(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104504342 A
(43)申请公布日 2015.04.08

(21)申请号 201410733815.0

(22)申请日 2014.12.04

(71)申请人 中国科学院信息工程研究所
地址 100093 北京市海淀区闵庄路甲89号

(72)发明人 吴槟 易小伟 赵险峰 冯凯
何晓磊

(74)专利代理机构 北京君尚知识产权代理事务所(普通合伙) 11200
代理人 余长江

(51)Int.Cl.
G06F 21/62(2013.01)

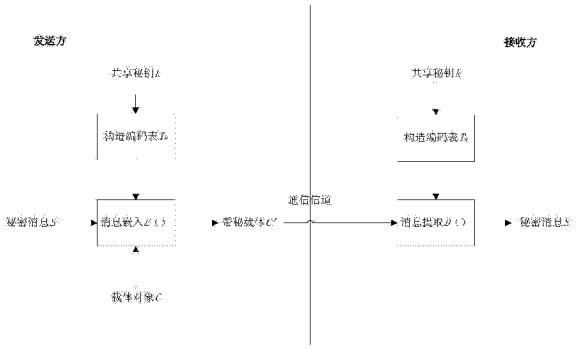
权利要求书2页 说明书6页 附图3页

(54)发明名称

基于Unicode编码利用不可见字符隐藏信息的方法

(57)摘要

本发明提出了一种基于Unicode编码的利用不可见字符隐藏信息的方法,主要包括消息嵌入算法和消息提取算法。通过利用不可见字符的Unicode编码特点来隐藏秘密信息,在保证安全性的前提下提高秘密信息的嵌入容量。利用本发明,能够在采用Unicode编码的含有空格等不可见字符的文本载体上隐藏秘密信息;能够根据编码表灵活地改变秘密信息的嵌入形式来保证信息的安全;能够有效提高秘密信息的嵌入容量。



CN 104504342 A

1. 一种基于 Unicode 编码的利用不可见字符隐藏信息的方法，包括以下步骤：

(1) 发送发和接收方协定秘钥，并利用 Unicode 码中的不可见字符分别根据秘钥构造编码表；

(2) 发送方选择基于 Unicode 编码的文本数据作为载体对象，根据步骤 (1) 中生成的编码表对载体对象中的不可见字符进行重新编码，将秘密消息嵌入到载体对象中，得到带秘载体；

(3) 发送方将步骤 (2) 中得到的带秘载体通过通信信道传输到接收方；

(4) 接收方根据步骤 (1) 中生成的编码表，从步骤 (3) 中接收到的带秘载体中将秘密消息提取出来，得到秘密消息。

2. 如权利要求 1 所述的方法，其特征在于，步骤 (1) 通过发送方和接收方共享的秘钥 k 控制编码表 T_k 的生成，构造编码表的具体实现步骤为：

(1-1) 密钥 k 必须是 2048 比特的二进制字符串，以十进制的方式表示成：

$k = (n_0, n_1, \cdots, n_i, \cdots, n_{255}), n_i \in N \cap [0,255], i = 0,1,\cdots 255,$

此外，密钥 k 的每个分量 n_i 满足如下条件：

$$\forall i,j \in \{0,1,\ldots,255\}, i \neq j \Rightarrow n_i \neq n_j,$$

以上描述表明，密钥 k 可以表述为 0 到 255 整数序列的一个置换，如下式所示：

$k = perms(0,1,2,\cdots,255)$

其中 perms() 为置换函数；

(1-2) 根据 Unicode 码的特征，发送方和接收方事先查找出 256 个不可见字符的可选码字，作为双方共享的编码表的构造基础；

(1-3) 根据步骤 (1-1) 和步骤 (1-2)，建立从密钥到可选码字的编码表 T_k。

3. 如权利要求 1 或 2 所述的方法，其特征在于：步骤 (2) 通过执行嵌入算法 E() 将秘密消息嵌入到载体对象中，嵌入算法 E() 表示为

$c' = E(c,m,T_k)$,

其中，输入参数为载体对象 c、待嵌入的秘密消息 m 和嵌入秘钥 k 生成的编码表 T_k，输出是包含秘密消息的带秘载体 c′。

4. 如权利要求 3 所述的装置，其特征在于：载体对象 c 为基于 Unicode 编码的文本数据，包括 TXT 格式文本和 Word、PDF、XML、HTML 复合文档中的文本数据；秘密消息 m 作为字节流方式进行处理，m 是 TXT 文本数据或者 JPEG 图像数据。

