将文本转换成 0-1 比特流，如表 1 所示，然后使用 $n$ 位 0-1 比特流作为定位标签。这样设计的好处在于所有的定位标签在文本中的分布很均匀，从而能够保证信息隐藏的成功率。

**表 1　　汉字转 0-1 比特流**

| 汉字 | Unicode | 奇偶性 | 转换后的值 |
|---|---|---|---|
| 我 | \u6211 | 奇 | 1 |
| 爱 | \u7231 | 奇 | 1 |
| 祖 | \u7956 | 偶 | 0 |
| 国 | \u56fd | 奇 | 1 |

在执行秘密信息嵌入的操作时，需要在文本大数据中搜索同时包含“定位标签”和“关键词”的载体文本。为了快速从海量文本中检索出满足条件的所有文本，需要构建倒排索引。在索引结构中应包含 3 个参数，分别是“定位标签”“关键词”“文本路径”。在信息隐藏的检索过程中，首先考虑的应该是文本库中某篇文本是否包含当前的定位标签，然后判断该定位标签后的4个汉字分词后的第一个词是否是要隐藏的关键词。因此，三级倒排索引的结构如图 1 所示。

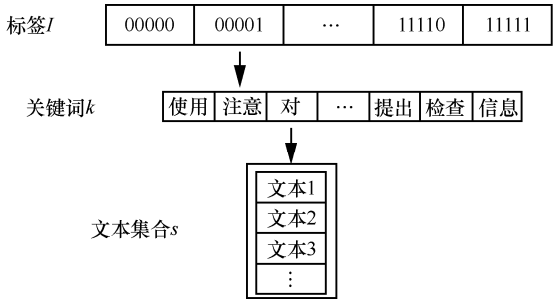



图 1　倒排索引结构

### 2.2　基于知网的词汇语义相似度计算

知网[17]是董振东先生建设的一个大型中英文知识库，目前依然在不断发展中。其有 2 个主要概念：义项（又称概念）和义原。义项是对词汇语义的一种描述，义原是用来描述义项的基本单位。知网中描述了义原的 8 种关系，其中最重要的是上下位关系。根据义原的上下位关系，所有的义原构成了一个义原层次体系。

2002 年，刘群等[17]在基于知网的词汇语义相似度计算的研究中，改进了将词汇相似度转化成义项相似度的方法，根据语义表达式的特点，将计算义项相似度更改为计算语义表达式相似度。通过计算 4 种语义描述式相似度并加权求和，得到了 2 个词汇间的语义相似度计算式为

$$\mathrm{Sim}(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}\mathrm{Sim}_j(S_1,S_2) \qquad (1)$$

其中，$\beta_i(1\leqslant i\leqslant 4)$ 是 4 个调节参数，满足 $\beta_1+\beta_2+\beta_3+\beta_4=1$ 且 $\beta_1\leqslant\beta_2\leqslant\beta_3\leqslant\beta_4$。在文献[17]的方法中，4 个调节参数分别为 $\beta_1=0.5$，$\beta_2=0.2$，$\beta_3=0.17$，$\beta_4=0.13$。

$\mathrm{Sim}_j(1\leqslant j\leqslant 4)$ 表示语义描述式中特定描述之间的相似度，分别是第一独立义原描述式、其他独立义原描述式、关系义原描述式和符号义原描述式。义原描述式是根据其中包含的义原之间的相似度计算得来，义原的相似度采用基于路径的相似度算法，计算式为

$$\mathrm{Sim}(p_1,p_2)=\frac{a}{d+a} \qquad (2)$$

其中，$p_1$、$p_2$ 是义原；$d$ 是 $p_1$、$p_2$ 在义原层次体系中的最短路径长度；$a$ 是一个可调节参数，在文献[17]中，$a=1.6$。

### 2.3　基于 RNN 和 Huffman 编码的文本自动生成隐写

基于文本自动生成的隐写方法不需要构建大型的载体文本库，该类方法直接以秘密信息生成含密文本，适合用来传递长度较短的秘密信息。文献[13]提出了一种基于 RNN 和 Huffman 编码的文本自动生成隐写方法，其流程如图 2 所示。

该方法在执行秘密信息嵌入过程前，首先训练包含多个隐藏层的 RNN 模型来提取文本库中的高维文本特征，最终得到满足训练样本统计特征的语言模型。在嵌入的过程中，根据已经生成的单词和训练好的 RNN 模型，计算下一时刻生成单词的概率分布；对所有候选单词按概率降序排序，选择前 $n(n=2^m,m\in N)$ 个单词构建候选池，对侯选池构建 Huffman 树，并按照左 0 右 1 进行编码，最后根据秘密信息序列，从树的根节点开始搜索，找到叶节点并输出其对应的候选词。将候选词作为下一次的输入，重复上述步骤，直到秘密信息全部被嵌入。

提取操作是嵌入的逆过程。接收方在收到文本后，使用网络结构和参数相同的语言模型计算每个单词在对应时刻的条件概率分布，构造与发送方相同的候选池，同样采用 Huffman 编码对候选池中的单词进行编码。最终根据文本中的每个单词还原出部分秘密序列，拼接得到完整的秘密信息。

