ISSN 2313-1527 (PRINT)
ISSN 2313-1535 (ONLINE)

IJEIE

International Journal of Electronics and Information Engineering

Vol. 4, No. 1 (Mar. 2016)

INTERNATIONAL JOURNAL OF ELECTRONICS & INFORMATION ENGINEERING

PUBLISHING OFFICE

Min-Shiang Hwang
Department of Computer Science & Information Engineering, Asia University, Taichung 41354, Taiwan, R.O.C.
Email: mshwang@asia.edu.tw

International Journal of Electronics and Information Engineering is published both in traditional paper form (ISSN 2313-1527) and in Internet (ISSN 2313-1535) at http://ijeie.jalaxy.com.tw

PUBLISHER: Candy C. H. Lin

23-75, P.O. Box, Taichung, Taiwan 40199, R.O.C.

International Journal of Electronics and Information Engineering

Vol. 4, No. 1 (Mar. 1, 2016)

The Encryption Algorithm AES-RFWKIDEA32-1 Based on Network RFWKIDEA32-1

Aripov Mersaid, Tuychiev Gulom
(Corresponding author: Gulom Tuychiev)

National University of Uzbekistan, Republic of Uzbekistan, Tashkent
(Email: mirsaidaripov@mail.ru, blasterjon@gmail.com)
(Received Aug. 12, 2015; revised and accepted Oct. 2, 2015)

Abstract

In this article, we developed a new block encryption algorithm based on network RFWKIDEA32-1 using of the transformations of the encryption algorithm AES, which is called AES-RFWKIDEA32-1. The block's length of this encryption algorithm is 256 bits, the number of rounds are 10, 12 and 14. The advantages of the encryption algorithms are that, when encryption and decryption process used the same algorithm. In addition, the encryption algorithm AES-RFWKIDEA32-1 encrypts faster than AES.

Keywords: Advanced encryption standard, Feystel network, Lai-Massey scheme, round function, round keys, output transformation

1 Introduction

In September 1997, the National Institute of Standards and Technology issued a public call for proposals for a new block cipher to succeed the Data Encryption Standard [37]. Out of 15 submitted algorithms the Rijndael cipher by Daemen and Rijmen [22] was chosen to become the new Advanced Encryption Standard in November 2001 [15]. The Advanced Encryption Standard is a block cipher with a fixed block length of 128 bits. It supports three different key lengths: 128 bits, 192 bits, and 256 bits. Encrypting a 128-bit block means transforming it in n rounds into a 128-bit output block. The number of rounds n depends on the key length: n =10 for 128-bit keys, n =12 for 192-bit keys, and n=14 for 256-bit keys. The 16-byte input block (t_0, t_1, ..., t_{15}) which is transformed during encryption is usually written as a 4x4 byte matrix, the called AES *State*.

t_0	t_4	t_8	t_{12}
t_1	t_5	t_9	t_{13}
t_2	t_6	t_{10}	t_{14}
t_3	t_7	t_{11}	t_{15}

The structure of each round of AES can be reduced to four basic transformations occurring to the elements of the *State*. Each round consists in applying successively to the *State* the SubBytes(), ShiftRows(), MixColumns() and AddRoundKey() transformations. The first round does the same with an extra AddRoundKey() at the beginning whereas the last round excludes the MixColumns() transformation.

The SubBytes() transformation is a nonlinear byte substitution that operates independently on each byte of the *State* using a substitution table (S-box). Figure 1 illustrates the SubBytes() transformation on the *State*.

In the ShiftRows() transformation operates on the rows of the *State*; it cyclically shifts the bytes in each row by a certain offset. For AES, the first row is left unchanged. Each byte of the second row is shifted one to the left.

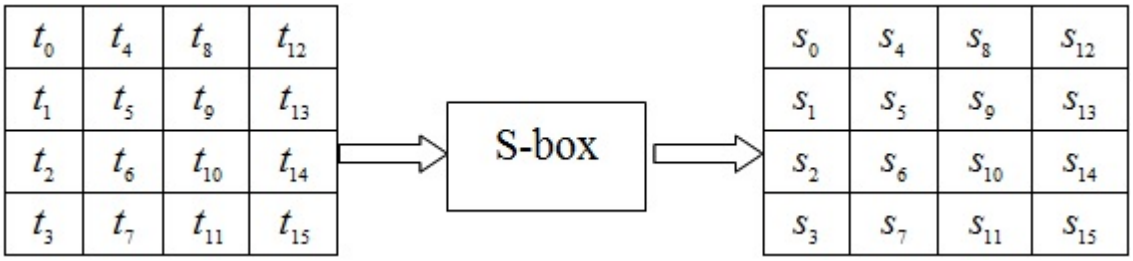

Figure 1: SubBytes() transformation

Similarly, the third and fourth rows are shifted by offsets of two and three respectively. Figure 2 illustrates the ShiftRows() transformation.

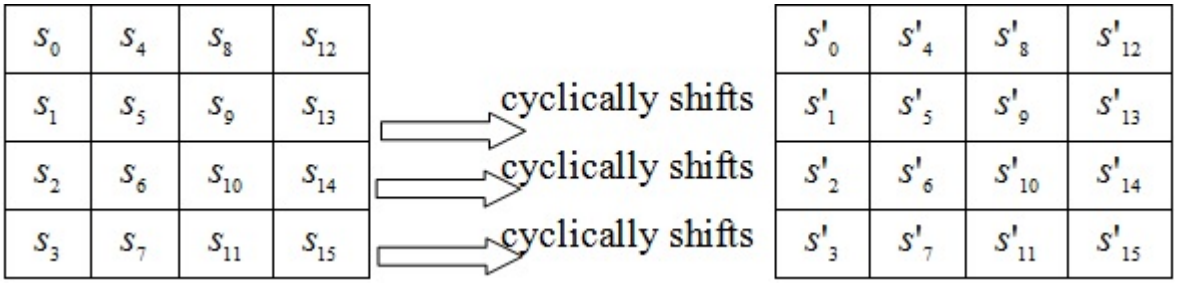

Figure 2: ShiftRows() transformation

The MixColumns() transformation operates on the *State* column-by-column, treating each column as a four-term polynomial. The columns are considered as polynomials over GF(2^8) and multiplied modulo $x^4 + 1$ with a fixed polynomial $a(x)$, given by $a(x) = 3x^2 + x^2 + x + 2$. Let $p = a(x) \otimes s'$:

$$\begin{bmatrix} p_{4i} \\ p_{4i+1} \\ p_{4i+2} \\ p_{4i+3} \end{bmatrix} = \begin{bmatrix} 02\ 03\ 01\ 01 \\ 01\ 02\ 03\ 01 \\ 01\ 01\ 02\ 03 \\ 03\ 01\ 01\ 02 \end{bmatrix} \begin{bmatrix} s'_{4i} \\ s'_{4i+1} \\ s'_{4i+2} \\ s'_{4i+3} \end{bmatrix}, \ i = \overline{0...3}$$

As a result of this multiplication, the four bytes in a column are replaced by the following:

$$\begin{aligned} y_{4i} &= (\{02\} \bullet s'_{4i}) \oplus (\{03\} \bullet s'_{4i+1}) \oplus s'_{4i+2} \oplus s'_{4i+3} \\ y_{4i+1} &= s'_{4i} \oplus (\{02\} \bullet s'_{4i+1}) \oplus (\{03\} \bullet s'_{4i+2}) \oplus s'_{4i+3} \\ y_{4i+2} &= s'_{4i} \oplus s'_{4i+1} \oplus (\{02\} \bullet s'_{4i+2}) \oplus (\{03\} \bullet s'_{4i+3}) \\ y_{4i+4} &= (\{03\} \bullet s'_{4i}) \oplus s'_{4i+1} \oplus s'_{4i+2} \oplus (\{02\} \bullet s'_{4i+3}). \end{aligned}$$

Figure 3 illustrates the MixColumns() transformation.

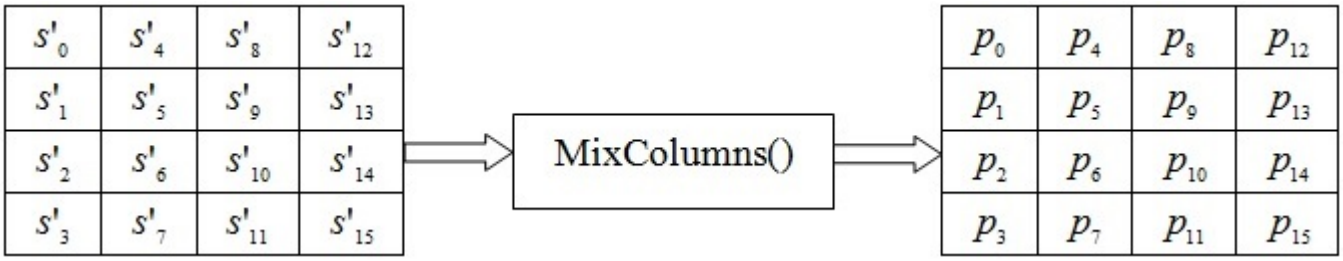

Figure 3: MixColumns() transformation

2 Analysis of AES, PES and IDEA

The first attack is a SQUARE attack suggested in [10] which uses 2^{128} - 2^{119} chosen plaintexts and 2^{120} encryptions. The second attack is a meet-in-the-middle attack proposed in [13] that requires 2^{32} chosen plaintexts and has a time complexity equivalent to almost 2128 encryptions. Recently, another attack on 7-round AES-128 was presented in [19]. The new attack is an impossible differential attack that requires $2^{117.5}$ chosen plaintexts and has a running time of 2^{121} encryptions. Similar results, but with better attack algorithms and lower complexities were reported in [3]. The resulting impossible differential attack on 7-round AES-192 has a data complexity of 292 chosen plaintexts and time complexity of 2^{162} encryptions, while the attack on AES-256 uses $2^{116.5}$ chosen plaintexts and running time of $2^{247.5}$ encryptions.

There are several attacks on AES-192 [3, 9, 10, 16, 19, 24]. The two most notable ones are the SQUARE attack on 8-round AES-192 presented in [10] that requires almost the entire code book and has a running time of 2^{188} encryptions and the meet in the middle attack on 7-round AES-192 in [24] that requires 2^{34+n} chosen plaintexts and has a running time of $2^{208-n} + 2^{82+n}$ encryptions. Legitimate values for n in the meet in the middle attack on AES-192 are 94 ¡ n ¡ 17, thus, the minimal data complexity is 2^{51} chosen plaintexts (with time complexity equivalent to exhaustive search), and the minimal time complexity is 2^{146} (with data complexity of 2^{97} chosen plaintexts). AES-256 is analyzed in [3, 9, 10, 19, 24]. The best attack is the meet in the middle attack in [24] which uses 2^{32} chosen plaintexts and has a total running time of 2^{209} encryptions. Finally, we would like to note the existence of many related-key attacks on AES-192 and AES-256. As the main issue of this paper is not related-key attacks, and as we deal with the single key model, we do not elaborate on the matter here, but the reader is referred to [42] for the latest results on related-key impossible differential attacks on AES and to [17] for the latest results on related-key rectangle attacks on AES.

The strength of AES with respect to impossible differentials was challenged several times. The first attack of this kind is a 5-round attack presented in [5]. This attack is improved in [7] to a 6-round attack. In [16], an impossible differential attack on 7-round AES-192 and AES-256 is presented. The latter attack uses 2^{92} chosen plaintexts (or $2^{92.5}$ chosen plaintexts for AES-256) and has a running time of 2^{186} encryptions (or $2^{250.5}$ encryptions for AES-256). The time complexity of the latter attack was improved in [3] to 2^{162} encryptions for AES-192. In [19] a new 7-round impossible differential attack was presented. The new attack uses a different impossible differential, which is of the same general type as the one used in previous attacks (but has a slightly different structure). Using the new impossible differential leads to an attack that requires $2^{117.5}$ chosen plaintexts and has a running time of 2^{121} encryptions. This attack was later improved in [3, 20] to use $2^{115.5}$ chosen plaintexts with time complexity of 2^{119} encryptions.

The last application of impossible differential cryptanalysis to AES was the extension of the 7-round attack from [19] to 8-round AES-256 in [3]. The extended attack has a data complexity of $2^{116.5}$ chosen plaintexts and time complexity of $2^{247.5}$ encryption. We note that there were three more claimed impossible differential attacks on AES in [40, 41, 43]. However, as all these attacks are flawed [2]. In paper [6] present a new attack on 7-round AES-128, a new attack on 7-round AES-192, and two attacks on 8-round AES-256. The attacks are based on the attacks proposed in [16, 19] but use additional techniques, including the early abort technique and key schedule considerations.

The best attack we present on 8-round AES-256 requires $2^{89.1}$ chosen plaintexts and has a time complexity of $2^{129.7}$ memory accesses. These results are significantly better than any previously published impossible differential attack on AES. We summarize results along with previously known results in Table 1.

The Proposed Encryption Standard (PES) is a 64-bit block cipher, using a 128-bit key, designed by Lai and Massey in 1990 (see [11]) and was a predecessor to IDEA (International Data Encryption Algorithm) [8]. IDEA was originally called IPES (Improved PES). PES iterates eight rounds plus an output transformation. The cryptanalysis of PES and IDEA presented on Table 2 and Table 3.

On the basis of encryption algorithm IDEA and Lai-Massey scheme developed the networks IDEA32-1 and RFWKIDEA32-1, consisting from one round function [27, 36]. In the networks IDEA32-1 and RFWKIDEA32-1, similarly as in the Feistel network, when it encryption and decryption using the same algorithm. In the networks used one round function having 16 input and output blocks and as the round function can use any transformation.

Using transformation SubBytes(), ShiftRows(), MixColumns(), AddRoundKey() AES encryption algorithm as a round function networks IDEA8-1 [28], RFWKIDEA8-1 [28], PES8-1 [29], RFWKPES8-1 [30], IDEA16-1 [26], created encryption algorithms AES-IDEA8-1 [33], AES-RFWKIDEA8-1 [35], AES-PES8-1 [34], AES-RFWKPES8-1 [31], AES-IDEA16-1 [32].

In this paper developed block encryption algorithm AES-RFWKIDEA32-1 based network RFWKIDEA32-1 [36] using transformation of the encryption algorithm AES. The length of block of the encryption algorithms is 256 bits, the number of rounds n equal to 10, 12, 14 and the length of key is variable from 256 bits to 1024 bits in steps 128 bits, i.e., key length is equal to 256, 384, 512, 640, 768, 896 and 1024 bits.