5. 如权利要求 3 所述的方法，其特征在于，所述嵌入算法的具体实现步骤为：

(2-1) 对输入的载体对象 c 进行预处理，用 0x2000 替换 c 中出现在编码表 T_k 的码字；

(2-2) 顺序地读取 c 的 2 个字节数据 xx，并判断 xx 值是否为 0x0000；

(2-3) 如果步骤 (2-2) 中结果为 no，则执行步骤 (2-4)，否则执行步骤 (2-12a)；

(2-4) 判断 xx 是否为不可见字符，包括半角 / 全角空格和制表符，它们对应的 Unicode 码表示分别为 0x2000、0x0030 和 0x0900；

(2-5) 如果步骤 (2-4) 中结果为 no，则跳转到步骤 (2-2)，否则执行步骤 (2-6)；

(2-6) 对输入的秘密消息 s，读取 1 个字节数据 y；

(2-7) 判断 y 值是否为 EOF；

(2-8) 如果步骤 (2-7) 中结果为 no,则执行步骤 (2-9),否则执行步骤 (2-12b) ;

(2-9) 对输入的编码表 T_k,查找 T_k 中 y 值对应的码字 zz ;

(2-10) 使用步骤 (2-9) 中的 zz 值替换步骤 (5) 所得到 c 中的 xx 值 ;

(2-11) 重复执行步骤 (2-2) ;

(2-12a) 输出载体嵌入容量不足的提示信息。

(2-12b) 输出改变后的 c,即为带秘载体 c′ 。

6. 如权利要求 5 所述的方法,其特征在于 :步骤 (4) 通过执行提取算法 D() 从带秘载体中将秘密消息提取出来,提取算法 D() 表示为

$m = D(c', T_{k'})$,

其中,输入参数为含秘密消息的带秘载体 c′ 和提取秘钥 k′ 生成的编码表 $T_{k'}$,输出是提取得到的秘密消息 m ;为了保持编码表 T_k 和 $T_{k'}$ 的一致性,嵌入秘钥 k 与提取密钥 k′ 相同。

7. 如权利要求 6 所述的方法,其特征在于,所述提取算法的具体实现步骤为 :

(4-1) 对输入的带秘载体 c′ ,读取 2 个字节数据 yy ;

(4-2) 判断 yy 值是否为 EOF ;

(4-3) 如果步骤 (4-2) 中结果为 no,则执行步骤 (4-4),否则执行步骤 (4-8) ;

(4-4) 根据输入编码表 $T_{k'}$,判断步骤 (4-2) 中的 yy 值是否为 $T_{k'}$ 中的码字 ;

(4-5) 如果步骤 (4-4) 中结果为 no,则跳转到步骤 (4-1),否则执行步骤 (4-6) ;

(4-6) 查找 $T_{k'}$,获得 yy 值所对应的 m 值 ;

(4-7) 重复执行步骤 (4-1) ;

(4-8) 输出秘密消息 m。

基于 Unicode 编码利用不可见字符隐藏信息的方法

技术领域

[0001] 本发明涉及一种基于 Unicode 编码的利用不可见字符隐藏信息的方法，属于信息隐藏技术领域。

背景技术

[0002] 随着计算机应用的普及和互联网的迅猛发展，人们对网络通信安全提出了更高的要求。与信息加密技术不同，信息隐藏技术通过掩盖传输消息行为的存在性来保障通信安全性。当前，随着文本文档的应用越来越广泛，利用文本文档隐藏秘密消息成为信息安全领域的一个研究方向。在计算机科学领域，Unicode 编码是一种通用字符集编码标准，被广泛地应用于现代操作系统。几乎所有的文字处理软件都支持对文本数据进行 Unicode 编码、解析和存储。

[0003] 不可见字符指文档打印时不可见的字符集合，包括空格、制表符和换行符等。通常地，利用不可见字符来隐藏秘密消息具有不易察觉、操作简单和隐蔽性强等优势。WbStego4 软件是一款利用不可见字符进行隐藏信息的开源工具，它支持 TXT、HTML、PDF 等多种数据格式。但是它不支持中文文本数据载体的信息隐藏，并且信息嵌入容量较为有限，在实际应用中存在很大的局限性。

发明内容

[0004] 本发明所要解决的技术问题是克服现有 WbStego4 软件技术的不足，提供一种基于 Unicode 编码的利用不可见字符隐藏信息的方法，适用于基于 Unicode 编码的含有不可见字符的中英文文本载体来隐藏消息，在保证安全性前提下提升隐藏信息的容量，能够较好地满足隐蔽通信的要求。