## 3　方案设计

为了兼顾文本的隐藏容量和生成文本的质量，本文提出了一种基于同义词扩展和标签传递机制的无载体信息隐藏方法，其流程架构如图 3 所示。

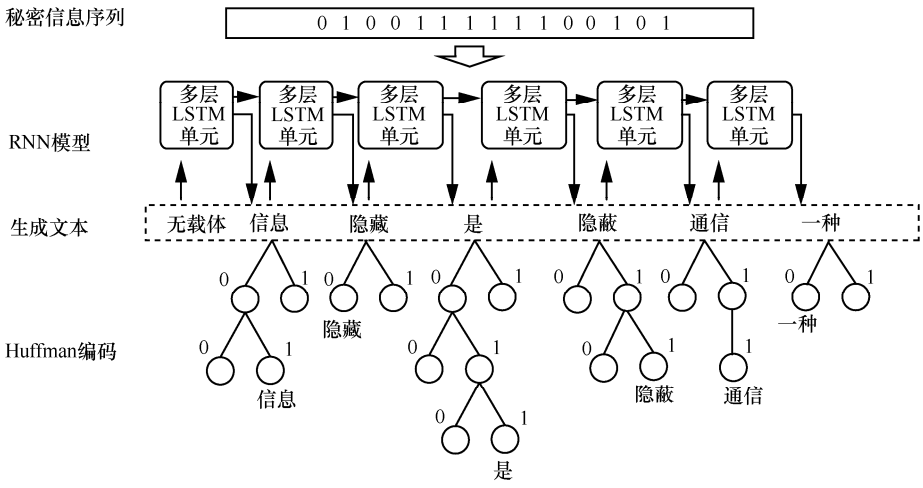



图 2　基于 RNN 和 Huffman 编码的文本自动生成隐写方法流程

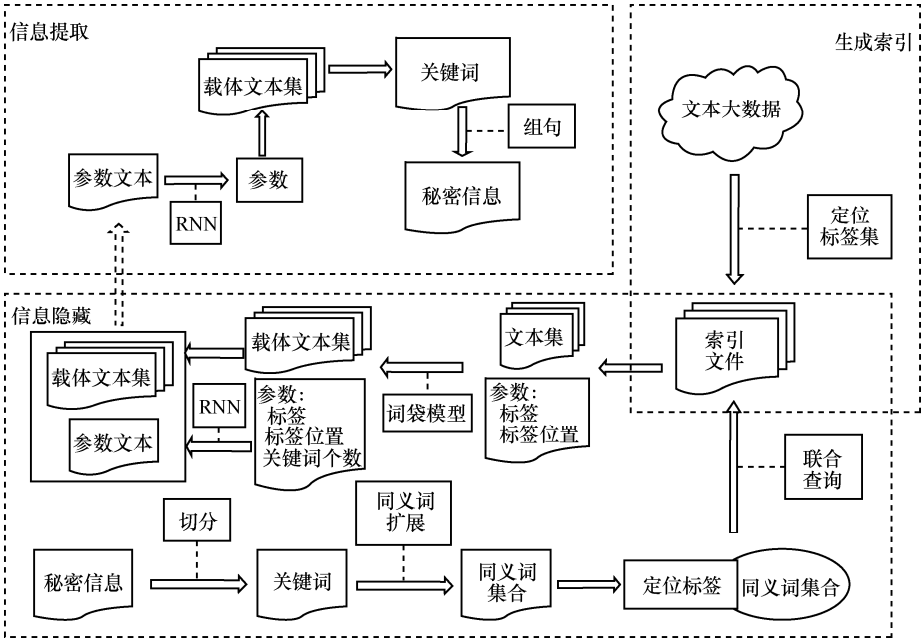



图 3　基于同义词扩展和标签传递机制的无载体信息隐藏流程架构

本文方法在嵌入时包含以下 2 个步骤：1) 使用原始的“定位标签+关键词”的方法将秘密信息隐藏到载体文本中，同时记录下所用的参数；2) 将所有参数转换成固定格式的二进制比特，使用基于 RNN 和 Huffman 编码的文本自动生成方法生成含密载体。步骤 1)中，为了在原有模型的基础上提高隐藏容量，本文方法对秘密信息分词后，首先使用哈尔滨工业大学提供的同义词词林将每个关键词扩充为同义词集合，接着使用刘群等[17]提出的词汇语义相似度方法计算集合中所有同义词与关键词的相似度，选择相似度大于 0.5 的同义词构成最终的同义词扩展集合，如图 4 所示。其目的是将更多的关键词或其同义词隐藏在同一篇载体文本中，以提高隐藏容量。

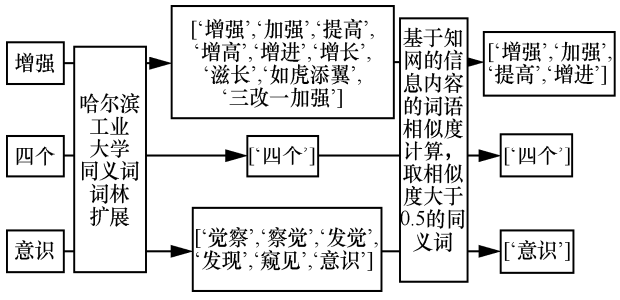



图 4　同义词扩展实例

本文方法的具体步骤如下所述。

### 3.1 构建索引

为了能在嵌入过程中快速找到满足“定位标签+关键词”的文本，本文方法首先需要创建索引文件，创建的具体过程如下所述。

1) 遍历文本库中的每一个文本 $T$，从文本的起始位置 IP 开始，取出 $n$ 个汉字根据其 Unicode 的奇偶性转换成二进制序列作为标签 $L$。

2) 选取定位标签后的 4 个汉字，对其进行分词操作，选取分词结果的第一个词或字作为关键词 $K$。

3) 创建名为 $L$ 的索引文件并存储步骤 2) 得到的关键词 $K$ 和文本 $T$ 的文本路径。

4) 与之前方法不同的是，本文方法允许载体文本中定位标签重复出现，因此只需要按顺序遍历文本，起始位置 IP=IP+1，重复上述步骤，直到 IP+$n$ 等于文本长度。