Table 1: A summary of the attacks on AES

Number of rounds	complexity		Attack type
	Data (CP)	Time	
AES-128			
7	2^{128} - 2^{119}	2^{120}	Square [10]
7	$2^{117.5}$	2^{121}	Impossible Differential [10]
7	$2^{117.5}$	2^{119}	Impossible Differential [20, 3]
7	2^{32}	2^{128}	Meet in the middle [13]
7	$2^{112.2}$	$2^{117.2}$ MA	Impossible Differential [6]
AES-192			
7	2^{32}	2^{184}	Square [9]
7	$19 \cdot 2^{32}$	2^{155}	Square [10]
7	2^{92}	$2^{186.2}$	Impossible Differential [16]
7	$2^{115.5}$	2^{119}	Impossible Differential [3]
7	2^{92}	2^{162}	Impossible Differential [3]
7	2^{34+n}	$2^{208-n} + 2^{82+n}$	Meet in the middle [24]
8	$2^{128} - 2^{119}$	2^{188}	Square [10]
7	$2^{113.8}$	$2^{118.8}$ MA	Impossible Differential [6]
7	$2^{91.2}$	$2^{139.2}$	Impossible Differential [6]
AES-256			
7	2^{32}	2^{200}	Square [9]
7	$21 \cdot 2^{32}$	2^{172}	Square [10]
7	$2^{92.5}$	$2^{250.5}$	Impossible Differential [16]
7	2^{32}	2^{208}	Meet in the middle [24]
7	2^{34+n}	$2^{208-n} + 2^{82+n}$	Meet in the middle [24]
7	$2^{115.5}$	2^{119}	Impossible Differential [3]
8	$2^{116.5}$	$2^{247.5}$	Impossible Differential [3]
8	$2^{128} - 2^{119}$	2^{204}	Square [10]
8	2^{32}	2^{209}	Meet in the middle [24]
7	$2^{113.8}$	$2^{118.8}$ MA	Impossible Differential [6]
7	2^{92}	2^{163} MA	Impossible Differential [6]
8	$2^{111.1}$	$2^{227.8}$ MA	Impossible Differential [6]
8	$2^{89.1}$	$2^{229.7}$ MA	Impossible Differential [6]

Table 2: A summary of the attacks on IDEA

Attack Type	Year	Attacked Rounds	Key Bits round	Chosen Plaintext	Time
Differential [12]	1993	2	32	2^{10}	2^{42}
Differential [38]	1993	2.5	32	2^{10}	2^{32}
Differential [12]	1993	2.5	96	2^{10}	2^{106}
Related-Key Differential [39]	1996	3	32	6	$6 \cdot 2^{32}$
Differential-Linear [21]	1996	3	32	2^{30}	2^{44}
Differential [1]	1996	3	32	2^{30}	$0.75 \cdot 2^{44}$
Truncated Differential [23, 21]	1997	3.5	48	2^{56}	2^{67}
Miss-in-the-middle [25]	1998	3.5	64	$2^{38.5}$	2^{53}
Miss-in-the-middle [25]	1998	4	69	2^{37}	2^{70}
Related-Key Differential-Linear [4]	1998	4	15	38.3	-
Miss-in-the-Middle [25]	1998	4.5	80	2^{64}	2^{112}
Square attack [18]	2000	2.5	77	$3 \cdot 2^{16}$	$2^{63} + 2^{47}$
Square attack [18]	2000	2.5	31	2^{32}	2^{62}
Square [18]	2000	2.5	31	2^{48}	2^{79}
Related-Key Square [18]	2001	2.5	32	2	2^{41}

Table 3: A summary of the attacks on PES

Attack Type	Year	Attacked Rounds	Key Bits round	Chosen Plaintext	Time
Differential [14]	1991	7	96	2^{64}	2^{160}
Square [18]	2000	2.5	31	2^{17}	2^{47}
Square [18]	2001	2.5	31	2^{32}	2^{63}
Related-Key Square [18]	2001	2.5	32	2	241

3 The Encryption Algorithm AES-RFWKIDEA32-1

3.1 The Structure of the Encryption Algorithm AES-RFWKIDEA32-1

In the encryption algorithm AES-RFWKIDEA32-1 as the round function used SubBytes(), ShiftRows(), MixColumns() transformation of encryption algorithm AES. The scheme n-rounded encryption algorithm AES-RFWKIDEA32-1 shown in Figure 4, and the length of subblocks X^0, X^1, ..., X^{31}, length of round keys $K_{32(i-1)}$, $K_{32(i-1)+1}$, ..., $K_{32(i-1)+31}$,, $i = \overline{1...n+1}$ and K_{32n+32}, K_{32n+33}, ..., K_{32n+95} are equal to 8-bits.

Consider the round function of the encryption algorithm AES-RFWKIDEA32-1. Initially 32-bit subblocks t_0, t_1, ..., t_{15} are written into the *State* array and are executed the above transformations SubBytes(), ShiftRows(), MixColumns(). After the MixColumns() transformation we obtain 8-bits subblocks y_0, y_1, ..., y_{15}, where y_0=p_0, y_1=p_1, ..., y_{15}=p_{15}.

The S-box SubBytes() transformation shown in Table 4 and is the only nonlinear transformation. The length of the input and output blocks S-box is eight bits. For example, if the input value the S-box is equal to 0xE7, then the output value is equal 0x79, i.e. selected elements of intersection row 0xE and column 0x7.

Table 4: The S-box of encryption algorithm AES-RFWKIDEA32-1

	0x0	**0x1**	**0x2**	**0x3**	**0x4**	**0x5**	**0x6**	**0x7**	**0x8**	**0x9**	**0xA**	**0xB**	**0xC**	**0xD**	**0xE**	**0xF**
0x0	0x87	0x1C	0x05	0x06	0x13	0x86	0x84	0xC9	0x3F	0xEF	0x85	0xA6	0x10	0x41	0xA2	0x15
0x1	0xD2	0xF3	0xCA	0x0C	0x12	0x4E	0xC5	0x1B	0xA8	0x59	0xB3	0xA0	0x78	0xB9	0x17	0xDB
0x2	0x21	0x08	0x63	0xB5	0x35	0x24	0x01	0xD8	0x3D	0xA9	0x89	0x0B	0x0F	0x5A	0x2F	0x6D
0x3	0xFD	0xC1	0xA7	0xC3	0x7E	0x71	0xED	0x72	0xE5	0x77	0xFB	0x93	0x82	0xA5	0x33	0x0D
0x4	0xEE	0xE3	0xBC	0x76	0x66	0x94	0x56	0xBB	0x57	0x26	0x51	0x23	0xAE	0x83	0xA4	0xF9
0x5	0x47	0x4B	0xFF	0x88	0xBF	0x18	0x2B	0x46	0x96	0xC2	0x30	0x2E	0xD6	0xDC	0x5E	0xC0
0x6	0x5B	0x80	0xB2	0x02	0xC7	0xCC	0x27	0xE9	0xCD	0x0A	0xF7	0x04	0x5F	0x3C	0x60	0xBA
0x7	0x4F	0xA3	0xDF	0xE0	0x73	0x68	0x3E	0x09	0x38	0x31	0x52	0xAF	0x7F	0x00	0x03	0x53
0x8	0xC8	0xFC	0x67	0x98	0x44	0x61	0xDD	0x65	0xD9	0xA1	0x14	0x2C	0x9D	0x4C	0x6E	0x07
0x9	0x9F	0xEB	0xC4	0x58	0xB7	0xB6	0x7B	0xFA	0xD5	0x90	0x3A	0x7D	0x50	0x54	0xE6	0x42
0xA	0x9B	0x37	0x36	0xF6	0xCE	0xF5	0xBD	0x5C	0xD3	0x43	0xB8	0x97	0x6B	0x69	0x99	0x0E
0xB	0x81	0xDA	0x25	0x8C	0xE8	0x49	0xD4	0xAA	0x9C	0x55	0x19	0x92	0x8D	0x16	0xB0	0xFE
0xC	0x32	0x1E	0xAD	0xB4	0x7C	0xB1	0x39	0xD1	0x9A	0x48	0x1D	0x64	0xC6	0x28	0xE2	0xF2
0xD	0x1F	0x34	0x29	0x95	0xDE	0xE7	0x11	0xF4	0x8F	0x2D	0x45	0x2A	0xF1	0xCB	0x6C	0x70
0xE	0x8B	0x1A	0x7A	0x6F	0x8E	0x4A	0xF0	0x79	0x62	0x74	0xE1	0x8A	0xD0	0x4D	0xBE	0x40
0xF	0xF8	0xAB	0xEA	0xEC	0x20	0x91	0xD7	0x9E	0xCF	0x6A	0xAC	0xE4	0x3B	0x5D	0x22	0x75

Consider the encryption process of encryption algorithm AES-RFWKIDEA32-1. Initially the 256-bit plaintext X partitioned into subblocks of 8-bits X_0^0, X_0^1, ..., X_0^{31}, and performs the following steps:

1) Subblocks X_0^0, X_0^1, ..., X_0^{31} summed by XOR respectively with round keys K_{32n+32}, K_{32n+33}, ..., K_{32n+63}: $X_0^j = X_0^j \oplus K_{32n+32+j}$, $j = \overline{0...31}$.

2) Subblocks X_0^0, X_0^1, ..., X_0^{31} multiplied and summed respectively with the round keys $K_{32(i-1)}$, $K_{32(i-1)+1}$,

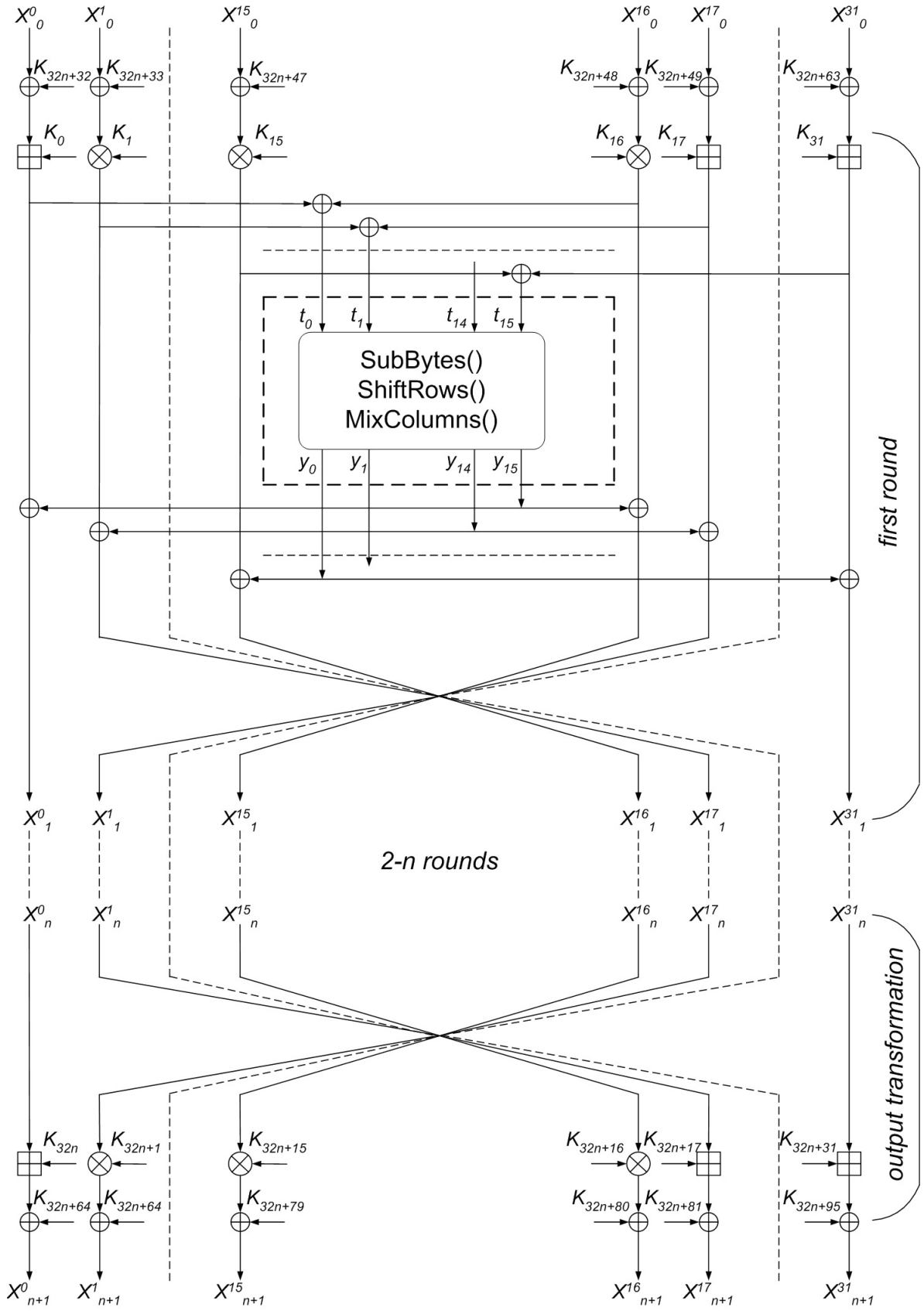

Figure 4: The scheme n-rounded encryption algorithm AES-RFWKIDEA32-1

$\ldots$, $K_{32(i-1)+31}$ and calculated 8-bit subblocks t_0, t_1, $\ldots$, t_{15}. This step can be represented as follows:

$$\begin{aligned}
t_0 &= (X_{i-1}^0 + K_{32(i-1)}) \oplus (X_{i-1}^{16} \cdot K_{32(i-1)+16}),\\
t_1 &= (X_{i-1}^1 \cdot K_{32(i-1)+1}) \oplus (X_{i-1}^{17} + K_{32(i-1)+17}),\\
t_2 &= (X_{i-1}^2 + K_{32(i-1)+2}) \oplus (X_{i-1}^{18} \cdot K_{32(i-1)+18}),\\
t_3 &= (X_{i-1}^3 \cdot K_{32(i-1)+3}) \oplus (X_{i-1}^{19} + K_{32(i-1)+19}),\\
t_4 &= (X_{i-1}^4 + K_{32(i-1)+4}) \oplus (X_{i-1}^{20} \cdot K_{32(i-1)+20}),\\
t_5 &= (X_{i-1}^5 \cdot K_{32(i-1)+5}) \oplus (X_{i-1}^{21} + K_{32(i-1)+21}),\\
t_6 &= (X_{i-1}^6 + K_{32(i-1)+6}) \oplus (X_{i-1}^{22} \cdot K_{32(i-1)+22}),\\
t_7 &= (X_{i-1}^7 \cdot K_{32(i-1)+7}) \oplus (X_{i-1}^{23} + K_{32(i-1)+23}),\\
t_8 &= (X_{i-1}^8 + K_{32(i-1)+8}) \oplus (X_{i-1}^{24} \cdot K_{32(i-1)+24}),\\
t_9 &= (X_{i-1}^9 \cdot K_{32(i-1)+9}) \oplus (X_{i-1}^{25} + K_{32(i-1)+25}),\\
t_{10} &= (X_{i-1}^{10} + K_{32(i-1)+10}) \oplus (X_{i-1}^{26} \cdot K_{32(i-1)+26}),\\
t_{11} &= (X_{i-1}^{11} \cdot K_{32(i-1)+11}) \oplus (X_{i-1}^{27} + K_{32(i-1)+27}),\\
t_{12} &= (X_{i-1}^{12} + K_{32(i-1)+12}) \oplus (X_{i-1}^{28} \cdot K_{32(i-1)+28}),\\
t_{13} &= (X_{i-1}^{13} \cdot K_{32(i-1)+13}) \oplus (X_{i-1}^{29} + K_{32(i-1)+29}),\\
t_{14} &= (X_{i-1}^{14} + K_{32(i-1)+14}) \oplus (X_{i-1}^{30} \cdot K_{32(i-1)+30}),\\
t_{15} &= (X_{i-1}^{15} \cdot K_{32(i-1)+15}) \oplus (X_{i-1}^{31} + K_{32(i-1)+31}), i = 1.
\end{aligned}$$

3) Performed SubBytes(), ShiftRows(), MixColumns() transformation. Output subblocks of the round function of the encryption algorithm are y_0, y_1, $\ldots$, y_{31}.