[0005] 本发明的技术解决方案是一种基于 Unicode 编码的利用不可见字符隐藏信息的方法，它主要包括消息嵌入算法和提取算法两个部分 ：

[0006] 消息嵌入算法，根据编码表对载体对象中的不可见字符进行重新编码，将秘密消息嵌入到载体对象中，其中载体对象是基于 Unicode 编码的中英文文本数据，不可见字符包括半角 / 全角空格和制表符 (对应的 Unicode 码表示分别为 0x20 00、0x00 30 和 0x0900)。嵌入算法的输入数据有载体对象、秘密消息和编码表，输出数据是带秘载体。

[0007] 消息提取算法，是嵌入算法的逆向算法，它根据编码表从带秘载体中恢复出秘密消息。提取算法的输入数据有带秘载体和编码表，输出数据是秘密消息。

[0008] 本发明的基于 Unicode 编码的利用不可见字符隐藏信息的方法，通过利用不可见字符的编码特点在文本载体中隐藏秘密消息，本方法包括以下步骤 ：

[0009] (1) 发送发和接收方协定秘钥，并分别根据秘钥构造编码表 ；

[0010] (2) 发送方选择载体对象，根据步骤 (1) 中生成的编码表，通过信息嵌入将秘密消息嵌入到载体对象中，得到带秘载体 ；

[0011] (3) 发送方将步骤 (2) 中得到的带秘载体通过通信信道传输到接收方 ；

[0012] (4) 接收方根据步骤 (1) 中生成的编码表，通过提取算法从步骤 (3) 中接收到的带秘载体中将秘密消息提取出来，得到秘密消息。

[0013] 本发明与现有技术相比的有益效果在于：

[0014] (1) 本发明中，载体对象的选择更宽泛。适用于隐藏秘密消息的载体对象可以选择基于 Unicode 编码的文本数据，以满足实际应用的需求。

[0015] (2) 本发明中，通过利用秘钥生成器所生成的秘钥来保证算法的安全性。秘钥可控制编码表的生成，在保证信息隐藏算法安全的前提下，将消息嵌入 / 提取操作与消息加密 / 解密操作联合起来，降低了实际应用中操作的复杂性和能量消耗。

[0016] (3) 本发明中，通过利用编码表以提升嵌入信息的容量。在编制码表时，利用 Unicode 码中冗余的不可见字符码字设计了由 256 个一一映射关系组成的编码表，其中编码表的映射关系受到秘钥的控制。编码表将每次嵌入信息率从 1 比特提升到每次可嵌入 8 比特。

附图说明

[0017] 图 1 是本发明方法实施例的实现流程图；

[0018] 图 2 是本发明方法中消息嵌入算法的实施流程图；

[0019] 图 3 是本发明方法中消息提取算法的实施流程图。

具体实施方式

[0020] 为使本发明的上述目的、特征和优点能够更加明显易懂，下面通过具体实施例和附图，对本发明做进一步说明。

[0021] 如图 1 所示，是本发明的实现流程示意图，该信息隐藏方法可以表示为一个六元组，即 $\Sigma = \langle C, S, T_k, C', E_k, D_k \rangle$，其中 C 为载体对象集合、S 为秘密消息集合、$T_k$ 为根据秘钥 k 构造的编码表、C′ 为载体对象隐藏秘密消息后所得到的带秘载体集合、E_k 为消息嵌入算法、D_k 为消息提取算法。在信息隐藏方法 Σ 中，包括 2 个主要的算法模块：消息嵌入算法模块和消息提取算法模块，分别由发送方和接收方调用。各模块的功能描述如下：

[0022] 1、消息嵌入算法

[0023] 发送方根据秘钥 k 构造编码表 T_k，将发送方输入的秘密信息 S 嵌入到所选择的载体对象 C 中，输出带秘载体 C′。然后，发送方将该带秘载体 C′ 通过通信信道传输到接收方。执行消息嵌入算法 E_k 的过程可以表示为：

[0024] $E_k: C \times M \times T_k \rightarrow C'$

[0025] 2、消息提取算法

[0026] 接收方根据提取秘钥 k 构造编码表 T_k，将从发送方处得到的带秘载体 C′ 通过提取算法提取出秘密信息 M。执行消息提取算法 D_k 的过程可以表示为：