创建索引的伪代码如算法 1 所示。

**算法 1**　创建索引伪代码

**输入**　文本库中的所有文本，定位标签长度 $n$

**输出**　以各标签命名的哈希表 $\text{HashIndex} = [\text{Index}\,L_1, \text{Index}\,L_2, \cdots, \text{Index}\,L_n]$

1)while 文本库没有遍历结束：

2)　取出一篇文本 $T$，剔除 $T$ 中非汉字符，统计汉字的总数 $C$，将 $T$ 的起始位置 IP 置 0；

3)　if　IP <=$C-n$：

4)　　选取从 IP 开始的 $n$ 个汉字；

5)　　根据 Unicode 编码的奇偶性转换 $n$ 个汉字为二进制序列作为标签 $L$；

6)　　对标签后的 4 个汉字分词；

7)　　取分词后的第一个词作为关键词 $K$；

8)　　if 以 $L$ 命名的哈希表不存在：

9)　　创建一张哈希表并以 $L$ 命名；

10)　　将关键词$K$和文本路径存入以$L$为名的哈希表中；

11)　　end if

12)　　IP = IP + 1；

13)　end if

14)return 多张哈希表

### 3.2 RNN 模型训练

为了使自动生成的文本与载体库中的自然文本尽可能相似，本文使用文献[13]提出的多层 RNN 模型提取文本库中的文本特征。训练集中的每一个句子 $S$ 可表示为 $w_1, w_2, \cdots, w_n$。由于需要将文本转化为神经网络可以学习的语言，首先需要把单词转化为高维词向量，以此来表征词语深层的语义信息。本文方法假定一个句子中包含 $l$ 个词语，词向量维度为 300，句子 $S$ 可表示为

$$S = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_l \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,300} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,300} \\ \vdots & \vdots & \ddots & \vdots \\ x_{l,1} & x_{l,2} & \cdots & x_{l,300} \end{bmatrix} \tag{3}$$

多个隐藏层的 RNN 模型包含多个隐藏层，每一个隐藏层又包含多个 LSTM 单元，因此，第 $j$ 个隐藏层的每一个单元在 $t$ 时刻的输出 $o_{i,t}^{j}$，可以通过式(4)计算。

$$o_{i,t}^{j} = f_{\text{LSTM}}\left(\sum_{k=1}^{300} W_{i,k}^{j} x_{j,k} + b_{i,t}^{j}\right) \tag{4}$$

其中，$W_{i,k}^{j}$ 表示第 $j$ 个隐藏层的权重，$b_{i,t}^{j}$ 表示在 $t$ 时刻第 $j$ 个隐藏层的偏置项。RNN 在 $t$ 时刻每一个隐藏层都可以表示为一个集合 $O_t^j = \left[o_{1,t}^{j}, o_{2,t}^{j}, \cdots, o_{N,t}^{j}\right]$，最终输出值 $y$ 可用式(5)计算。

$$y_i = \sum_{k=1}^{m} w_{k,i}^{p} o_{i,t}^{l} + b_{i,t}^{p} \tag{5}$$

其中，$w_{k,i}^{p}$ 表示预测概率的权值，$o_{i,t}^{l}$ 表示一个特征的重要程度，$b_{i,t}^{p}$ 表示偏置项。接着使用 softmax 函数对输出值进行归一化，计算 $t$ 时刻出现该词的概率 $p_w$，计算式为

$$p_w = \frac{\exp(y_i)}{\sum_{j=1}^{N} \exp(y_i)} \tag{6}$$

其中，$N$ 表示训练集中不重复的单词数。本文模型中采用损失函数反向传播的方式更新参数，损失函数的计算式为

$$\text{Loss} = -\log(p(S)) = -\sum_{t=1}^{L} \log\left(p_{w_t}\right) \tag{7}$$

在经过数轮迭代更新后，当损失函数最小化时，可以得到满足训练样本统计特征的语言模型。

### 3.3 隐藏方法

本文方法主体上采用“定位标签+关键词”的形式完成信息隐藏，标签在隐藏和提取操作中都扮演着不可或缺的角色。与之前研究不同的是，本文方法允许在一篇载体文本中定位标签多次出现，因此在隐藏过程中还需要记录定位标签的位置，即使用“标签+标签位置”来确定一个关键词。同时，为了实现接收

端完整还原秘密信息的要求，每个关键词在原始秘密信息中的位置信息也需要被记录。上述一系列位置信息参数都将转变为固定格式的二进制参数，嵌入生成文本中传递给接收方。具体隐藏方法如下。

1) 对秘密信息 $M$ 进行切分，对于切分后长度大于 4 的关键词继续切分，直到每个关键词长度不超过 4，得到关键词集合 $K=\{k_1,k_2,\cdots,k_n\}$（$n$ 为最终切分完成后的关键词个数）。

2) 使用哈尔滨工业大学提供的同义词词林将每个关键词扩展为同义词集合 $S$，接着使用刘群等[17]提出的词汇语义相似度方法筛选集合中与原关键词相似度大于 0.5 的词汇，构成最终的同义词扩展集合 $S'=\{s_1,s_2,\cdots,s_n\}$（$s_i=\{w_1,w_2,\cdots\}$ 为最终扩展的同义词集合）。

3) 对于 $S'$ 中的每个关键词集合 $s_i$，遍历 $s_i$ 中每个关键词 $w_j$，在所有索引文件中检索关键词 $w_j$，获得所有满足条件的文本集合 $t_j=\{\text{txt}_1,\text{txt}_2,\cdots\}$，并记录索引文件名中的标签 $l_j=\{\text{tag}_1,\text{tag}_2,\cdots\}$，以及定位标签的位置信息构成位置信息集合 $c_j=\{d_1,d_2,\cdots\}$。