4) Subblocks y_0, y_1, $\ldots$, y_{31} are summed to XOR with subblocks X_{i-1}^0, X_{i-1}^1, $\ldots$, X_{i-1}^{31}, i.. $X_{i-1}^j = X_{i-1}^j \oplus y_{15-j}$, $X_{i-1}^{j+16} = X_{i-1}^{j+16} \oplus y_{15-j}$, $j = \overline{0...15}$, $i = 1$.

5) At the end of the round subblocks X_{i-1}^j and X_{i-1}^{31-j}, $j = \overline{1...15}$ swapped, i.., $X_i^0 = X_{i-1}^0$, $X_i^1 = X_{i-1}^{30}$, $X_i^2 = X_{i-1}^{29}$, $X_i^3 = X_{i-1}^{28}$, $X_i^4 = X_{i-1}^{27}$, $X_i^5 = X_{i-1}^{26}$, $X_i^6 = X_{i-1}^{25}$, $X_i^7 = X_{i-1}^{24}$, $X_i^8 = X_{i-1}^{23}$, $X_i^9 = X_{i-1}^{22}$, $X_i^{10} = X_{i-1}^{21}$, $X_i^{11} = X_{i-1}^{20}$, $X_i^{12} = X_{i-1}^{19}$, $X_i^{13} = X_{i-1}^{18}$, $X_i^{14} = X_{i-1}^{17}$, $X_i^{15} = X_{i-1}^{16}$, $X_i^{16} = X_{i-1}^{15}$, $X_i^{17} = X_{i-1}^{14}$, $X_i^{18} = X_{i-1}^{13}$, $X_i^{19} = X_{i-1}^{12}$, $X_i^{20} = X_{i-1}^{11}$, $X_i^{21} = X_{i-1}^{10}$, $X_i^{22} = X_{i-1}^{9}$, $X_i^{23} = X_{i-1}^{8}$, $X_i^{24} = X_{i-1}^{7}$, $X_i^{25} = X_{i-1}^{6}$, $X_i^{26} = X_{i-1}^{5}$, $X_i^{27} = X_{i-1}^{4}$, $X_i^{28} = X_{i-1}^{3}$, $X_i^{29} = X_{i-1}^{2}$, $X_i^{30} = X_{i-1}^{1}$, $X_i^{31} = X_{i-1}^{31}$, $i = 1$.

6) Repeating steps 2-5 n times, i.e., $i = \overline{2...n}$ obtain subblocks X_n^0, X_n^1, $\ldots$, X_n^{31}.

7) In output transformation round keys are multiplied and summed into subblocks, i.e. $X_{n+1}^0 = X_n^0 + K_{32n}$, $X_{n+1}^1 = X_n^{30} \cdot K_{32n+1}$, $X_{n+1}^2 = X_n^{29} + K_{32n+2}$, $X_{n+1}^3 = X_n^{28} \cdot K_{32n+3}$, $X_{n+1}^4 = X_n^{27} + K_{32n+4}$, $X_{n+1}^5 = X_n^{26} \cdot K_{32n+5}$, $X_{n+1}^6 = X_n^{25} + K_{32n+6}$, $X_{n+1}^7 = X_n^{24} \cdot K_{32n+7}$, $X_{n+1}^8 = X_n^{23} + K_{32n+8}$, $X_{n+1}^9 = X_n^{22} \cdot K_{32n+9}$, $X_{n+1}^{10} = X_n^{21} + K_{32n+10}$, $X_{n+1}^{11} = X_n^{20} \cdot K_{32n+11}$, $X_{n+1}^{12} = X_n^{19} + K_{32n+12}$, $X_{n+1}^{13} = X_n^{18} \cdot K_{32n+13}$, $X_{n+1}^{14} = X_n^{17} + K_{32n+14}$, $X_{n+1}^{15} = X_n^{16} \cdot K_{32n+15}$, $X_{n+1}^{16} = X_n^{15} \cdot K_{32n+16}$, $X_{n+1}^{17} = X_n^{14} + K_{32n+17}$, $X_{n+1}^{18} = X_n^{13} \cdot K_{32n+18}$, $X_{n+1}^{19} = X_n^{12} + K_{32n+19}$, $X_{n+1}^{20} = X_n^{11} \cdot K_{32n+20}$, $X_{n+1}^{21} = X_n^{10} + K_{32n+21}$, $X_{n+1}^{22} = X_n^{9} \cdot K_{32n+22}$, $X_{n+1}^{23} = X_n^{8} + K_{32n+23}$, $X_{n+1}^{24} = X_n^{7} \cdot K_{32n+24}$, $X_{n+1}^{25} = X_n^{6} + K_{32n+25}$, $X_{n+1}^{26} = X_n^{5} \cdot K_{32n+26}$, $X_{n+1}^{27} = X_n^{4} + K_{32n+27}$, $X_{n+1}^{28} = X_n^{3} \cdot K_{32n+28}$, $X_{n+1}^{29} = X_n^{2} + K_{32n+29}$, $X_{n+1}^{30} = X_n^{1} \cdot K_{32n+30}$, $X_{n+1}^{31} = X_n^{31} + K_{32n+31}$;

8) Subblocks X_{n+1}^0, X_{n+1}^1, $\ldots$, X_{n+1}^{31} are summed to XOR with the round key K_{32n+64}, K_{32n+65}, $\ldots$, K_{32n+95}: $X_{n+1}^j = X_{n+1}^j \oplus K_{32n+64+j}$, $j = \overline{0...31}$. As ciphertext plaintext X receives the combined 16-bit subblocks $X_{n+1}^0 || X_{n+1}^1 || ... || X_{n+1}^{31}$.

3.2 Key Generation of the Encryption Algorithm AES-RFWKIDEA32-1

In n-rounded encryption algorithm AES-RFWKIDEA32-1 in each round we applied sixteen (32) round keys of the 8-bit and output transformation thirty two round keys of the 8-bit. In addition, before the first round and after the output transformation we used thirty two round keys of 8-bits. Total number of 8-bit round keys is equal to $32n+96$. In Figure 4 encryption used encryption round keys K_i^c instead of K_i, while decryption used decryption round keys K_i^d. If n=10 then need 416 to generate round keys, if n=12, you need to generate 480 round keys and if n=14 need 544 to generate round keys.

When generating round keys like the AES encryption algorithm uses an array Rcon: Rcon=[0x01, 0x02, 0x04, 0x08, 0x10, 0x20, 0x40, 0x80].

The key encryption algorithm K of length l ($256 \le l \le 1024$) bits is divided into 8-bit round keys K_0^c, K_1^c,..., $K_{Lenght-1}^c$, $Lenght = l/8$, here $K = \{k_0, k_1, ..., k_{l-1}\}$, $K_0^c = \{k_0, k_1, ..., k_7\}$, $K_1^c = \{k_8, k_9, ..., k_{15}\}$,..., $K_{Lenght-1}^c = \{k_{l-8}, k_{l-7}, ..., k_{l-1}\}$ and $K = K_0^c || K_1^c || ... || K_{Lenght-1}^c$. Then we calculate $K_L = K_0^c \oplus K_1^c \oplus ... \oplus K_{Lenght-1}^c$. If $K_L = 0$ then K_L is chosen as 0xC5, i.e. $K_L =$ 0xC5. When generating a round keys K_i^c, $i = \overline{Lenght...32n+95}$, we used transformation *SubBytes()* and *RotWord8()*, here *SubBytes()*-is transformation 8-bit subblock into S-box and *RotWord8()*-cyclic shift to the left of 1 bit of the 8-bit subblock. When the condition $imod3 = 1$ is true, then the round keys are computed as $K_i^c = SubBytes(K_{i-Lenght+1}^c) \oplus SubBytes(RotWord8(K_{i-Lenght}^c)) \oplus Rcon[imod8] \oplus K_L$, otherwise $K_i^c = SubBytes(K_{i-Lenght}^c) \oplus SubBytes(K_{i-Lenght+1}^c) \oplus K_L$. After each round key generation the value K_L is cyclic shift to the left by 1 bit.

Decryption round keys are computed on the basis of encryption round keys and decryption round keys of the output transformation associate with of encryption round keys as follows:

$$\begin{aligned}
&(K_{32n}^d, K_{32n+1}^d, K_{32n+2}^d, K_{32n+3}^d, K_{32n+4}^d, K_{32n+5}^d, K_{32n+6}^d, K_{32n+7}^d, K_{32n+8}^d, K_{32n+9}^d, K_{32n+10}^d, K_{32n+11}^d,\\
&\quad K_{32n+12}^d, K_{32n+13}^d, K_{32n+14}^d, K_{32n+15}^d, K_{32n+16}^d, K_{32n+17}^d, K_{32n+18}^d, K_{32n+19}^d, K_{32n+20}^d, K_{32n+21}^d, K_{32n+22}^d,\\
&\quad K_{32n+23}^d, K_{32n+24}^d, K_{32n+25}^d, K_{32n+26}^d, K_{32n+27}^d, K_{32n+28}^d, K_{32n+29}^d, K_{32n+30}^d, K_{32n+31}^d)\\
=\ &(-K_0^c, (K_1^c)^{-1}, -K_2^c, (K_3^c)^{-1}, -K_4^c, (K_5^c)^{-1}, -K_6^c, (K_7^c)^{-1}, -K_8^c, (K_9^c)^{-1}, -K_{10}^c, (K_{11}^c)^{-1},\\
&\quad -K_{12}^c, (K_{13}^c)^{-1}, -K_{14}^c, (K_{15}^c)^{-1}, (K_{16}^c)^{-1}, -K_{17}^c, (K_{18}^c)^{-1}, -K_{19}^c, (K_{20}^c)^{-1}, -K_{21}^c, (K_{22}^c)^{-1}, -K_{23}^c, (K_{24}^c)^{-1},\\
&\quad -K_{25}^c, (K_{26}^c)^{-1}, -K_{27}^c, (K_{28}^c)^{-1}, -K_{29}^c, (K_{30}^c)^{-1}, -K_{31}^c).
\end{aligned}$$

For example, if the number of rounds n is 10 the formula is as follows:

$$\begin{aligned}
&(K_{320}^d, K_{321}^d, K_{322}^d, K_{323}^d, K_{324}^d, K_{325}^d, K_{326}^d, K_{327}^d, K_{328}^d, K_{329}^d, K_{330}^d, K_{331}^d, K_{332}^d, K_{333}^d, K_{334}^d, K_{335}^d, K_{336}^d,\\
&\quad K_{337}^d, K_{338}^d, K_{339}^d, K_{340}^d, K_{341}^d, K_{342}^d, K_{343}^d, K_{344}^d, K_{345}^d, K_{346}^d, K_{347}^d, K_{348}^d, K_{349}^d, K_{350}^d, K_{351}^d)\\
=\ &(-K_0^c, (K_1^c)^{-1}, -K_2^c, (K_3^c)^{-1}, -K_4^c, (K_5^c)^{-1}, -K_6^c, (K_7^c)^{-1}, -K_8^c, (K_9^c)^{-1}, -K_{10}^c, (K_{11}^c)^{-1}, -K_{12}^c,\\
&\quad (K_{13}^c)^{-1}, -K_{14}^c, (K_{15}^c)^{-1}, (K_{16}^c)^{-1},\\
&\quad -K_{17}^c, (K_{18}^c)^{-1}, -K_{19}^c, (K_{20}^c)^{-1}, -K_{21}^c, (K_{22}^c)^{-1}, -K_{23}^c, (K_{24}^c)^{-1},\\
&\quad -K_{25}^c, (K_{26}^c)^{-1}, -K_{27}^c, (K_{28}^c)^{-1}, -K_{29}^c, (K_{30}^c)^{-1}, -K_{31}^c).
\end{aligned}$$

Decryption round keys of the first round associates with the encryption round keys as follows:

$$\begin{aligned}
&(K_0^d, K_1^d, K_2^d, K_3^d, K_4^d, K_5^d, K_6^d, K_7^d, K_8^d, K_9^d, K_{10}^d, K_{11}^d, K_{12}^d, K_{13}^d, K_{14}^d, K_{15}^d, K_{16}^d, K_{17}^d, K_{18}^d, K_{19}^d, K_{20}^d, K_{21}^d,\\
&\quad K_{22}^d, K_{23}^d, K_{24}^d, K_{25}^d, K_{26}^d, K_{27}^d, K_{28}^d, K_{29}^d, K_{30}^d, K_{31}^d)\\
=\ &(-K_{32n}^c, (K_{32n+1}^c)^{-1}, -K_{32n+2}^c, (K_{32n+3}^c)^{-1}, -K_{32n+4}^c, (K_{32n+5}^c)^{-1}, -K_{32n+6}^c, (K_{32n+7}^c)^{-1}, -K_{32n+8}^c,\\
&\quad (K_{32n+9}^c)^{-1}, -K_{32n+10}^c, (K_{32n+11}^c)^{-1}, -K_{32n+12}^c, (K_{32n+13}^c)^{-1}, -K_{32n+14}^c, (K_{32n+15}^c)^{-1}, (K_{32n+16}^c)^{-1},\\
&\quad -K_{32n+17}^c, (K_{32n+18}^c)^{-1}, -K_{32n+19}^c, (K_{32n+20}^c)^{-1}, -K_{32n+21}^c, (K_{32n+22}^c)^{-1}, -K_{32n+23}^c, (K_{32n+24}^c)^{-1},\\
&\quad -K_{32n+25}^c, (K_{32n+26}^c)^{-1}, -K_{32n+27}^c, (K_{32n+28}^c)^{-1}, -K_{32n+29}^c, (K_{32n+30}^c)^{-1}, -K_{32n+31}^c).
\end{aligned}$$