[0027] $D_k: C' \times T_k \rightarrow M$

[0028] 如图 1 所示，本发明的具体实现过程如下：

[0029] 1、编码表构造

[0030] 本发明通过发送方和接收方共享的秘钥 k 控制编码表 T_k 的生成，在保证编码安全性的前提下提高隐藏信息的容量，以满足实际应用的要求。构造编码表的具体实现步骤

为：

[0031] (1) 密钥 k 必须是 2048 比特的二进制字符串，以十进制的方式可以表示成

[0032] $k = (n_0, n_1, \cdots, n_i, \cdots, n_{255}), n_i \in N \cap [0, 255], i = 0, 1, \cdots 255$

[0033] 此外，密钥 k 的每个分量 n_i 还必须满足如下条件

[0034] $\forall i, j \in \{0,1,\ldots,255\}, i \neq j \Rightarrow n_i \neq n_j$

[0035] 以上描述表明：密钥 k 可以表述为 0 到 255 整数序列的一个置换，如下式所示

[0036] $k = perms(0, 1, 2, \cdots, 255)$

[0037] 其中 perms() 为置换函数。

[0038] (2) 根据 Unicode 码的特征，发送方和接收方事先查找出 256 个不可见字符的可选码字，作为双方共享的编码表的构造基础（每个码字必须是 2 个字节的 Unicode 编码）。如表 1 所示给出了一个 256 个可选码字的十六进制表示示例。

[0039] (3) 根据步骤 (1) 和步骤 (2)，建立从密钥到可选码字的编码表 T_k。表 1 给出的即是在密钥 $k = (0,1,2, \cdots,255)$ 时所构造出的编码表。

[0040] 表 1. 编码表

[0041]

m	码字	m	码字	m	码字	m	码字	m	码字	m	码字	m	码字
00	00 D8	25	01 E4	4A	02 F8	6F	04 E3	94	05 F7	B9	07 E2	DE	08 F6
01	00 D9	26	01 E5	4B	03 D8	70	04 E4	95	05 F8	BA	07 E3	DF	08 F7
02	00 DA	27	01 E6	4C	03 D9	71	04 E5	96	06 D8	BB	07 E4	E0	08 F8
03	00 DB	28	01 E7	4D	03 DA	72	04 E6	97	06 D9	BC	07 E5	E1	09 D8
04	00 DC	29	01 F0	4E	03 DB	73	04 E7	98	06 DA	BD	07 E6	E2	09 D9
05	00 DD	2A	01 F1	4F	03 DC	74	04 F0	99	06 DB	BE	07 E7	E3	09 DA
06	00 DE	2B	01 F2	50	03 DD	75	04 F1	9A	06 DC	BF	07 F0	E4	09 DB
07	00 DF	2C	01 F3	51	03 DE	76	04 F2	9B	06 DD	C0	07 F1	E5	09 DC
08	00 E0	2D	01 F4	52	03 DF	77	04 F3	9C	06 DE	C1	07 F2	E6	09 DD
09	00 E1	2E	01 F5	53	03 E0	78	04 F4	9D	06 DF	C2	07 F3	E7	09 DE
0A	00 E2	2F	01 F6	54	03 E1	79	04 F5	9E	06 E0	C3	07 F4	E8	09 DF
0B	00 E3	30	01 F7	55	03 E2	7A	04 F6	9F	06 E1	C4	07 F5	E9	09 E0
0C	00 E4	31	01 F8	56	03 E3	7B	04 F7	A0	06 E2	C5	07 F6	EA	09 E1