4) 当上述过程中出现同一个关键词扩展集合 $s_i$ 中不同关键词在同一篇的文本中存在的情况时，只记录第一次出现的关键词相关信息。最终得到每个 $s_i$ 对应文本集的集合 $T=\{t_1,t_2,\cdots,t_n\}$、对应标签集的集合 $L=\{l_1,l_2,\cdots,l_n\}$、对应关键词位置集的集合 $C=\{c_1,c_2,\cdots,c_n\}$。若在遍历完 $s_i$ 后，文本集的集合 $t_i$ 为空，则将原秘密关键词切分成单个汉字继续隐藏（此时不对单个汉字做同义词扩展）。

5) 对于文本集的集合 $T=\{t_1,t_2,\cdots,t_n\}$，构建词袋模型，选出频率最高的文本 $\text{txt}_i$，记录其中包含的所有关键词在秘密信息中的位置 $m'=\{m'_1,m'_2,\cdots\}$、标签集合 $L'=\{l'_1,l'_2,\cdots\}$ 和标签位置 $d'=\{d'_1,d'_2,\cdots\}$，将 $\text{txt}_i$ 作为载体文本发送给接收方，$m'$、$L'$、$d'$ 按固定格式转化成二进制参数序列 $e$ 存储。

6) 剔除 $T$ 中在步骤 5)已经隐藏的关键词扩展集合，再次执行步骤 5)，直到所有关键词都隐藏完成。

7) 使用 RNN 模型计算候选池中单词的概率分布，使用 Huffman 编码按条件概率对候选词进行编码，根据二进制参数 $e$ 选择合适的候选词作为下一轮输入，直到参数 $e$ 完全被嵌入，最终生成文本 $\text{txt}'$。

为了更直观地了解隐藏过程，本文设计了如图 5 所示的详细流程和如算法 2 所示的伪代码。

**算法 2** 秘密信息隐藏方法

**输入** 定位标签长度 $n$，秘密信息 $M$

**输出** 载体文本集合 $\{\text{txt}_1,\text{txt}_2,\cdots\}$ 和生成的参数文本 $\text{txt}'$

1) 创建索引表 HashIndex；

2) 根据 $n$ 生成 $2^n$ 种标签的集合 LabelSet；

3) 训练 RNN 模型；

4) 对秘密信息 $M$ 分词、去停用词，得到关键词集合 $K$，对于 $K$ 中的每个关键词，若长度大于 4，则继续按长度 2 切分，直到所有关键词长度都不超过 4，得到集合 $K'$；

5) 对 $K'$ 中每个关键词进行扩展，得到同义词

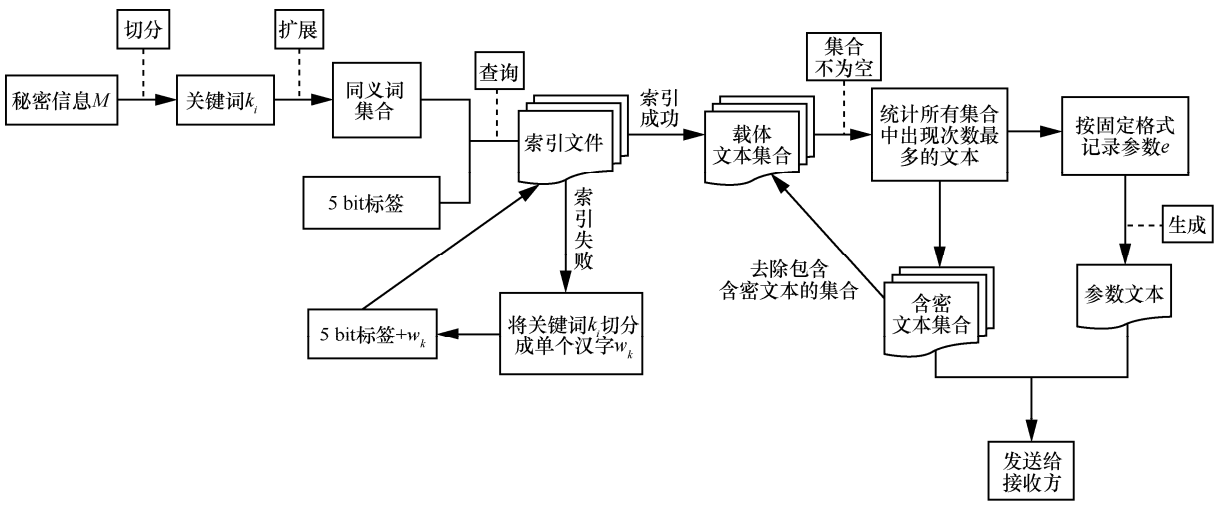



图 5 秘密信息隐藏过程

扩展集合 $S'$；

6) 定义文本集的集合 $T$ ；

7) for $s$ in $S'$：

8)　for synonyms in $s$:

9)　在所有索引表中检索 synonyms，将检索到的文本集合存入集合 $t$；

10)　对 $t$ 集合中的文本进行去重；

11)　if $t$ 为空集：

12)　将 $s$ 集合对应的关键词 $k$ 按单个汉字切分；

13)　for word in $k$：

14)　在所有索引表中检索 word，将检索到的文本集合存入集合 $t$

15)　if 检索结果为空：

16)　print “隐藏失败”

17)　continue；

18)　T.append($t$)；

19)　end if

20)　end for

21)　else

22)　T.append($t$)；

23)　end if

24)　end for

25) end for

26) while $T$ 不为空：

27)对 $T$ 构建词袋模型，取出出现频率最高的文本 txt，记录该文本中所有隐藏的关键词对应的标签集合 $L'$、关键词在秘密信息中的位置集合 $m'$；

28)根据标签集合，在文本中根据关键词扩展集合检索标签位置 $d'$；

29) 将标签集合 $L'$、关键词在秘密信息中的位置集合 $m'$ 和标签位置 $d'$ 按固定格式转化为二进制比特 $e$ 并存储；

30) 将文本 txt 发送给接收方；

31) 在 $T$ 中剔除上述文本中已经隐藏的关键词扩展子集；

32) end while

33) 使用 RNN 模型计算候选池中单词的概率分布，使用 Huffman 编码按条件概率对候选词进行编码，根据二进制参数 $e$ 选择合适的候选词作为下一轮输入，直到参数 $e$ 完全被嵌入，最终生成文本 txt′。