Likewise, the decryption round keys of the second, third and n-round associates with the encryption round keys as follows:

$$\begin{aligned}
&(K_{32(i-1)}^d, K_{32(i-1)+1}^d, K_{32(i-1)+2}^d, K_{32(i-1)+3}^d, K_{32(i-1)+4}^d, K_{32(i-1)+5}^d, K_{32(i-1)+6}^d, K_{32(i-1)+7}^d, K_{32(i-1)+8}^d,\\
&\quad K_{32(i-1)+9}^d, K_{32(i-1)+10}^d, K_{32(i-1)+11}^d, K_{32(i-1)+12}^d, K_{32(i-1)+13}^d, K_{32(i-1)+14}^d, K_{32(i-1)+15}^d, K_{32(i-1)+16}^d,\\
&\quad K_{32(i-1)+17}^d, K_{32(i-1)+18}^d, K_{32(i-1)+19}^d, K_{32(i-1)+20}^d, K_{32(i-1)+21}^d, K_{32(i-1)+22}^d, K_{32(i-1)+23}^d, K_{32(i-1)+24}^d,\\
&\quad K_{32(i-1)+25}^d, K_{32(i-1)+26}^d, K_{32(i-1)+27}^d, K_{32(i-1)+28}^d, K_{32(i-1)+29}^d, K_{32(i-1)+30}^d, K_{32(i-1)+31}^d)\\
=\ &(-K_{32(n-i+1)}^c, (K_{32(n-i+1)+30}^c)^{-1}, -K_{32(n-i+1)+29}^c, (K_{32(n-i+1)+28}^c)^{-1}, -K_{32(n-i+1)+27}^c, (K_{32(n-i+1)+26}^c)^{-1},\\
&\quad -K_{32(n-i+1)+25}^c, (K_{32(n-i+1)+24}^c)^{-1}, -K_{32(n-i+1)+23}^c, (K_{32(n-i+1)+22}^c)^{-1}, -K_{32(n-i+1)+21}^c,\\
&\quad (K_{32(n-i+1)+20}^c)^{-1}, -K_{32(n-i+1)+19}^c, (K_{32(n-i+1)+18}^c)^{-1}, -K_{32(n-i+1)+17}^c, (K_{32(n-i+1)+16}^c)^{-1},\\
&\quad (K_{32(n-i+1)+15}^c)^{-1}, -K_{32(n-i+1)+14}^c, (K_{32(n-i+1)+13}^c)^{-1}, -K_{32(n-i+1)+12}^c, (K_{32(n-i+1)+11}^c)^{-1},\\
&\quad -K_{32(n-i+1)+10}^c, (K_{32(n-i+1)+9}^c)^{-1}, -K_{32(n-i+1)+8}^c, (K_{32(n-i+1)+7}^c)^{-1}, -K_{32(n-i+1)+6}^c, (K_{32(n-i+1)+5}^c)^{-1},\\
&\quad -K_{32(n-i+1)+4}^c, (K_{32(n-i+1)+3}^c)^{-1}, -K_{32(n-i+1)+2}^c, (K_{32(n-i+1)+1}^c)^{-1}, -K_{32(n-i+1)+31}^c), i = \overline{2...n}
\end{aligned}$$

Decryption round keys applied to the first round and after the output transformation associated with the encryption round keys as follows: $K^d_{32n+32+j} = K^c_{32n+64+j}$, $K^d_{32n+64+j} = K^c_{32n+32+j}$, $j = \overline{0...31}$.

4 Results

Using the transformations SubBytes(), ShiftRows(), MixColumns() of the encryption algorithm AES as the round function network RFWKIDEA32-1 we developed encryption algorithm AES-RFWKIDEA32-1. In the algorithm, the number of rounds of encryption and key's length is variable and the user can select the number of rounds and the key's length in dependence of the degree of secrecy of information and speed encryption.

As in the encryption algorithms based on the Feistel network, the advantages of the encryption algorithm AES-RFWKIDEA32-1 are that, when encryption and decryption process used the same algorithm. In the encryption algorithm AES-RFWKIDEA32-1 in decryption process encryption round keys are used in reverse order, thus on the basis of operations necessary to compute the inverse. For example, if the round key is multiplied by the subblock, while decryption is is necessary to calculate the multiplicative inverse, if summarized, it is necessary to calculate the additive inverse.

It is known that the resistance of AES encryption algorithm is closely associated with resistance S-box, applied in the algorithm. In the S-box's encryption algorithm AES algebraic degree of nonlinearity deg = 7, nonlinearity $NL = 112$, resistance to linear cryptanalysis $\lambda = 32/256$, resistance to differential cryptanalysis $\delta = 4/256$, strict avalanche criterion SAC = 8, bit independence criterion BIC = 8.

In the encryption algorithm AES-RFWKIDEA32-1 resistance S-box is equal to resistance S-box's encryption algorithm AES, i.e., deg = 7, $NL = 112$, $\lambda = 32/256$, $\delta = 4/256$, SAC= BIC=8.

5 Conclusions

It is known that as a algorithms based of Feistel network, the resistance algorithm based on networks RFWKIDEA32-1 closely associated with resistance round function. Therefore, selecting the transformations SubBytes(), ShiftRows(), MixColumns() of the encryption algorithm AES, based on round function network RFWKIDEA32-1 we developed relatively resistant encryption algorithm.

References

[1] J. Borst, "Differential-linear cryptanalysis of idea," *Department of Electrical Engineering, ESATCOSIC Technical Report 96/2, 14 pages.*

[2] J. Chen, "Personal communications," august 2008.

[3] D. Feng, W. Zhang, W. Wu, "New results on impossible differential cryptanalysis of reduced AES," in *Proceedings of ICISC 2007*, LNCS 4817, pp. 239–250, 2007.

[4] P. Hawkes, "Differential-linear weak key classes of idea," in *Advances in Cryptology (Eurocrypt98)*, LNCS 1403, pp. 112–126, Springer, 1998.

[5] N. Keller, E. Biham, "Cryptanalysis of reduced variants of rijndael," *unpublished manuscript*, 1999.

[6] N. Keller, J. Kim, J. Lu, O. Dunkelman, "New impossible differential attacks on AES," in *Indocrypt 2008*, LNCS 5365, pp. 279–293, Springer, 2008.

[7] K. Kim, J. Y. Lee, S. Kang, J. Cheon, M. Kim, "Improved impossible differential cryptanalysis of rijndael and crypton," in *Proceedings of Information Security and Cryptology (ICISC'01)*, LNCS 2288, pp. 39–49, Springer, 2002.

[8] X. Lai, "On the design and security of block ciphers," *Doctoral Theses, Hartung-Gorre*, 1992.

[9] S. Lucks, "Attacking seven rounds of rijndael under 192-bit and 256-bit keys," in *Proceedings of the Third AES Candidate Conference (AES3)*, pp. 215–229, 2000.

[10] S. Lucks, B. Schneier, M. Stay, D. Wagner, D. Whiting, N. Ferguson, J. Kelsey, "Improved cryptanalysis of rijndael," in *Proceedings of Fast Software Encryption 7*, LNCS 1978, pp. 213–230, Springer, 2001.

[11] J. L. Massey, X. Lai, "A proposal for a new block encryption standard," in *Advances in Cryptology (Eurocrypt90)*, LNCS 473, pp. 389–404, Springer, 1990.

[12] W. Meier, "On the security of the idea block cipher," in *Advances in Cryptology (Eurocrypt93)*, LNCS 765, pp. 371–385, Springer, 1994.

[13] M. Minier, H. Gilbert, "A collision attack on 7 rounds of rijndael," in *Proceedings of the Third AES Candidate Conference (AES3)*, pp. 230–241, 2000.

[14] S. Murphy, X. Lai, J. L. Massey, "Markov ciphers and differential cryptanalysis," in *Advances in Cryptology (Eurocrypt91)*, LNCS 547, pp. 17–38, Springer, 1991.

[15] National Institute of Standards and Technology. Announcing the Advanced Encryption Standard (AES), "Federal information processing standards publication 197," 2001. (http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf)
[16] R. Ch W. Phan, "Impossible differential cryptanalysis of 7-round advanced encryption standard (AES)," *Information Processing Letters*, vol. 91, no. 1, pp. 33–38, 2004.
[17] B. Preneel, J. Kim, S. Hong, "Related-key rectangle attacks on reduced AES-192 and AES-256," in *Proceedings of Fast Software Encryption 14*, LNCS 4593, pp. 225–241, Springer, 2007.
[18] B. Preneel, J. Vandewalle, Y. Kim, J. Nakahara, P. S. L. M. Barreto, "Square attacks on reduced-round pes and idea block ciphers," in *23rd Symposium on Information Theory*, pp. 187–195, 2002.
[19] A. M. Reza, B. Bahrak, "A novel impossible differential cryptanalysis of AES," in *Proceedings of the Western European Workshop on Research in Cryptology 2007*, 2007.
[20] A. M. Reza, B. Bahrak, "Impossible differential attack on seven-round AES-128," *IET Information Security Journal*, vol. 2, no. 2, pp. 28–32, 2008.
[21] V. Rijmen, J. Borst, L. Knudsen, "Two attacks on reduced idea (extended abstract)," in *Advances in Cryptology (Eurocrypt97)*, LNCS 1233, pp. 1–13, Springer, 1997.
[22] V. Rijmen, J. Daeman, "AES proposal: Rijndael, version 2," 1999. (http://csrc.nist.gov/archive/aes/rijndael/Rijndael-ammended.pdf)
[23] V. Rijmen, L. R. Knudsen, "Truncated differentials of idea," *Department of Electrical Engineering, ESATCOSIC Technical Report 97/1*.
[24] A. Selcuk, H. Demirci, "A meet-in-the-middle attack on 8-round AES," in *Proceedings of Fast Software Encryption 15*, LNCS 5806, pp. 116–126, Springer, 2008.
[25] A. Shamir, E. Biham, A. Biryukov, "Miss-in-the-middle attacks on idea, khufu and khafre," in *6th Fast Software Encryption Workshop*, LNCS 1636, pp. 124–138, Springer, 1999.
[26] G. N. Tuychiev, "About networks idea16-4, idea16-2, idea16-1, created on the basis of network idea16-8," *Compilation of theses and reports republican seminar Information security in the sphere communication and information. Problems and their solutions*, 2014.
[27] G. N. Tuychiev, "About networks idea328, idea324, idea322, idea321, created on the basis of network idea3216," *Infocommunications: NetworksTechnologiesSolutions*, vol. 30, no. 2, pp. 45–50, 2014.
[28] G. N. Tuychiev, "About networks idea8-2, idea8-1 and rfwkidea8-4, rfwkidea8-2, rfwkidea8-1 developed on the basis of network idea8-4," *Uzbek mathematical journal*, no. 3, pp. 104–118, 2014.
[29] G. N. Tuychiev, "About networks pes8-2 and pes8-1, developed on the basis of network pes8-4," *Transactions of the international scientific conference Modern problems of applied mathematics and information technologies-Al-Khorezmiy 2014*, vol. 2, pp. 28–32, 2014.
[30] G. N. Tuychiev, "About networks rfwkpes8-4, rfwkpes8–2, rfwkpes8-1, developed on the basis of network pes8-4," *Transactions of the international scientific conference Modern problems of applied mathematics and information technologies-Al-Khorezmiy 2014* , vol. 2, pp. 32–36, 2014.
[31] G. N. Tuychiev, "New encryption algorithm based on network rfwkpes8-1 using of the transformations of the encryption algorithm AES," *International Journal of Multidisciplinary in Cryptology and Information Security*, vol. 3, no. 6, pp. 31–34, 2014.
[32] G. N. Tuychiev, "New encryption algorithm based on network idea16-1 using of the transformation of the encryption algorithm AES," *IPASJ International Journal of Information Technology*, vol. 3, pp. 6–12, 2015.
[33] G. N. Tuychiev, "New encryption algorithm based on network idea8-1 using of the transformation of the encryption algorithm AES," *IPASJ International Journal of Computer Science*, vol. 3, pp. 1–6, 2015.
[34] G. N. Tuychiev, "New encryption algorithm based on network pes8-1 using of the transformations of the encryption algorithm AES," *International Journal of Multidisciplinary in Cryptology and Information Security*, vol. 4, no. 1, pp. 1–5, 2015.
[35] G. N. Tuychiev, "New encryption algorithm based on network rfwkidea8-1 using transformation of AES encryption algorithm," *International Journal of Computer Networks and Communications Security*, vol. 2, no. 3, pp. 43–47, 2015.
[36] G. N. Tuychiev, "To the networks rfwkidea3216, rfwkidea328, rfwkidea324, rfwkidea322 and rfwkidea321, based on the network idea3216," *International Journal on Cryptography and Information Security (IJCIS)*, vol. 5, no. 1, pp. 9–20, 2015.
[37] U. S. Department of Commerce/National Institute of Standards and Technology. Data Encryption Standard (DES), "Federal information processing standards publication 46-3," 1979. (http://csrc.nist.gov/publications/fips/fips46-3/fips46-3.pdf)
[38] J. Vandewalle, J. Daemen, R. Govaerts, "Cryptanalysis of 2.5 rounds of IDEA (extended abstract)," *Department of Electrical Engineering, ESATCOSIC Technical Report 93/1*, pp. 1–6, 1993.
[39] D. Wagner, J. Kelsey, B. Schneier, "Key-schedule cryptanalysis of idea, g-des,gost, safer and triple-des," in *Advances in Cryptology (Crypto96)*, LNCS 1109, pp. 237–251, Springer, 1996.

[40] Y. Wei, J. Chen, Y. Hu, "A new method for impossib le differential cryptanalysis of 8-round adanced encryption standard," *Wuhan Univeristy Journan of National Sciences*, vol. 11, no. 6, pp. 1559–1562, 2006.
[41] Y. Wei, J. Chen, Y. Hu, "A new method for impossible differential cryptanalysis of 7-round advanced encryption standard," in *Proceedings of IEEE International Conference on Communications, Circuits and Systems Proceedings 2006*, vol. 3, pp. 1577–1579, 2006.
[42] L. Zhang, D. Feng, W. Zhang, W. Wu, "Improved related-key impossible differential attacks on reduced-round AES-192," in *Proceedings of Selected Areas in Cryptography 2006*, LNCS 4356, pp. 15–27, Springer, 2007.
[43] Y. Zhang, J. Chen, Y. Hu, "Impossible differential cryptanalysis of advanced encryption standard," *Science in China Series F: Information Sciences*, vol. 50, no. 3, pp. 342–350, 2007.