0D	00 E5	32	02 D8	57	03 E4	7C	04 F8	A1	06 E3	C6	07 F7	EB	09 E2
0E	00 E6	33	02 D9	58	03 E5	7D	05 D8	A2	06 E4	C7	07 F8	EC	09 E3
0F	00 E7	34	02 DA	59	03 E6	7E	05 D9	A3	06 E5	C8	08 D8	ED	09 E4
10	00 F0	35	02 DB	5A	03 E7	7F	05 DA	A4	06 E6	C9	08 D9	EE	09 E5
11	00 F1	36	02 DC	5B	03 F0	80	05 DB	A5	06 E7	CA	08 DA	EF	09 E6
12	00 F2	37	02 DD	5C	03 F1	81	05 DC	A6	06 F0	CB	08 DB	F0	09 E7
13	00 F3	38	02 DE	5D	03 F2	82	05 DD	A7	06 F1	CC	08 DC	F1	09 F0
14	00 F4	39	02 DF	5E	03 F3	83	05 DE	A8	06 F2	CD	08 DD	F2	09 F1
15	00 F5	3A	02 E0	5F	03 F4	84	05 DF	A9	06 F3	CE	08 DE	F3	09 F2
16	00 F6	3B	02 E1	60	03 F5	85	05 E0	AA	06 F4	CF	08 DF	F4	09 F3
17	00 F7	3C	02 E2	61	03 F6	86	05 E1	AB	06 F5	D0	08 E0	F5	09 F4
18	00 F8	3D	02 E3	62	03 F7	87	05 E2	AC	06 F6	D1	08 E1	F6	09 F5
19	01 D8	3E	02 E4	63	03 F8	88	05 E3	AD	06 F7	D2	08 E2	F7	09 F6
1A	01 D9	3F	02 E5	64	04 D8	89	05 E4	AE	06 F8	D3	08 E3	F8	09 F7
1B	01 DA	40	02 E6	65	04 D9	8A	05 E5	AF	07 D8	D4	08 E4	F9	09 F8
1C	01 DB	41	02 E7	66	04 DA	8B	05 E6	B0	07 D9	D5	08 E5	FA	0A D8
1D	01 DC	42	02 F0	67	04 DB	8C	05 E7	B1	07 DA	D6	08 E6	FB	0A D9
1E	01 DD	43	02 F1	68	04 DC	8D	05 F0	B2	07 DB	D7	08 E7	FC	0A DA
1F	01 DE	44	02 F2	69	04 DD	8E	05 F1	B3	07 DC	D8	08 F0	FD	0A DB
20	01 DF	45	02 F3	6A	04 DE	8F	05 F2	B4	07 DD	D9	08 F1	FE	0A DC
21	01 E0	46	02 F4	6B	04 DF	90	05 F3	B5	07 DE	DA	08 F2	FF	0A DD
22	01 E1	47	02 F5	6C	04 E0	91	05 F4	B6	07 DF	DB	08 F3		
23	01 E2	48	02 F6	6D	04 E1	92	05 F5	B7	07 E0	DC	08 F4		
24	01 E3	49	02 F7	6E	04 E2	93	05 F6	B8	07 E1	DD	08 F5		

[0042] 在本发明中，编码表是作为消息嵌入算法模块和提取算法模块的输入，且发送方和接收方通过秘钥 k 构造的编码表是相同的，所以输入到嵌入算法模块的编码表 T_k 与输入到提取算法模块的编码表 T_k 一致。

[0043] 2、消息嵌入

[0044] 本发明中消息嵌入算法模块执行嵌入算法 E() 可以表示成

[0045] $c' = E(c, m, T_k)$

[0046] 其中，模块输入参数有载体对象 c、待嵌入的秘密消息 m 和嵌入秘钥 k 生成的编码表 T_k，模块输出是包含秘密消息的带秘载体 c′。此外，本发明要求嵌入信息的载体 c 为基于 Unicode 编码的文本数据，包括 TXT 格式文本和 Word、PDF、XML、HTML 复合文档中的文本数据；并且在嵌入模块中 m 作为字节流方式进行处理，所以 m 可以是 TXT 文本数据和 JPEG 图像数据。

[0047] 嵌入算法的实施流程图如图 2 所示，具体实现步骤为：

[0048] (1) 对输入的载体对象 c 进行预处理，用 0x2000 替换 c 中出现在编码表 T_k 的码字；

[0049] (2) 顺序地读取 c 的 2 个字节数据 xx，并判断 xx 值是否为 0x0000；

[0050] (3) 如果步骤 (2) 中结果为 no，则执行步骤 (4)，否则执行步骤 (12a)；

[0051] (4) 判断 xx 是否为不可见字符，包括半角 / 全角空格和制表符，它们对应的 Unicode 码表示分别为 0x2000、0x0030 和 0x0900；