上述二进制参数格式如图 6 所示。

分词数：根据秘密信息被切分成关键词的个数 $n_{\text{kws}}$，计算分词数 $a$ 的值满足 $2^{a-1} \leqslant n_{\text{kws}} \leqslant 2^{a}$，用固定 6 bit 记录分词数 $a$ 的值。

最大隐藏数：选择隐藏最多的文本，记录隐藏的关键词个数 $\max_{\text{kws}}$，计算最大隐藏数 $c$ 的值满足 $2^{c-1} \leqslant \max_{\text{kws}} \leqslant 2^{c}$，用固定 5 bit 记录最大隐藏数 $c$ 的值。

$k_i$ 表示第 $i$ 篇文本中隐藏的关键词个数，每个关键词需要“标签、标签位置、秘密信息位置”3 个参数才能解密，因此每一篇文本载体，需要 $c+k_i(5+6+a)$ 长度的参数来提取其中的关键词，而 $a$ 和 $c$ 作为 2 个变量，分别用固定的 6 bit 和 5 bit 保存。

关键词个数：第 $i$ 篇文本中关键词的个数，用 $c$ bit 表示。

标签：定位标签，用 5 bit 表示。

标签位置：与标签搭配使用，表示在该文本中取某标签下的第几个关键词，用 6 bit 表示。

秘密信息位置：对应关键词在原秘密信息中的位置，用 $a$ bit 表示。

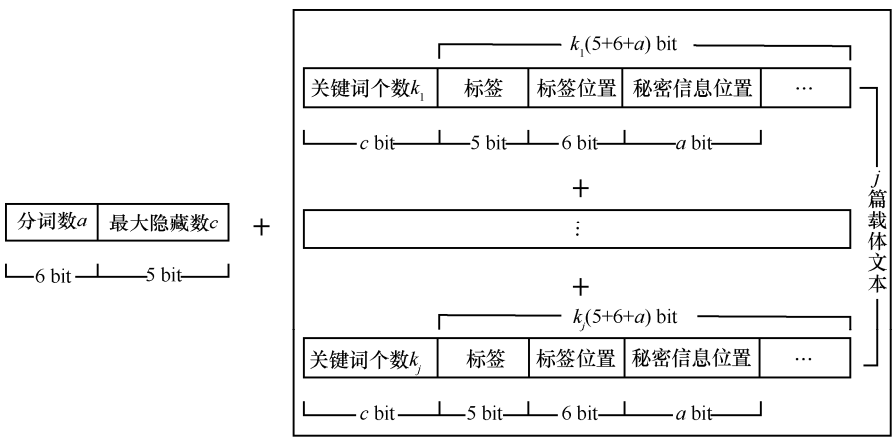



图 6　二进制参数格式

### 3.4 提取方法

秘密信息提取过程是隐藏的逆过程，接收方通过维护和发送方同样的 RNN 语言模型，从收到的生成文本中解出二进制序列，按照固定格式解析得到每篇文本中所有关键词的标签集合 $L'$、关键词在秘密信息中的位置集合 $m'$ 和标签位置 $d'$，接着使用 $L'+d'$ 从载体文本中解出关键词，使用 $m'$ 将关键词还原成原始秘密信息，具体提取过程如下所述。

1) 接收方维护与发送方相同的 RNN 模型。对于收到的最后一个文本 txt′，使用 RNN 模型计算每个单词在每个时刻的概率分布，根据计算出来的条件概率使用 Huffman 编码方法对文本中的词语进行编码，解出二进制参数 $e$。

2) 对于参数 $e$，按图 5 所示的格式解析，得到各篇文本中包含的关键词个数，以及每个关键词对应的标签 tag、标签位置 $d$ 和秘密信息位置 $m$。

3) 根据标签 tag 和标签位置 $d$ 在对应的文本中提取关键词，最后根据秘密信息位置 $m$ 将关键词进行排序组句，得到最终完整秘密信息。

为了更直观地了解提取过程，本文设计了如图 7 所示的详细流程。提取过程的伪代码如算法 3 所示。

**算法 3** 秘密信息提取方法

**输入** 生成的参数文本 txt′，载体文本集合 $\{\text{txt}_1,\text{txt}_2,\cdots\}$

**输出** 还原后的秘密信息 $M'$

1) 使用与发送方相同参数训练 RNN 模型；

2) 使用 RNN 模型计算 txt′ 中每个单词在每个时刻的概率分布，根据计算出来的条件概率使用 Huffman 编码方法对文本中的词语进行编码，解出二进制参数 $e$；

3) 按照固定格式解析 $e$，得到每篇文本 txt 中对应的标签集合 $L'$、标签位置 $d'$ 和关键词在秘密信息中的位置集合 $m'$；

4) 根据 $L'+d'$ 从载体文本 txt 中解出关键词，根据 $m'$ 将关键词还原成秘密信息 $M'$；

5) return $M'$。

## 4 实验分析

实验中构建索引所用的文本数据库大小为 422 MB，包含 216 160 篇文本。这些文本分别来自不同的领域，包括古典文学、简书文章、新闻、小说，按照 1:1:1:1 的比例从网络中随机抽取。为了保证实验结果的可靠公正，测试数据集从搜狗实验室中随机选取。

### 4.1 秘密信息相似度分析

本文方法在隐藏过程中使用同义词替换原始关键词以提升隐藏容量，因此最终还原出的秘密信息有可能与原始秘密信息存在一定的差异。为了评估该差异对原始秘密信息语义表达的影响，本文首先进行嵌入信息与原始秘密信息的相似度计算。