Aripov Mersaid Doctor of Phys. Math. Science, Professor of National University of Uzbekistan.

Tuychiev Gulom candidate technical Sciences (Ph.D.), National University of Uzbekistan.

Defense Against Selfish PUEA in Cognitive Radio Networks Based on Hash Message Authentication Code

Walid Ramadan Ghanem, Mona Shokair, and Moawad I. Dessouky
(Corresponding author: Walid Ramadan Ghanem)

The Department of Electronic and Electrical Communication, Menoufia University, Egypt
(Email: walid.ghanem.eng@ieee.org)
(Received Dec. 11, 2015; revised and accepted Dec. 26, 2015)

Abstract

Cognitive radio network (CRN) is a very astute technology developed to solve the spectrum shortage problem in wireless communication by utilizing the unused bands, where a secondary user (SU) utilizes the free spectrum of the primary user (PU) in an opportunistic manner. The CRN defines the free spectrum sessions using an intelligent and sophisticated process called a spectrum sensing. The spectrum sensing encounters a security problem called primary user emulation attack (PUEA). In this problem, an attacker mimics the PU signal to force the SUs to leave the free band. In this paper, a proposed model based on hash message authentication code (HMAC) is used to detect the PUEA in CRN. HMAC is used to trusting the PU transmission, which is not clarified until now. A shared secret key is used between the SU and the PU to achieve an accurate identification of the PU signal from the attacker. The effectiveness of the proposed approach is analyzed through both theoretical analysis and Simulation. Results show that the proposed method is completely defeated the selfish PUEA and achieves efficient spectrum sharing, moreover, it provides a good detection of the PU when error correcting codes are used.

Keywords: Cognitive radio network (CRN), hash message authentication code (HMAC), physical layer authentication, primary user emulation Attack (PUEA)

1 Introduction

Cognitive radio (CR) has attracted a strong attention recently to solve the spectrum shortage problem [19]. Spectrum scarcity becomes a serious challenge to the emerging wireless technologies. In licensed networks, the primary users operate in their allocated licensed bands. It is recognized that the licensed bands are generally underutilized and their occupation fluctuates temporary and geographically in the range of 15-85%. In a typical Cognitive Radio (CR) system, the PU is the spectrum license holder and the SU is an unlicensed user who intends to use the spectrum opportunistically. CR is based on Dynamic Spectrum Access (DSA), where the Priority is given to the PU in the sense that SU can only transmit, if its transmission is deemed to be a harmless to the PU. The SU is not allowed to transmit when the PU is transmitting [1] . Unused bands (White spaces) are identified through spectrum sensing process [24], then they are used by the SUs for data transmission, the spectrum sensing process is continuously performed to determine the white spaces, if the primary user is detected, then the SU must vacate the band for him. The CR networks are exposed to many attacks one of these attacks is called PUEA. The PUEA is considered as a physical layer attack, where a selfish user (attacker) mimics the PU signal to confuse the SUs to leave the band for him. This leads to low spectrum utilization in CRN.

PUEA has been studied in many researches. R. Chen proposed to use the location of the primary user to identify the PUEA in [9] . S. Annand made an analytical model based on Fenton's approximation and Markov inequality in [4]. Z. Jin et al. Presented a NeymanPearson composite hypothesis test in [15]. and a Wald's sequential probability ratio test to detect the PUEA was described in [16]. Z. Chen showed how the attacker can emulate the PU signal to confuse the SU and used an advanced strategy called variance detection to mitigate the effect of an attacker using the difference between the communication channel of PUEA and PU in [10]. C. Chen et al made a joint position verification method to enhance the positioning accuracy in [14] . Cooperative Spectrum Sensing in CRN in the Presence of the PUEA is proposed in [8] . Feign Bao et al studied the PUEA with national secondary users in CRN

and using a hybrid method based on Energy Detection (ED) and Variance Detection in [5] . Kapil M. Borle has developed a physical layer authentication scheme for wireless communication in [7]. Advanced encryption standard was used to mitigate PUEA in [2]. More researches discussed this problem will be found in [3, 6, 12, 13, 21, 22].

In this paper, a proposed Physical layer authentication based on hash message authentication code (HMAC) is used to detect the PUEA in cognitive radio networks which is not clarified until now. Moreover, different coding techniques will be applied. The HMAC is used to generate a tag at the transmitter, this tag is appended to the message and sent over the channel. At the receiver the secondary user separates the message and tag and regenerates a new tag from the shared key and the received message. By comparing the two tags, the SU determines if the signal comes from the PU or from the attacker. The proposed method provide a good detection of the PU and it is completely defeated the PUEA under any condition and ables a high spectrum usage and high efficiency. The performance of the system is measured using the probability of false alarm and probability of detection.

The rest of this paper is organized as follows In Section 2, the problem formulation will be explained. In Section 3, the system model of the system is introduced. In Section 4 the performance metrics will be presented. In Section 5, the analytical Model Evolution will be explained. In Section 6 numerical and simulation results will be included. Finally conclusions will be drawn in Section 7

2 Problem Formulation

The PUEA is considered one of the main threats in CRNs. In this problem, an attacker emulates the PU signal to confuse the SU to leave the free spectrum session as shown in Figure 1. The attacker receives the signal from the primary user and emulates it and retransmits the signal again to the SUs. The SUs suppose that the transmission comes from the PU and therefore, they leave the free spectrum. The PUEA destroys the spectrum sensing process. In Figure 1, there are three users as follow:

- **Primary User**: A user who has higher priority or legacy rights for the usage of a specific part of the Spectrum.
- **Secondary User**: A user who has a lower priority and therefore exploits the spectrum in such a way that it does not cause any interference to PUs.
- **Selfish PUEA**: the aim of this attacker is to maximize the spectrum usage for himself, by taking the free band and preventing the others SUs from using it.

This attacker can mimic the primary user power, modulation, signal characteristics and any characteristics of the PU signal. Thus the detection of the attacker becomes extremely difficult. The authentication is a proper method to solve this problem. By authenticating the transmission between the PU and the SU, this problem will be solved easily.

3 System Model

In this section, the system model will be described, then the block diagram of the proposed system, which includes the attackers will be explained. Finally the flow chart of the system will be done as shown in figure 2. The PU message runs into the HMAC algorithm to produce the first HMAC (TAG A) [23], then this tag is padded with the message using TDMA, both of them are encoded, modulated and transmitted to the SUs as shown in figure .2. The SU intern runs the message portion of the transmission through the HMAC algorithm using the same key that was used by the PU, producing a second HMAC (TAG B), The SU compares the two tags. If they are identical the transmission is assumed to be from the PU and the SU must stop its transmission, otherwise a PUEA is in progress. The attacker don not have the same key used by the PU and assumed to be not intelligent enough to extract the key, therefore, he cannot authenticate his transmission. If the attacker uses another key, the SU defines him correctly.

In Figure 3, the flow chart of the proposed model is described as follows, first, a random data represents the PU message is generated, this data is applied to the HMAC function to generate the TAG A. The TAG A and the message are added together using TDMA, then the message is encoded, modulated and transmitted on the channel. The SU receives the data demodulates and decodes it. Then SU separates the TAG from the message, and applies the message part to HMAC to produces a second tag (TAG B). And compares the two tags, if the two tags are identical, then the data is considered from the PU, otherwise, the attacker is in progress and the SU must punish him.

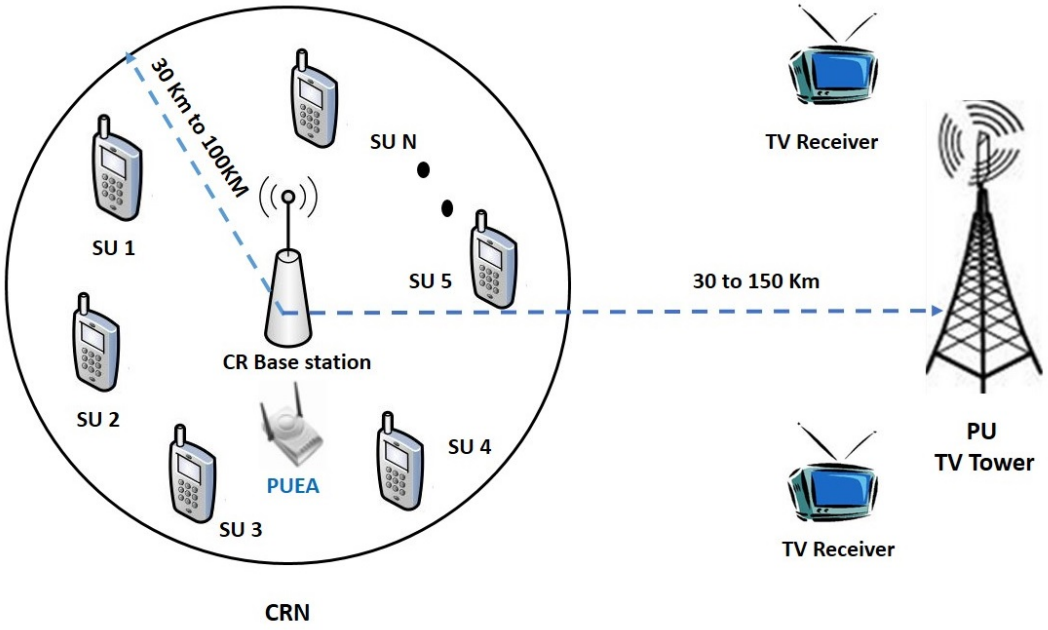

Figure 1: The basic concept of PUEA in CRN

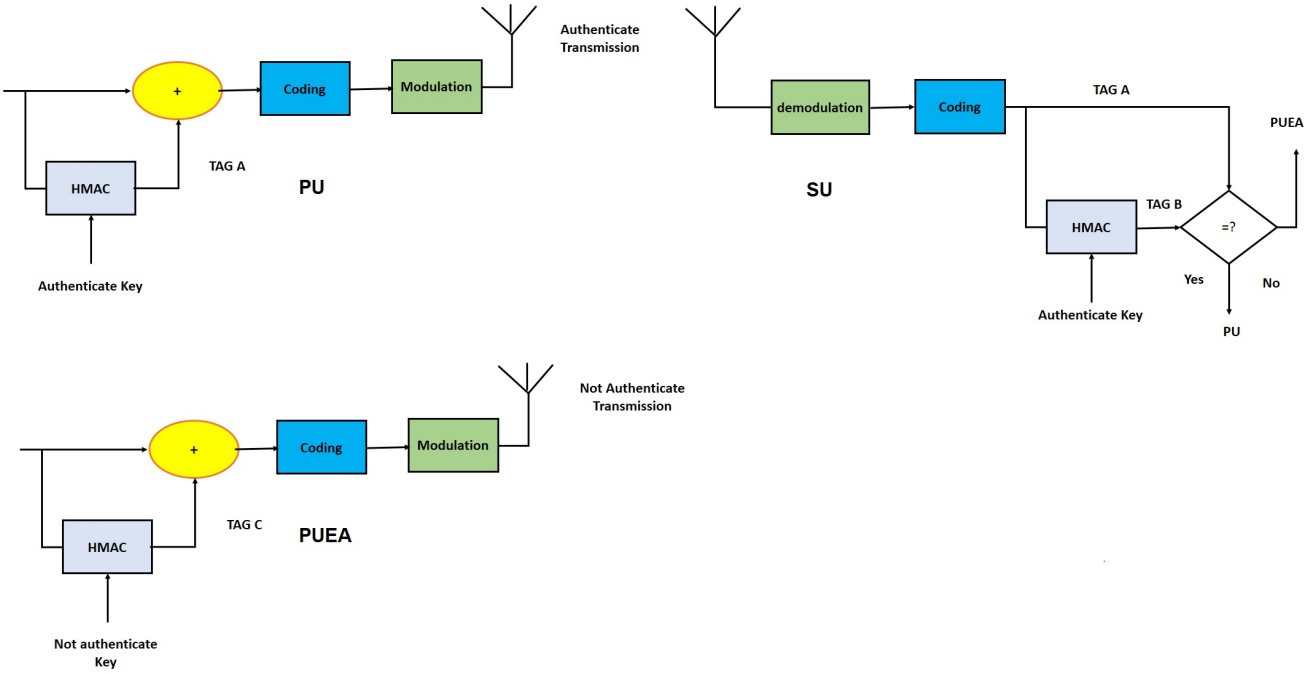

Figure 2: Block diagram of the proposed system

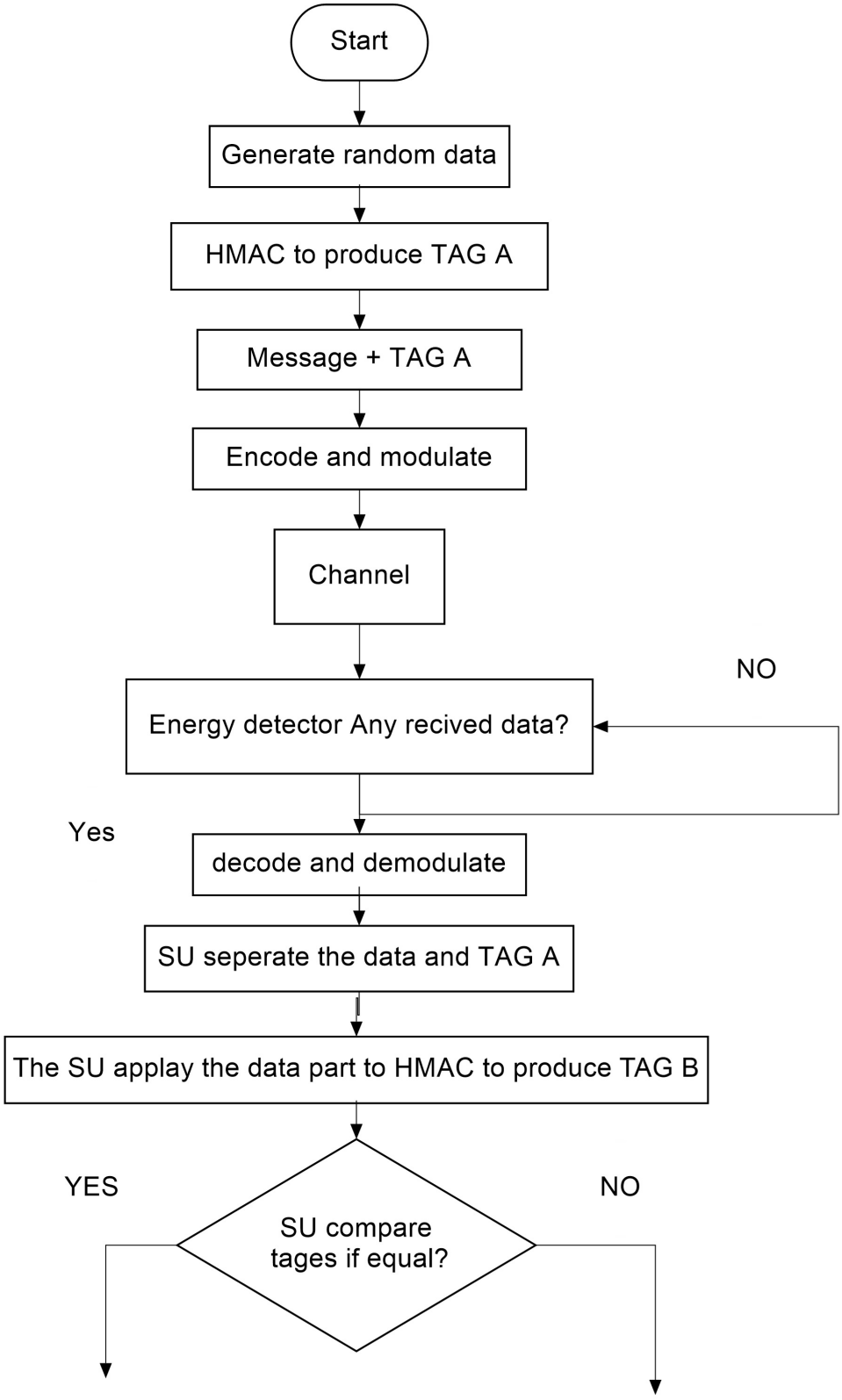

Figure 3: Flow chart of the proposed model

4 Performance Matrices

In this section, the system performance is analyzed by using false alarm probability and detection probability [10]. Most existing works on cognitive sensing focused on performing a hypothesis testing to decide the presence of the primary user. H0: the signal is from the primary user H1: the signal is from the attacker