[0052] (5) 如果步骤 (4) 中结果为 no，则跳转到步骤 (2)，否则执行步骤 (6)；

[0053] (6) 对输入的秘密消息 s，读取 1 个字节数据 y；

[0054] (7) 判断 y 值是否为 EOF；

[0055] (8) 如果步骤 (7) 中结果为 no，则执行步骤 (9)，否则执行步骤 (12b)；

[0056] (9) 对输入的编码表 T_k，查找 T_k 中 y 值对应的码字 zz；

[0057] (10) 使用步骤 (9) 中的 zz 值替换步骤 (5) 所得到 c 中的 xx 值；

[0058] (11) 重复执行步骤 (2)；

[0059] (12a) 输出载体嵌入容量不足的提示信息。

[0060] (12b) 输出改变后的 c，即为带秘载体 c′。

[0061] 3、消息提取

[0062] 本发明中消息提取算法模块执行提取算法 D()，可以表示成

[0063] $m = D(c', T_{k'})$

[0064] 其中，模块输入参数有含秘密消息的带秘载体 c′ 和提取秘钥 k′ 生成的编码表 $T_{k'}$，模块输出是提取得到的秘密消息 m。特别地，为了保持编码表 T_k 和 $T_{k'}$ 的一致性，本发明中嵌入秘钥 k 与提取密钥 k′ 是相同的。

[0065] 提取算法的实施流程图如图 3 所示，具体实现步骤为：

[0066] (1) 对输入的带秘载体 c′，读取 2 个字节数据 yy；

[0067] (2) 判断 yy 值是否为 EOF；

[0068] (3) 如果步骤 (2) 中结果为 no，则执行步骤 (4)，否则执行步骤 (8)；

[0069] (4) 根据输入编码表 $T_{k'}$，判断步骤 (2) 中的 yy 值是否为 $T_{k'}$ 中的码字；

[0070] (5) 如果步骤 (4) 中结果为 no，则跳转到步骤 (1)，否则执行步骤 (6)；

[0071] (6) 查找 $T_{k'}$，获得 yy 值所对应的 m 值；

[0072] (7) 重复执行步骤 (1)；

[0073] (8) 输出秘密消息 m。

[0074] 以上实施例仅用以说明本发明的技术方案而非对其进行限制，本领域的普通技术人员可以对本发明的技术方案进行修改或者等同替换，而不脱离本发明的精神和范围，本发明的保护范围应以权利要求所述为准。

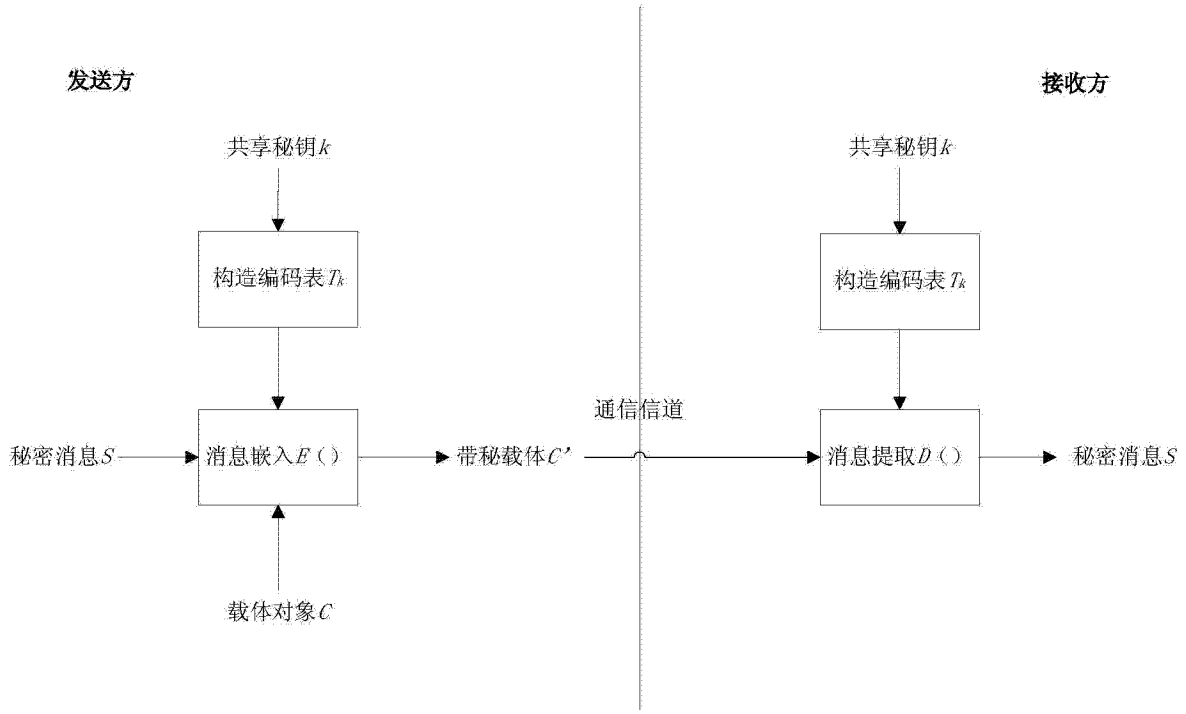

发送方
接收方
共享秘钥k
共享秘钥k
构造编码表Tk
构造编码表Tk
通信信道
秘密消息S
消息嵌入E()
带秘载体C'
消息提取D()
秘密消息S
载体对象C