本文实验使用逆文本频率（idf）和余弦相似度（cosine）来计算一个秘密信息的相似度。句子的含义通常都由其中的核心词决定，因此在计算句子相似度时，核心词之间的相似度将占据很大的比重。idf 是衡量一个词语普遍重要性的参数，这样可以将每个词的 idf 值视为权重来计算句向量。计算单个词汇的 idf 值和句向量的计算式为

$$\text{idf}(w)=\log\left(\frac{D}{D_w+1}\right) \tag{8}$$

其中，$D$ 为语料库中文档的数量，$D_w$ 为出现词汇 $w$ 的文档的数量。

$$\text{vector}(s)=\sum_i^m v(w_i)\text{idf}(w_i) \tag{9}$$

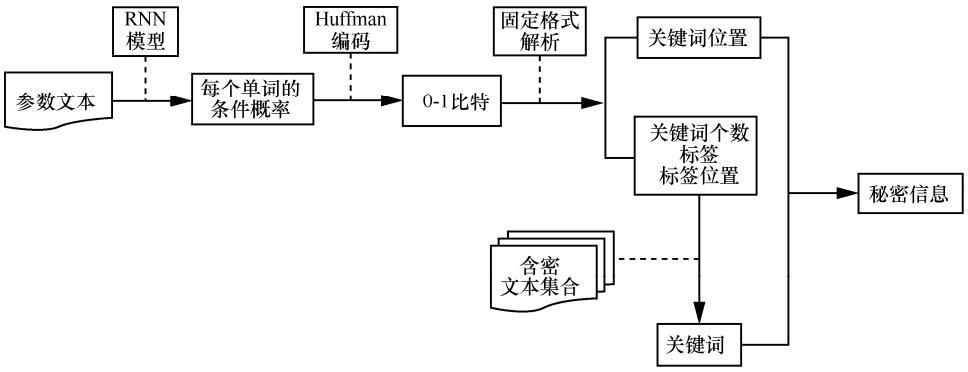



图 7 秘密信息提取过程

其中，$v(w_i)$为句子第$i$个词$w_i$的向量，$\mathrm{idf}(w_i)$为词$w_i$的 idf 值。2 个句向量之间的相似度可以用余弦相似度来衡量。余弦相似度通过计算 2 个向量夹角的余弦值来度量它们之间的相似性。假设$\boldsymbol{v}_1$和$\boldsymbol{v}_2$是 2 个 $n$ 维向量，即 $\boldsymbol{v}_1=(x_1,x_2,\cdots,x_n)$, $\boldsymbol{v}_2=(y_1,y_2,\cdots,y_n)$，它们夹角的余弦值的计算式为

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{10}$$

当 2 个句子相似度越高时，它们的向量余弦夹角越小，其余弦值越接近 1；当余弦值为负数时，表示 2 个向量负相关。

本文实验从搜狗实验室的新闻语料库中随机挑选了 4 组文本，第一组包含 50 篇 1 KB 文本，第二组包含 50 篇 2 KB 文本，第三组包含 100 篇 1 KB 文本，第四组包含 100 篇 2 KB 文本。以每一篇文本作为秘密信息，计算每一组中的嵌入信息与原秘密信息的平均相似度$\beta$，其计算式为

$$\beta=\frac{\sum_{i}^{n}\cos(\theta)_i}{n} \tag{11}$$

其中，$\cos(\theta)_i$表示每组中第$i$篇测试文本嵌入信息与原秘密信息的相似度，$n$表示该组文本的数量。实验结果如图 8 所示。

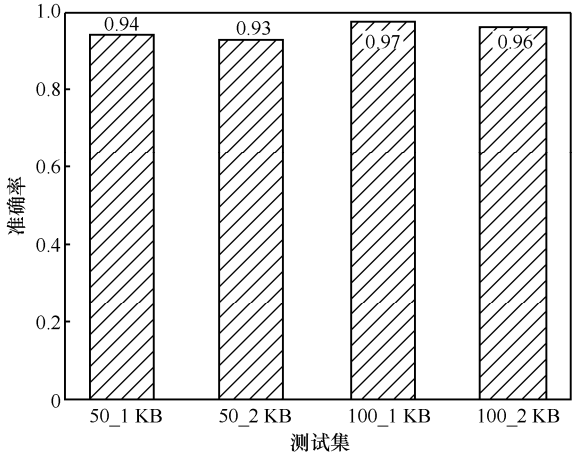



图 8　嵌入信息与原始信息相似度对比

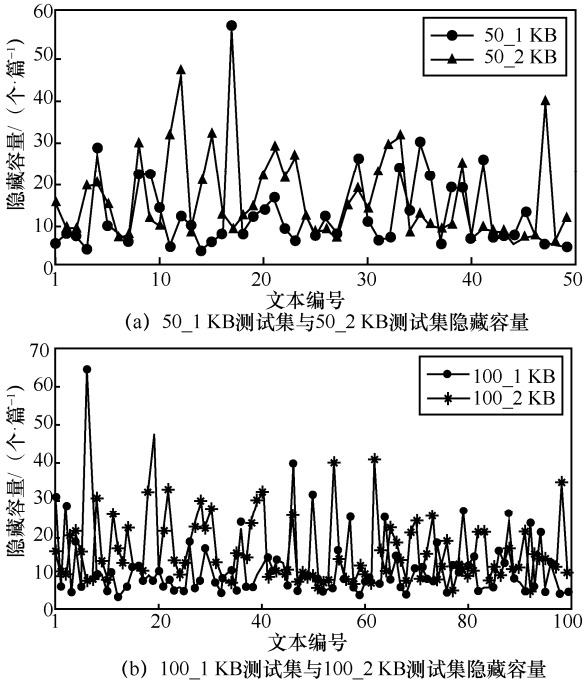
上述实验结果表明，经过同义词替换后的秘密信息与原始的秘密信息在语义上依然保持较高的相似度。因此，同义词替换操作对于原始秘密信息语义的表达几乎没有影响。