- **Probability of false alarm (PFA)**: When the signal is considered from the primary user, the probability that the SU falsely Identify the signal as from the attacker is defined by [10]:

$$P_{FA} = Pr(H_1 \backslash H_0). \tag{1}$$

  If this case happens, the SU will attempt to access the network and cause interference to the PU. Then the SU may be punished as an attacker. Hence the SU may use a strategy to make PFA as small as possible while the attacker wants to make PFA as large as possible.

- **Probability of Primary User detection (PDP)**: When the signal is deeded from the PU, the probability that the SU classifies it as from the primary use,

$$P_{DP} = Pr(H_0 \backslash H_0) = 1 - P_{FA}. \tag{2}$$

  If this case happens, the SU will attempt to access the network and cause interference to the PU. Then the SU may be punished as an attacker. Hence the SU may use a strategy to make PFA as small as possible while the attacker wants to make PFA as large as possible.

- **Probability of Misdetection (PMD)**: When the signal is counted from the attacker, the probability that the SU classifies falsely it as from the primary user is detonated by,

$$P_{MD} = Pr(H_0 \backslash H_1). \tag{3}$$

  If this happens, the victim will give up accessing the network, although the spectrum band is vacant, and the attacker launches a successful PUEA and takes the spectrum resources. Another widely matrices is the probability of detection of the attacker PDA.

$$P_{DA} = Pr(H_1 \backslash H_1) = 1 - P_{MD}. \tag{4}$$

  The SU should take a strategy to make the PDA as large as possible.

5 Analytical Model Evolution

The authentication tag is formed using a Key based HMAC, even a single bit in the tag or the message will destroy the authentication between the SU and the PU. The probability of false alarm is calculated analytically for the uuencoded BPSK under AWGN.

The Probability of False Alarm for Uuencoded BPSK Under AWGN

The probability of false alarm is related to the probability of one bit occurs in the tag or the data. Assume the channel is a binary symmetric channel with error probability equal to P, the bit error occurs independently. Hence, the probability of m errors in a block of n bits is given by [20].

$$P(m,n) = \binom{n}{k} P^m (1-P)^{n-m} \tag{5}$$

The probability of false alarm is measured by calculating the probability of one error occur and more, therefore, the probability of false alarm is given by

$$P_{FA} = Pr(atleastonebiterror) \tag{6}$$

$$= \sum_{m=1}^{n} \binom{n}{k} P^m (1-P)^{n-m} \tag{7}$$

The channel error rate of BPSK under AWGN is given by [20]:

$$P = 0.5erfc(\sqrt{\frac{E_b}{N_o}}). \tag{8}$$

The probability of false alarm of uuencoded BPSK is done by put Equation (8) into Equation (6):

$$P_m = \sum_{m=1}^{n} \binom{n}{k} (0.5erfc(\sqrt{\frac{E_b}{N_o}})^m)(1 - (0.5erfc(\sqrt{\frac{E_b}{N_o}})))^{n-m} \tag{9}$$

For Coded BPSK Using a Linear Block Coding (BCH Code)

$$P_{CW}^{tag} \leq \sum_{i=t^{tag}+1}^{n} \binom{n^{tag}}{i} P^i (1-P)^{n-i} \tag{10}$$

$$P_b^{tag} \leq \frac{1}{n^{tag}} \sum_{i=t^{tag}+1}^{n} i \binom{n^{tag}}{i} P^i (1-P)^{n-i} \tag{11}$$

For simplicity equally will be used the probability of false alarm is given by:

$$\begin{aligned} P_{FA} &= 1 - (1 - P_{CW}^{tag})^{\frac{L}{K^{tag}}} \\ &= 1 - (1 - \sum_{i=t^{tag}+1}^{n} \binom{n^{tag}}{i} * P^i (1-P)^{n-i})^{\frac{L}{K^{tag}}} \end{aligned} \tag{12}$$

The probability of False Alarm of BPSK under AWGN using block code is done by substitute Equation (6) into Equation (12):

$$P_{FA} = 1 - (1 - \sum_{i=t^{tag}+1}^{n} \binom{n^{tag}}{i} (0.5erfc(\sqrt{\frac{E_b}{N_o}}))^m (1 - (0.5erfc(\sqrt{\frac{E_b}{N_o}}))^{n-i})^{\frac{L}{K^{tag}}} \tag{13}$$

The overall probability of detection of the primary user of BPSK under AWGN when error correcting block code block code is done by:

$$P_{DP} = (1 - \sum_{i=t^{tag}+1}^{n} \binom{n^{tag}}{i} (0.5erfc(\sqrt{\frac{E_b}{N_o}})))^m (1 - (0.5erfc(\sqrt{\frac{E_b}{N_o}}))^{n-i})^{\frac{L}{K^{tag}}} \tag{14}$$

More about authentication and encryption will be found in [11, 17, 18].

6 Simulation And Numerical Results

In this section, the effectiveness of the HMAC authentication method will be validated through analytical and simulation. The simulation parameters of the system will be tabulated in table 1. First the data is generated by the PU as a random data of length 18bytes, this data is prepared first to apply as the input of the HMAC function to produce an output with a length of 20 bytes by taking the left 18 bytes in the output to produce TAG A. The TAG A is appended to the message using TDMA algorithm, then encoded, modulated and sent both of them on the channel. The SU separates the message and TAG A. The received message is applied to HMAC to produce TAG B. The SU compares the two tags. If the two tags are identical, then the transmission is represented by PU, otherwise is done by the attacker. Mont Carol simulation is used, for every packet we run 10000 times and calculate the probability of the false alarm and detection.

In Figure 4 the probability of false alarm varies by changing the SNR dB using BPSK under AWGN channel model. As the SNR increases the probability of the PFA decreases. The simulation and the analytical solution gives the same results. It is also show that, using a good error correcting code decreases the PFA. At SNR=2dB, the PFA=100% for the uuencoded BPSK, when using BCH (7, 4, 1) the PFA is decrease by 15%, when using BCH (31, 16, 3) the PFA is decreased by 60%.

In Figure 5, the probability of false alarm varies by changing the SNR dB under AWGN channel model when the error correcting code is a reed solomn code. As the SNR increases the probability of the PFA is decreases. At

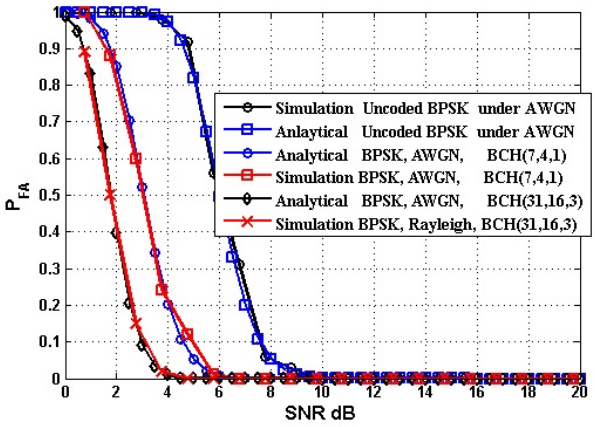

Figure 4: The PFA versus the SNR for a different BCH codes

Table 1: Simulation parameters

Simulation parameters	N=1000
Data length of the PU	16bytes
HMAC code length	16 bytes
Total PU data	36 bytes
Channel	AWGN
Modulation type	BPSK
S/N dB	0:20 dB
Encoding types	BCH, Reed solomn
Hash method used	SHA-1

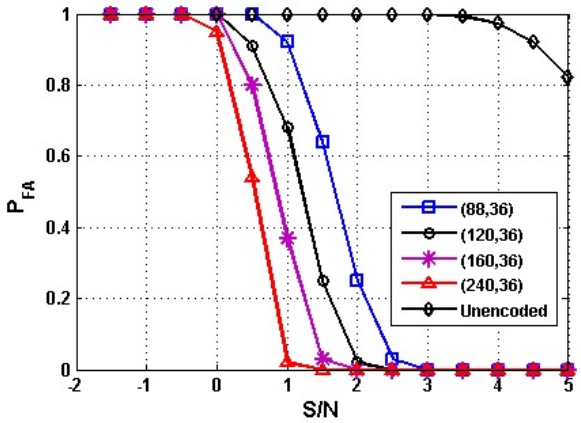

Figure 5: PFA V.s SNR for Uuencoded and encoded using RS (n, k) codes

SNR=1dB the PFA=100% for the uuencoded, when using RS (88, 36), RS (120, 36), RS (160, 36) and RS (240, 36) the PFA is decrease by 10%, 30 %, 60% and 98% respectively.

In Figure 6 the probability of detection (PDA) is varied according to SNR dB. The SU detect the attacker at each value of the SNR and achieve 100% detection of PUEA under all cases. In this simulation the attacker tries many keys to confuse the SU but he cannot. The two curves are identical.

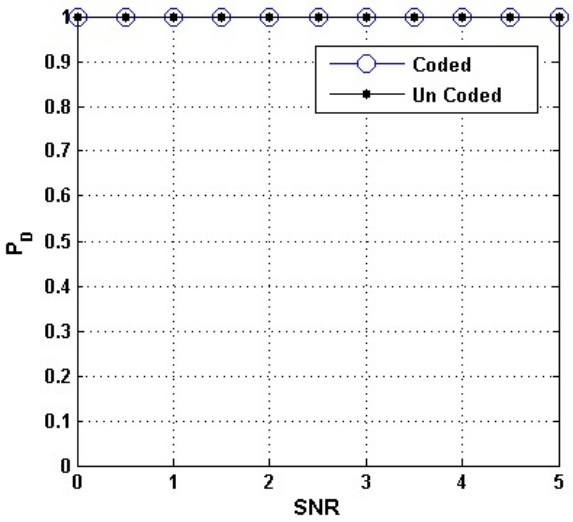

Figure 6: PDA of BPSK under AWGN

7 Conclusions

In this paper, a new method is proposed to solve the primary user emulation attack in cognitive radio networks. In this model, the primary user uses a hash message authentication code (HMAC) to authenticate its transmission. The new method helps the secondary user to defines the attackers, by appending a tag in the transmission and sends the tag and the message on the channel. At the receiver the SU separates the tag from the message and a new tag is generated from the received message and the shard key. If the two tags are the same, then the transmission is considered from the PU, otherwise, the PUEA is in progress. The simulation results show that the new method completely defeats the PUEA and provides good detection of the primary user. In future, more error correcting codes and channel models will be used.

Acknowledgments

The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "Nextgeneration/dynamic spectrum access/cognitive radio wireless networks," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[2] A. Alahmadi, M. Abdelhakim, J. Ren, and T. Li, "Defense against primary user emulation attacks in cognitive radio networks using advanced encryption standard," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 5, pp. 772–781, 2014.

[3] A. Alahmadi, T. Song, and T. Li, "Sub-band detection of primary user emulation attacks in ofdm-based cognitive radio networks," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP'14)*, pp. 1165–1169, Atlanta-GA, Dec. 2014.

[4] S. Anand, Z. Jin, and K. P. Subbalakshmi, "An analytical model for primary user emulation attacks in cognitive radio networks," in *3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN'08)*, pp. 1–6, Chicago, USA, Oct. 2008.