图 1

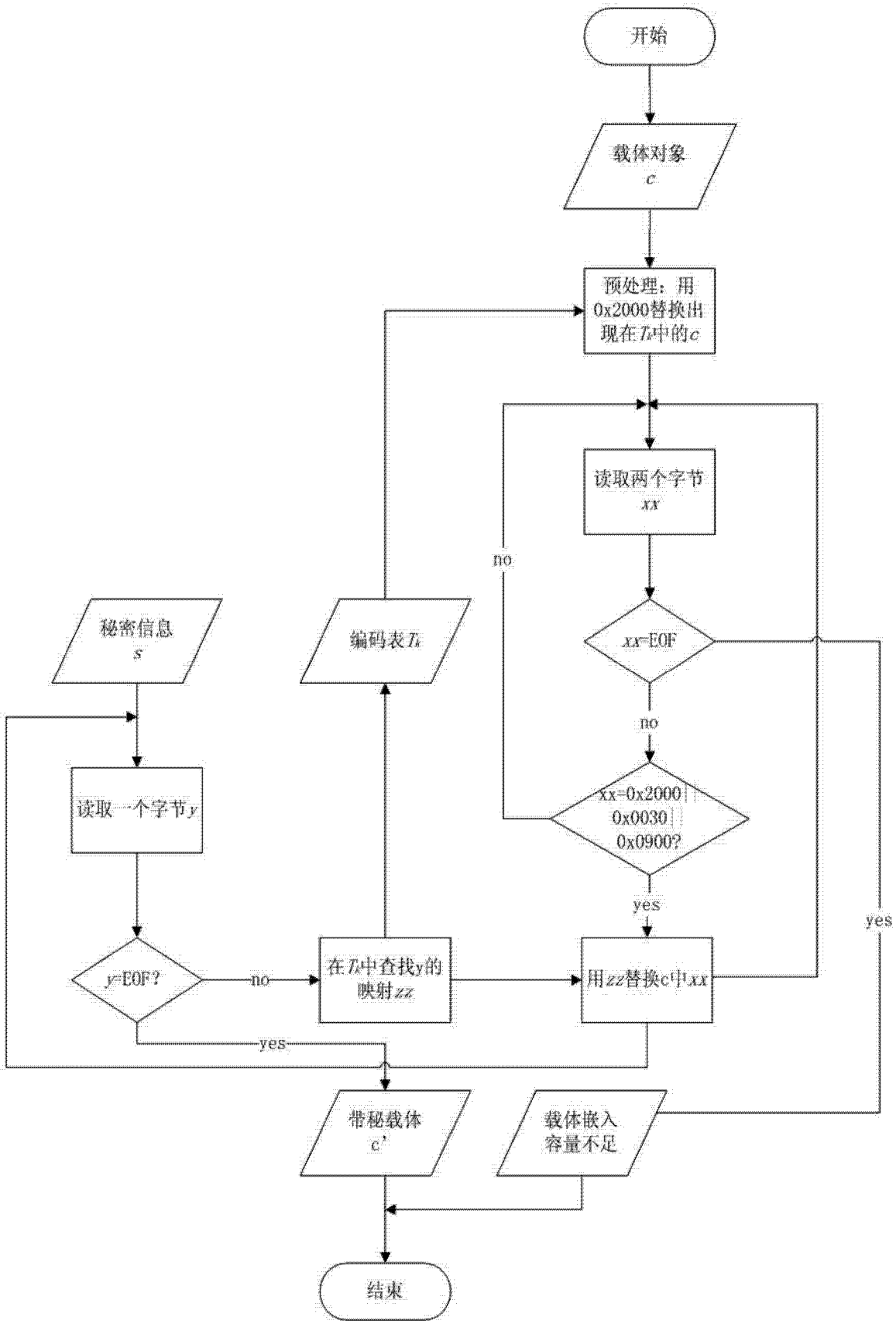

开始
载体对象
c
预处理：用
0x2000替换出
现在Tk中的c
读取两个字节
xx
no
xx=EOF
no
xx=0x2000
0x0030
0x0900?
yes
yes
秘密信息
s
编码表Tk
读取一个字节y
y=EOF?
no
在Tk中查找y的
映射zz
用zz替换c中xx
yes
带秘载体
c'
载体嵌入
容量不足
结束