### 4.2　隐藏容量

本文提出的方法需要发送多个载体文本，因此对隐藏容量的定义为被隐藏的关键词总数和最终传递的文本数量（包含一个生成文本）的比值。计算式为

$$y=\frac{\tau}{\rho} \tag{12}$$

其中，$\tau$为秘密信息传递过程中被隐藏的关键词总数，$\rho$为最终传递的文本数量。实验使用固定标签长度为 5，测试集依然使用上述 4 组文本集合，图 9 是每一组文本的隐藏容量结果。




(a) 50_1 KB测试集与50_2 KB测试集隐藏容量

(b) 100_1 KB测试集与100_2 KB测试集隐藏容量

图 9　不同测试集的隐藏容量

根据图 9 可以发现，对于不同的秘密信息，隐藏容量的波动幅度很大，主要原因是本文实验构建的文本数据库较小，虽然可以满足正常的隐藏和提取功能需求，但由于文本库中的文本与不同的秘密信息的内容相关度不同，导致某一种类的秘密信息在嵌入过程中达到非常高的隐藏容量。同时，秘密信息的长度也影响了隐藏容量。可以看到，实验中 2 KB 文本测试集中绝大部分文本的隐藏容量都大于 1 KB 文本测试集中文本的隐藏容量，由此推测出当秘密信息长度越长，单篇文本的隐藏容量也会越高。

即使本文方法在嵌入容量上有较大波动，但与其他传统的搜索式算法相比，本文方法在嵌入容量上有了较大的提升。本文实验的平均隐藏容量$\tilde{y}$计算式为

$$\tilde{\gamma}=\frac{\sum_{i}^{N}\gamma_i}{N} \tag{13}$$

其中，$N$ 为每一组测试文本的数量，$\gamma_i$ 为第 $i$ 个测试文本的隐藏容量。表 2 为本文方法与其他方法在平均隐藏容量上的对比情况。

**表 2　　各方法隐藏容量对比**

| 方法 | 50_1 KB | 50_2 KB | 100_1 KB | 100_2 KB |
|---|---|---|---|---|
| 本文方法 | **12.941** | **13.833** | **9.831** | **13.500** |
| 多关键词[5] | 1.005 | 1.000 | 1.000 | 1.000 |
| Unicode[6] | 1.086 | 1.063 | 1.081 | 1.081 |
| Word2Vec[7] | 1.021 | 1.016 | 1.013 | 1.013 |
| 词编码[18] | 1.054 | 1.065 | 1.063 | 1.055 |

根据表 2 可以看到，在使用相同文本数据库的前提下，由于加入了同义词扩展以及不受位置约束的载体文本合并操作，本文方法的平均隐藏容量远大于对比方法。

### 4.3　隐藏成功率

在信息隐藏的过程中，难免会存在不常用的词汇、汉字导致该关键词隐藏失败，因此本文实验定义隐藏成功率为 $\sigma$，当一篇测试文本中所有汉字都被成功隐藏才算隐藏成功，成功率计算式为

$$\sigma=\frac{x}{\chi}$$

其中，$x$ 为一组文本中成功隐藏的文本数量，$\chi$ 为一组文本中所有测试文本的数量（本文中 $\chi=50,100$）。表 3 为本文方法与其他方法在隐藏成功率上的对比情况。

**表 3　　各方法隐藏成功率对比**

| 方法 | 50_1 KB | 50_2 KB | 100_1 KB | 100_2 KB |
|---|---|---|---|---|
| 本文方法 | **1.00** | **0.92** | **0.95** | **0.94** |
| 多关键词[5] | 0 | 0 | 0 | 0 |
| Unicode[6] | 0.50 | 0.08 | 0.35 | 0.15 |
| Word2Vec[7] | 0.24 | 0.10 | 0.34 | 0.23 |
| 词编码[18] | 0.20 | 0.15 | 0.28 | 0.18 |

与其他方法相比，本文方法对文本库的要求较低，在同样使用小型文本库的情况下，本文方法依然可以较好地实现秘密信息的嵌入和提取。但其他方法由于标签形式的限制，不能充分利用载体文本中的冗余信息，必须构建并维护大型文本数据库才能保证较高的隐藏成功率，因此通信所需的开销也较昂贵。

### 4.4　隐蔽性

基于文本的隐写分析通常是使用基于统计的算法检测文本的修改痕迹，但本文方法并未对文本本身进行修改，使用自然文本实现秘密信息的传递，因此在通信过程中拥有较高的隐蔽性，主要可以体现在如下 2 个方面。

1) 抗检测。本文方法选取 5 位 0-1 比特流作为定位标签，在信息隐藏过程中使用检索的方法挑选自然文本发送给接收方，最终再使用相关参数生成与文本库相似的生成文本发送给接收方，整个过程中没有对载体文本执行任何修改操作，与普通文本并没有差异，因此机器或人眼很难察觉秘密信息的存在。

2) 抗非法提取。要从接收到的多篇文本中准确提取秘密信息，接收方首先要维护与发送方相同的 RNN 语言模型来解码生成文本，同时还需要了解参数格式才能解析关键词的各项参数，最终完成提取操作。