[5] F. Bao, H. Chen, and L. Xie, "Analysis of primary user emulation attack with motional secondary users in cognitive radio networks," in *2012 IEEE 23rd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC'12)*, pp. 956–961, Sydney-NSW, Sept. 2012.
[6] J. Blesa, E. Romero, A. Rozas, and A. Araujo, "Pue attack detection in cwsns using anomaly detecttion techniques.," *EURASIP Journal on Wireless Communications and Networking,*, vol. 2013, no. 1, pp. 1154–1164, 2013.
[7] K. Borle, B. Chen, and W. Du, "A physical layer authentication scheme for countering primary user emulation attacks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'13)*, pp. 2935–2939, Vancouver, BC, May 2013.
[8] C. Chen, H. Cheng, and Y. Yao, "Cooperative spectrum sensing in cognitive radio networks in the presence of the primary user emulation attack," *IEEE Transactions on Wireless Communication,*, vol. 10, no. 7, pp. 2135–2141, 2011.
[9] R. Chen, J. Park, and J. Reed, "Ensuring trustworthy spectrum sensing in cognitive radio networks," in *Proceedings of the IEEE Workshop Network Technology Software Defined Radio Networks(2006)*, pp. 110–119, Reston, USA, sept 2006.
[10] Z. Chen, T. Cooklev, C. Chen, and C. Plmalaza-Raez, "Modeling primary user emulation attacks and defenses in cognitive radio networks," in *IEEE International Conference on Communications (ICC'09)*, pp. 208–215, Scottsdale, AZ, Dec 2009.
[11] Y. Dou, W. Jiang, and C. Ma, "Improved fault attack against eta pairing," *International Journal of Network Security*, vol. 16, no. 1, pp. 71–77, 2014.
[12] W. R. Ghanem, M. Shokir, and MI Dessouky, "Investigation of puea in cognitive radio networks using energy detection in different channel model," *Circuits and Systems: An International Journal*, vol. 2, no. 2/3/4, pp. 1–11, 2015.
[13] M. Haghighat and S. M. S. Sadough, "Smart primary user emulation in cognitive radio networks: defence strategies against radioaware attacks and robust spectrum sensing.," *Transactions on Emerging Telecommunications Technologies,*, vol. 26, no. 9, pp. 1154–1164, 2015.
[14] L. Huang, L. Xie, H. Yu, W. Wang, and Y. Yao, "Anti-pue attack based on joint position verification in cognitive radio networks," in *2010 International Conference on Communications and Mobile Computing (CMC'10),* , pp. 169–173, Shenzhen, April 2010.
[15] Z. Jin, S. Anand, and K. P. Subbalakshmi, "Mitigating primary user emulation attacks in dynamic spectrum access networks using hypothesis testing," in *IEEE International Conference on Communications (ICC'09)*, pp. 1–5, Dresden, USA, June 2009.
[16] Z. Jin and K. P. Subbalakshmi, "Detecting primary user emulation attacks in dynamic spectrum access networks," in *IEEE International Conference on Communications (ICC'09)*, pp. 1–5, Dresden, USA, June 2009.
[17] C. C. Lee, C. H. Liu, and M. S. Hwang, "Guessing attacks on strong-password authentication protocol.," *International Journal of Network Security*, vol. 15, no. 1, pp. 64–67, 2013.
[18] C. Lin, Y. Lv K. Li, and C. C. Chang, "Ciphertext-auditable identity-based encryption.," *International Journal of Network Security*, vol. 17, no. 1, pp. 23–28, 2015.
[19] J. Mitola and G. Q. Maguire, "Cognitive radios: Making software radios more personal," *IEEE Personal Communication*, vol. 6, no. 4, pp. 13–18, 1999.
[20] J. G. Proakis and M. Salehi, *Digital Communications (5ed), English.* USA: McGraw-Hill Education, 2007.
[21] M. J. Saber and S. M. S. Sadough, "Optimal soft combination for multiple antenna energy detection under primary user emulation attacks," *AEU-International Journal of Electronics and Communications,*, vol. 69, no. 9, pp. 1181–1188, 2015.
[22] A. A. Sharifi, Morteza Sharifi, and Mir Javad Musevi Niya, "Secure cooperative spectrum sensing under primary user emulation attack in cognitive radio networks: Attack-aware threshold selection approach," *AEU-International Journal of Electronics and Communications,*, vol. 7, no. 2/3/4, pp. 95–104, 2016.
[23] W. Stallings, *Cryptography and Network Security: Principles and Practice (5ed).* USA: Prentice-Hall, 2010.
[24] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, 1967.

Walid R Ghanem received the B.Sc. degree in communication engineering from Electronic Engineering, Menoufia University, Egypt, in May 2011, and he is currently working toward the MSc degree in Electrical communication engineering. His current research interests are cognitive radio networks, Localization, wireless security, encryption and optimization algorithms.

Mona Shokair received the B.Sc., and M.Sc. degrees in electronics engineering from Menoufia University, Menoufia, Egypt, in 1993, and 1997, respectively. She received the Ph.D. degree from Kyushu University, Japan, in 2005. She received VTS chapter IEEE award from Japan, in 2003. She published about 70 papers until 2014. She received

the Associated Professor degree in 2011. Presently, she is an Associated Professor at Menoufia University. Her research interests include adaptive array antennas, CDMA system, WIMAX system, OFDM system, game theory, next generation networks and optimization algorithms.

Moawad I. Dessouky received the B.Sc. (Honors) and M.Sc. degrees from the Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt, in 1976 and 1981, respectively, and the Ph.D. from McMaster University, Canada, in 1986. He joined the teaching staff of the Department of Electronics and Electrical Communications, Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt, in 1986. He has published more than 200 scientific papers in national and International conference proceedings and journals. He has received the most cited paper award from Digital Signal Processing journal for 2008. His current research areas of interest include spectral estimation techniques, image enhancement, image restoration, super resolution reconstruction of images, satellite communications, and spread spectrum techniques.

A New Efficient Files Retrieval System Using Caching Search Engine

Mohammed Khaleel Hussein[1,2], Hazem M. El-Bakry[2], Ahmed A. Saleh[2]

(Corresponding author: Mohammed Khaleel)

Ministry of Higher Education and Scientific Research, Scholarships and Cultural Relations Directorate, Iraq[1]
Information Systems Department, Faculty of Computers and Information, Mansoura University, Egypt[2]
(Email: mkhaleel777@yahoo.com)

(Received Dec. 17, 2015; revised and accepted Jan. 7, 2016)

Abstract

Profitable cloud file search systems have to route client queries over massive Web indexes under certain stretched limitations. In training phase, to get low latency, massive outcomes are working and a segment of the query transfer is provided using earlier computed outcomes. In addition, file search systems are required to revise their indexes regularly in order to integrate modifications to the cloud. After each key change in database, the signification of cache entries may turn into old, thus reducing the newness of served outcomes. In this paper, a proposed framework which called Fast File Retrieval System (FFRS) is proposed. Such framework manipulates problems in caching. This is to achieve scalability in file explore systems and the ability to handle the changes in index. Furthermore, a new method that uses a (TTL) time-to-live amount and frequently used value is introduced. The idea is to set cache entrance to select revives cached outcomes by concerning revive queries to back-end search groups from database. Then sorting the entrances to be identified is performed according to the number of admittance with the age of a record in the cache. Moreover, to set the rate at which revive demands are mattered, a novel method that takes inactive cycles of execution servers is presented. Assessment using real data illustrates that the proposed framework can realize hit rate enhancement efficiently.

Keywords: Crawler, fast file retrieval system, file caching, file search engine

1 Introduction

Data centers, global enterprises, and distributed cloud storage all require sharing of huge amounts of data in a consistent, efficient, and reliable manner across a wide-area network called Internet. The two emerging trends of offloading data to a distributed storage cloud and using the Map-Reduce algorithm for building highly parallel data-intensive functions [5]. A massive amount of data is available in Internet, which is only helpful and useful if it can be accessed in its file. To access files from the Internet, we need fast and accurate search facility such as search engine.

Searchable web grows to be bigger and bigger, with further than 20 billion pages and over 100 billion files to index, handling a single query requires assessing great amounts of data. In such a setting, to realize a rapid response time and to enhance the query throughput, using a cache is critical. The main use of a cache is to accelerate calculation by utilizing recurrently or freshly used data, although reducing the workload to back-end servers is also a major goal.

Caching can be appropriated at dissimilar levels with growing response time or processing constraints. Caching of search results has long been identified as an essential optimization step in search systems [14]. To specify the high evaluated quantity of user demands, outcome caches materialize as critical functioning elements to decrease the query transfer to servers and also to decrease the standard demand execution time.

The engine must make a decision whether to reconsider frequent queries, thereby decreasing the efficiency of caching, or to save calculations at the risk of returning old results. Existing resolutions are as simple as preceding caching completely and concerning (TTL) time-to-live policies on cached entries to ensure worst-case moldiness bounds. A familiar surveillance from the information recovery is that query frequencies follow a power-law sharing [27].

With the evaluation of the usual trouble of caching in operating engines, the trouble of outcome caching in explore systems is not size of memory. For explore systems, it is achievable to store massive records on disk and yet get better query time. On common, it obtains tens of milliseconds to develop a query on a explore category, and obtaining from disk in similar response time, often lower. Additionally, using disks to accumulate earlier calculated outcomes

gives an occasion to reduce the capability of misses observed in small outcome caches that utilize only memory of RAM. One main weakness of huge outcome caches is newness. Explore system indexes modifies recurrently because of fresh batches of spidered articles. Therefore, it is probable that a significant portion of earlier computed outcomes in the cache become decayed in time, i.e., some of the top-matching outcomes in the existing key are not exist in cached records, thus may be corrupted the value of outcomes. In reality, we disagree that the newness dilemma grow to be more strict as the cache capability raises.

In this manuscript, we produces the plan of the cache outcomes applied in the Yahoo system. This cache establishes the perception of stimulating cache records and produces a realistic method for preferring records to refresh. To the finest of our information, we are the initial to believe the dilemma of refreshing outcome records in explore system caches. Moreover, our subscriptions are:

" Suggest a method for terminating cache records appropriate to a (TTL) time-to-live significance and a method for preserving the new and clean signification by concerning revive demands to back-end explore groups, according on availability of inactive rounds in those groups. " Suggest a method for preferring cache records to be revived based on the admission occurrence of records and the age of the cached record. " We estimate the functioning of our methods via 130 million demands. " We present some figures from suggested schemes, coverage on our experimental assistance in practice.

With the suggested methods, we are capable of obtaining two significant advantages in manufacture. Higher hit rates, which recover the typical reply time of the explore system. Decrease peak query traffic on back-end explore clusters, which reducing the hardware prices.

This paper prearranged as following: Section-2 illustrates related work. In Section-3, we review the scheme framework and proposed algorithms for this work. We produce several experimental results in Section-4. Section-5 concludes the paper. Finally, section-6 is the list of references.

2 Related Work

2.1 Basic Crawling Process

Crawling process is complex and need much time to complete searching in huge cloud databases. So, crawling process must be done in offline part of the system in order to complete complex tasks. Any crawling algorithm consists of four major stages [22]: a) Searching in URLs with some attributes in cloud databases, b) Download and Fetch the results, c) Results Filtration process with the remaining attributes, d) Caching and retrieving the final results. Figure 1. shows the major steps for crawling process. Figure 2 shows the global crawling Architecture

2.2 Web Crawling

Heydon and Najork [11] have provided information on several crawler modules and the options of the design and that was one of the early full explanations of a scalable Web crawler is that of Mercator. A distributed crawler was defined by Najork and Heydon according to Mercator [15]. Heritrix [9] is a modular open source and archival-quality crawler, established at the Internet Archive. The distributed Web crawler, UBICrawler, applied in Java that is working in a decentralized approach and using reliable hashing to divid the domains to crawl through the crawling servers, as revealed by Boldi et al. [20]. Lee et al. [12] define the main structures of the data and architecture of IRLBot, a crawler that implements DRUM (Disk Repository with Update Management) to check if a URL was seen previously. Using DRUM permits IRLBot to maintain the crawling rate high, even after crawling billions of Web pages.

When the Web changes and Web pages were deleted, modified, or created, [16], and in order to handle these changes, active crawling tactics are required. Cho and Garcia Molina [2] explain a gradual crawler in order to optimize the average freshness for the data of crawled Web. Olston and Pandey [17] define strategies for re-crawling for freshness optimizing according to the information longevity on Web pages. A parameterized algorithm to monitor the resources of a Web for optimizing completeness or timeliness, and updates based on application-specific needs was also introduced by Pandey and Olston [19].

2.3 Types of Web Crawler

1) Focused Web Crawler: Web crawler which attempts to download some pages which are relevant to each other called Focused Crawler [2]. It gathers specific documents that are related to the particular topic [13]. The focused crawler controls the following - Relevancy, Way forward. It regulates relationship of the certain page to the given topic and how to forward. Economically, focused web crawler is reasonable in terms of network and hardware resources, it is able to decrease the amount of downloads and network traffic, and that is the significance of it [10]. The search experience of focused web crawler is great as well [21].

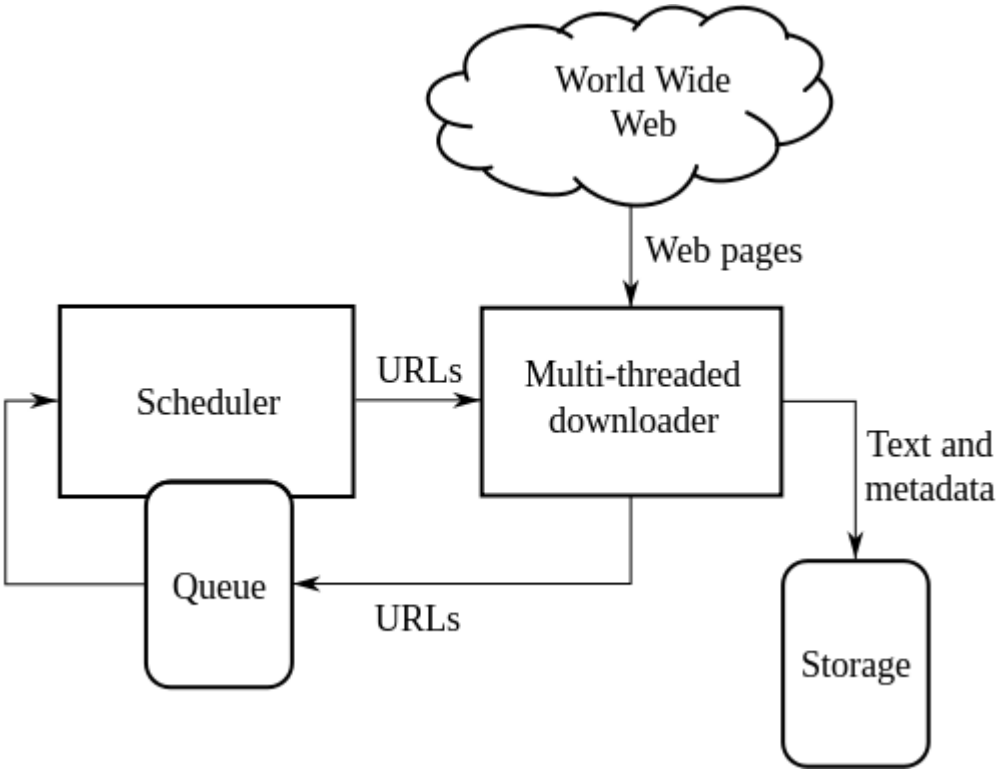

Figure 1: Major steps for crawling process

2) Gradual Crawler: A traditional crawler, for its collection to be refreshed, regularly replaces the old documents with the documents which were lately downloaded. On the opposite of that, a gradual crawler refreshes the present pages collection regularly via visiting them repeatedly; according to the evaluation as to how frequently pages are changed [2]. It also removers the less important pages and replaces them by new ones that are more important. It also provides a resolution the pages freshness problem. In the regular crawler, only valuable data is delivered to the user, and that is the significance of it. Therefore, data improvement is achieved and bandwidth of the network is saved [24, 26].

3) Distributed Crawler: A distributed computing system is called distributed web crawling. Several crawlers are operated to be distributed in the web crawling process, to get the web highest coverage. The synchronization and communication of the nodes are managed by a central server, as it is distributed geographically [21]. Basically, it is using Page rank algorithm due to its high quality search and efficiency. It is strong against crashes of the system and the other actions, and could be fitted to several applications of crawling, and that is one of the significance of distributed web crawler.