图 2

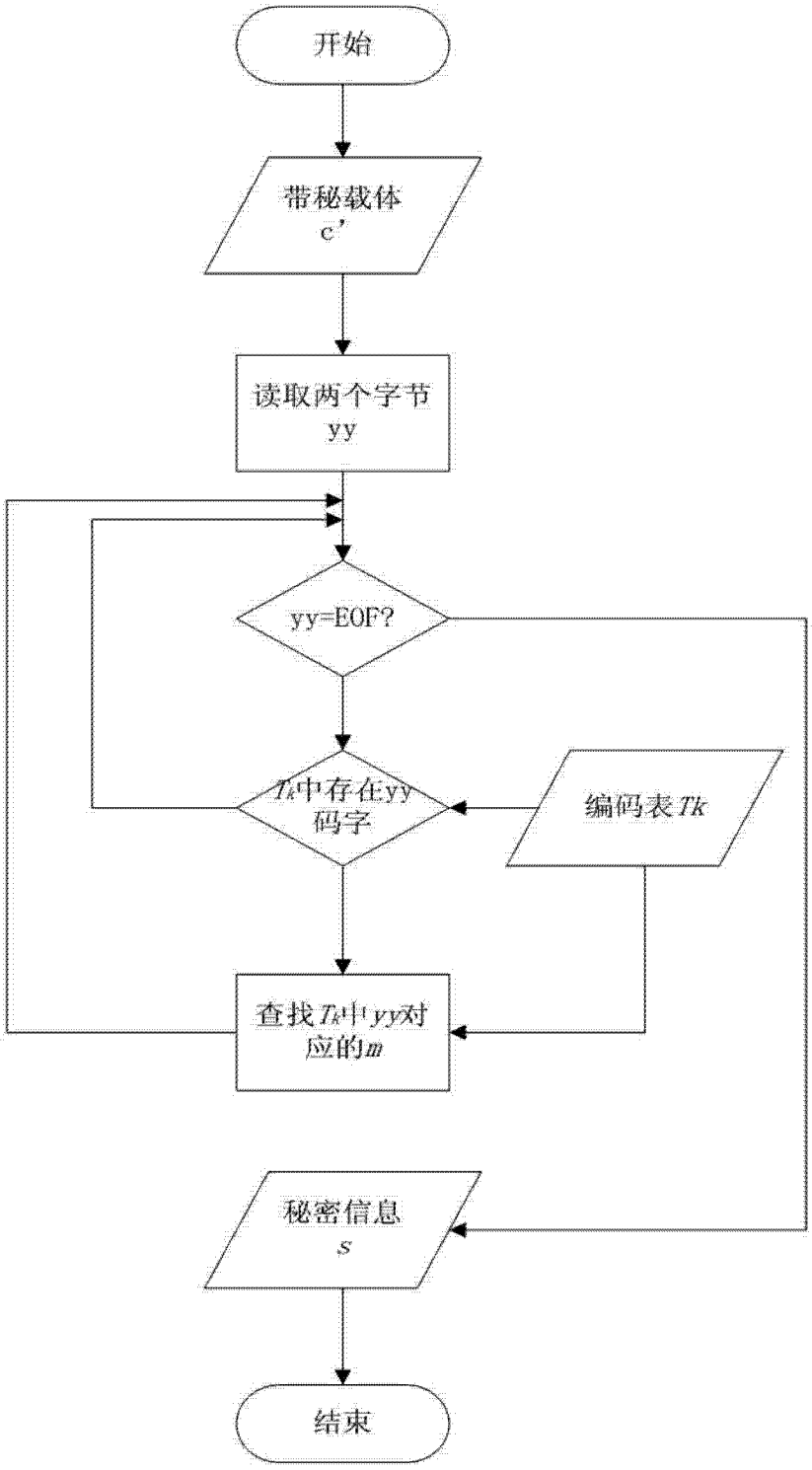

开始
带秘载体
c'
读取两个字节
yy
yy=EOF?
T_k中存在yy
码字
编码表Tk
查找T_k中yy对
应的m
秘密信息
s
结束

图 3