## 5　结束语

与之前基于搜索的无载体隐藏方法相比，本文提出的无载体隐藏方法具有较大的优势。借助文本自动生成的方法，将标签信息嵌入新的生成文本中，在保证隐蔽性的同时也增加了可传递参数的数量，进而有效地提升了自然文本载体的隐藏容量和信息传递的安全性。本文方法提出了搜索式无载体信息隐藏和生成式无载体信息隐藏方法相结合的思想，虽然隐藏容量有了很大的提升，但文本库的差异会导致文本嵌入率有较大的波动。另一方面，本文方法选取的标签形式在进行复用的情况下抵抗针对文本的增加、删除、修改攻击能力较弱。因此，在未来的研究中主要集中在改进嵌入算法，优化同义词匹配算法，探索合适的标签形式，从而降低对文本库之间的差异对隐藏结果的影响，在保证稳健性的前提下提高隐藏容量的稳定性。

## 参考文献：

[1] 张新鹏, 钱振兴, 李晟. 信息隐藏研究展望[J]. 应用科学学报, 2016, 34(5): 475-489.
ZHANG X P, QIAN Z X, LI S. Prospect of digital steganography research[J]. Journal of Applied Sciences, 2016, 34(5): 475-489.

[2] 吴国华, 龚礼春, 袁理锋, 等. 中文文本信息隐藏研究进展[J]. 通信学报, 2019, 40(9): 145-156.

WU G H, GONG L C, YUAN L F, et al. Review of information hiding on Chinese text[J]. Journal on Communications, 2019, 40(9): 145-156.

[3] BAO Y J, YANG H, YANG Z L, et al. Text steganalysis with attentional L STM-CNN[C]//2020 5th International Conference on Computer and Communication Systems. Piscataway: IEEE Press, 2020: 138-142.

[4] CHEN X Y, SUN H Y, TOBE Y, et al. Coverless information hiding method based on the Chinese mathematical expression[C]//Cloud Computing and Security. Piscataway: IEEE Press, 2015: 1-12.

[5] ZHOU Z L, MU Y, YANG C N, et al. Coverless multi-keywords information hiding method based on text[J]. International Journal of Security and Its Applications, 2016, 10(9): 309-320.

[6] CHEN X, CHEN S, WU Y. Coverless information hiding method based on the Chinese character encoding[J]. Journal of Internet Technology, 2017, 18(2): 313-320.

[7] LONG Y, LIU Y L. Text coverless information hiding based on Word2Vec[C]//Cloud Computing and Security. Berlin: Springer, 2018: 463-472.

[8] 彭博, 李晖. 融合多语言特点的无载体信息隐藏[J]. 微处理机, 2019, 40(1): 49-54.
PENG B, LI H. Coverless information hiding by fusing multilingual features[J]. Microprocessors, 2019, 40(1): 49-54.

[9] 王建业, 郭振波, 王开西. 基于汉字数学表达式的无载体文本隐写方法[J]. 青岛大学学报(自然科学版), 2019, 32(1): 81-86.
WANG J Y, GUO Z B, WANG K X. A method of text steganography without carrier based on mathematical expressions of Chinese characters[J]. Journal of Qingdao University (Natural Science Edition), 2019, 32(1): 81-86.

[10] LIU Y L, WU J, XIN G J. Multi-keywords carrier-free text steganography method based on Chinese Pinyin[J]. International Journal of Computational Science and Engineering, 2020, 21(2): 202-209.

[11] 张维, 王晓梅, 张晨旭. 一种基于声调特征映射的无载体信息隐藏方法[J]. 信息工程大学学报, 2021, 22(1): 38-43.
ZHANG W, WANG X M, ZHANG C X. Coverless text steganography method based on tone feature mapping[J]. Journal of Information Engineering University, 2021, 22(1): 38-43.

[12] SHNIPEROV A N, NIKITINA K A. A text steganography method based on Markov chains[J]. Automatic Control and Computer Sciences, 2016, 50(8): 802-808.

[13] YANG Z L, GUO X Q, CHEN Z M, et al. RNN-stega: linguistic steganography based on recurrent neural networks[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1280-1295.

[14] KANG H X, WU H Z, ZHANG X P. Generative text steganography based on LSTM network and attention mechanism with keywords[J]. Electronic Imaging, 2020(4): 291-1-291-8.

[15] YANG Z L, WEI N, LIU Q H, et al. GAN-TStega: text steganography based on generative adversarial networks[C]//Digital Forensics and Watermarking. Berlin: Springer, 2020: 18-31.

[16] ABDELNABI S, FRITZ M. Adversarial watermarking transformer: towards tracing text provenance with data hiding[J]. arXiv Preprint, arXiv: 2009.03015, 2020.

[17] 刘群, 李素建. 基于知网的词汇语义相似度计算[D]. 北京: 中国科学院计算技术研究所, 2002.
LIU Q, LI S J. Calculation of word semantic similarity based on HowNet [D]. Beijing: Institute of Computing Technology, Chinese Academy of Sciences, 2002.

[18] CHEN X Y, CHEN S. Text coverless information hiding based on compound and selection of words[J]. Soft Computing, 2019, 23(15): 6323-6330.

[作者简介]

**张祯**（1978− ），男，山西大同人，杭州电子科技大学副教授，主要研究方向为计算机应用、保密信息化、图形图像处理。

**倪嘉铭**（1995− ），男，江苏南通人，杭州电子科技大学硕士生，主要研究方向为信息内容安全、文本信息隐藏。

**姚晔**（1978− ），男，湖北随州人，博士，杭州电子科技大学副教授，主要研究方向为多媒体内容安全、视频图像智能分析。

**龚礼春**（1995− ），男，福建南平人，杭州电子科技大学硕士生，主要研究方向为信息内容安全、文本信息隐藏。

**王玉娟**（1972− ），女，浙江宁波人，杭州电子科技大学馆员，主要研究方向为信息化与标准化研究。

**吴国华**（1970− ），男，山东济南人，博士，杭州电子科技大学教授、博士生导师，主要研究方向为保密信息化、定密理论与实务。