4) Parallel Crawler: Normally, multiple crawlers are run in parallel and, thus, are called Parallel crawlers. Multiple crawling Processes are involved in a parallel crawler [3] known as C-procs that may run on workstations network [25]. The Parallel crawlers are subjected to Page Selection and Page freshness [4]. A Parallel crawler could be distributed at locations that are geographically distant or be on local network [21]. Upon downloading documents in a rational time, crawling system parallelization is very energetic [25].

2.4 Requirements for a Crawler

Flexibility: we would like to be able to use the system in a variety of scenarios, with as few modifications as possible. Low Cost and High Performance: The system should scale to at least several hundred pages per second and hundreds of millions of pages per run, and should run on low-cost hardware. Note that well-organized use of disk access is critical to maintain a high speed after the main data structures, such as the "URL seen" structure and crawl frontier, become too large for main memory. This will only happen after downloading several million pages. Robustness: There are several aspects here. First, since the system will cooperate with millions of servers, it has to accept bad HTML, strange server actions and configurations, and many other odd issues [6].

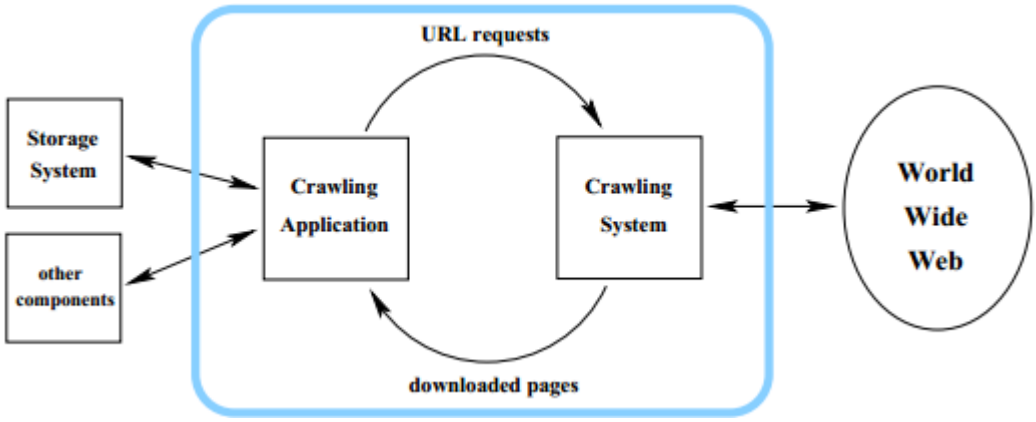

Figure 2: Global crawling architecture

2.5 Caching Mechanism

Caching has been considered widely in the circumstance of operating systems and memory paging, databases [7], Web servers and proxies, as well as Web search systems [23]. In the manuscript, we do not want to wrap the text comprehensively. As an alternative, we spotlight on outcomes linked to the range of the paper.

Caching obtains benefits of the pyramidal structural design of schemes and vicinity of mentions in jobs to allow fast access to pre-calculated or freshly used data. A cache is differentiate by its size and the strategy used for choosing the records to be eliminated when the cache fall full. Two recognizable strategies are grounded on throw out the " Least Recently Used (LRU)", or the " Least Frequently Used (LFU) " points from the cache. LRU is most favorable when the desires are strained from the LRU stack reserve sharing. Alternatively, LFU is most favorable when the desires are drawn from a Zip sharing, which corresponds to the independent reference model (IRM). While LRU is easy to implement, LFU is more challenging because the operational time for a request depends on the cache size, and the correct implementation of LFU requires a complete history of request frequencies [1].

Caching for Web explore systems also has exacting necessities. The position of client demands can be broken by using caching in Web explore systems [18]. Pervious work on caching for data recovery engines make the focal point on the diminution of the server consignment by caching data on the client side, changing the query assessment procedure appropriate to the cached data, or recovering the value of outcomes using a set of perseverant, most favorable demands [8].

3 Proposed Framework and Algorithms

Cloud search systems often analyze file content via so-called Meta Data or Meta Information, such as file name, file size, file keywords and file extension. Recent systems tend to filter query result of cloud searching with meta information into classes and ranking the result according to the log history of user queries. Nevertheless, such the filtration, clustering and ranking increase time and memory requirements together with complexity of the searching process.

3.1 Proposed Algorithms

3.1.1 Proposed Crawling Algorithm

Unlike search engine, we focus only with crawling and retrieving files. So, the attributes or properties that needed for first step of crawling process are file name, extension and size. Proposed algorithm shown in Algorithm1 has four attributes as input URL, Keyword, and File Extension respectively and array of files as output.

3.1.2 Proposed Caching Algorithm

Like all crawling systems, proposed system need to index the results into database, but this traditional process that make retrieving process slow down. So, we over this problem with caching process that make retrieving process more quickly. As we know, cache is smaller than database, so with must use cache for most important files only with

techniques of caching replacement policy. Caching replacement policies techniques that used in our system are LRU (Least Recently Used) and MFU (Most Frequently Used) and implemented according to Algorithm 2.

A perfect implementation of LRU requires a timestamp on each reference, and the system needs to keep a list of files ordered by the timestamp. MFU require the number of times that file is accessed.

Algorithm 1 File Crawling Proposed Algorithm

```
Input: URL, FKeyword, FExtension
Output: arrayOfFiles
Procedure File_Crawling
Begin
B_url = getBaseURL (URL)
P = Download(URL)
Urls = ExtractOutgoingURL in p with B_url
Furls = Fetch Urls format with FKeyword
Foreach Furls as Furl do
IF file extension is in Furls
    arrayOfFiles = Furl
Else
    Continue
End Foreach
IF arrayOfFiles equal null
    File_Crawling(Urls, FKeyword, FExtension)
End IF
```

Algorithm 2 LRU with MFU Caching Proposed Algorithm

```
Input: Class of Files, Time Stamp for each file
Output: Victim Files
If File needed in cache (Cache Hit)
    Access File from Cache with ID
Else File needed not in Cache (Cache Miss)
Search For File in Class (Cluster) in Database
IF File founded in DB
    Select From Cache the least file used with minimum accessed number and longest time stamp As a victim
    Replace it with the file founded from DB
END IF
END IF
```

3.2 Proposed Framework

The system is designed in a modular fashion and is logically composed of two separate phases, via, the retrieval phase (online phase) and the crawling phase (offline phase). This section describe in details the two phases that described in Figure 3 shows system functional specification.

3.3 System Phases

3.3.1 Offline Crawling Phase

The first phase is performed offline as a preprocessing phase in order to make all files can be searched. The First Step, crawl URLs from World Wide Web to find files about queries with most frequently used. In crawling module, File_Crawling algorithm is used, which discussed in Algorithm 1 and filtration process to reduce the amount of data retrieved.

The Second Step, results from the crawling process in the first step would be indexing into database. In indexing module, two algorithms are used. The first part, proposed system uses K-means algorithm to cluster the similar files of similar queries together in one category to decrease time required for retrieval process. The second part in the second step, proposed system uses cache replacement policies such as Least Recently Used (LRU) and Most

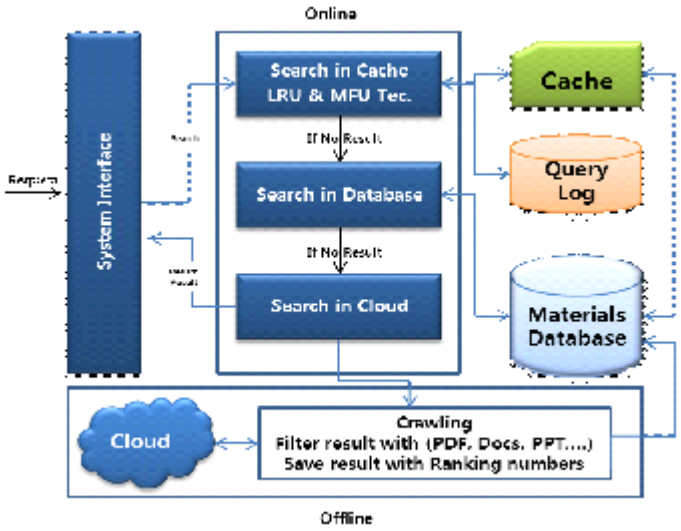

Figure 3: Block diagram of the proposed system

Frequently Used as discussed in Algorithm 2 to increase performance and decrease time needed to retrieve relevant files.

The access log files are collected from the favorite servers. Although there are different access log files based on the types of the server settings, all of these files stores similar information such as request sent time and date, requested resource name (URL), client IP address, and the request method.

The unstructured requests in the log file are processed to create the user sessions. A user session is a collection of web pages which are visited by the user. Exact identification of users and their sessions is very important in web personalization as the user models are built based on their behaviors. Identification of users based on the log file is a difficult task. This is because a typical user can use different hardware platforms and on the other hand different users might use a similar computer. There are various ways to identify the users, but many of them are a threat to security and personalization.

3.3.2 Online Retrieving Process

In the retrieval part, user requests query with additional information about his profile (username, password, IP address, gender and email) for tracking process to system interface. Results of the requested query may be come from one of the three regions. The first region is the cache if and only if category of the requested query in the cache because of that category is top category (Most Accessed Category and Most Recently Used). The second region, if query category is not in cache (cache miss) then get category from database. The third region, if query category is not in database (Database miss) then label this query as miss flag and get it from cloud by using crawling process in offline part.

4 Evaluation and Experiments

4.1 Machine Specifications

The specifications of the machine are Corei7 CPU, 2GB RAM, 500GB Hard Disk and Windows 7. The specifications of the Software are Apache Server (Localhost) with PHP version 5.3 and MYSQL Database version 5.5. The dataset contains 100 online users. In this experiment, we will show the performance of crawling process and time of query process.

4.2 Proposed System Evaluation

Performance of crawling process for 10 queries as sample (Most Frequently Used) is shown in Table 1 and chart Figure 4 with the total number of results that retrieved, irrelevant results and relevant results. Also, Time of that process is shown in Table 2 and chart Figure 5.Table1 is datasheet that describe the result of crawling process of 10

samples (top queries - most frequently queries) from proposed system users. Datasheet in Table 1 shows summation of total results for each query included relevant result (performance) and irrelevant (error) result.

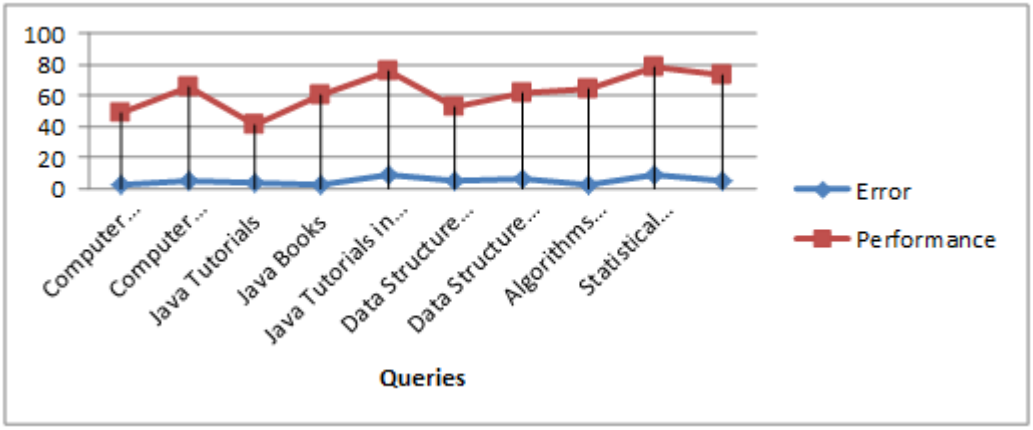

Figure 4: Performance and error chart for crawling with caching process

Table 1: Relevant and irrelevant results of crawling process on 10 queries as sample

Queries 10 Samples Top Queries	Total Results Retrieved After filtration process	Error (Irrelevant Result)	Performance (Relevant Result)
Computer Science Book	50	2	48
Computer Science Tutorial	70	5	65
Java Tutorials	44	3	41
Java Books	62	2	60
Java Tutorials in PDF and DOC	84	8	76
Data Structure Books PDF	57	5	52
Data Structure Algorithms	68	6	62
Algorithms Examples in computer science	66	2	64
Statistical Tutorials	88	9	79
Mobile Computing Tutorials in PDF	77	4	73

Table 2 is datasheet which explain total time of crawling process in seconds of 10 samples (top queries - most frequently queries) from suggested system users.

Table 2: Time for each query of crawling process on 10 queries as sample

Queries 10 Samples Top Queries	Total Time of Crawling and Filtering in second
Computer Science Book	3.94
Computer Science Tutorial	5.5
Java Tutorials	2.98
Java Books	4.55
Java Tutorials in PDF and DOC	6.20
Data Structure Books PDF	4.10
Data Structure Algorithms	4.58
Algorithms Examples in computer science	4.56
Statistical Tutorials	6.57
Mobile Computing Tutorials in PDF	5.90

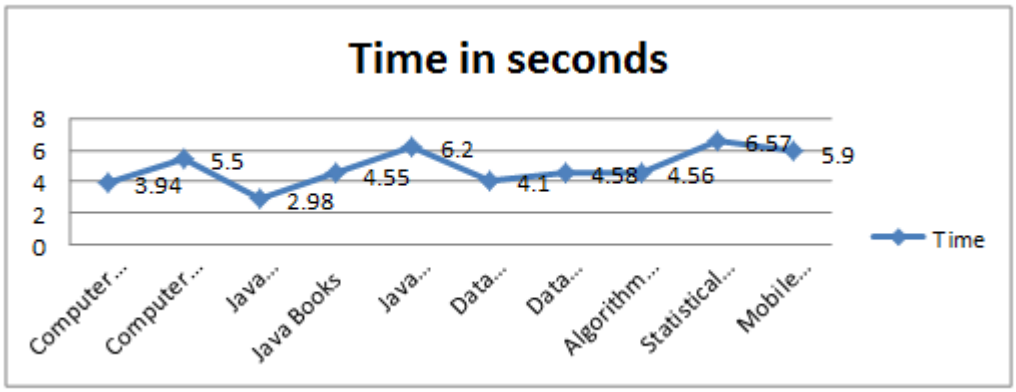

Figure 5: Time chart for crawling with caching process

5 Conclusion and Future Work

A proposed system for fast file searching in cloud environment has been presented. It has been proven that the proposed system is more accurate and faster than other systems. The methodology has used both Crawling and caching algorithms to retrieve file. The proposed system has combined two phases (online and offline) to overcome the challenges and drawbacks of other existing systems. In the future work, query processing algorithms will be integrated with caching to improve the performance of caching algorithms. Another interesting future work will be to apply batch query processing technique for the online case given that the largest search engine gets tens of thousands of queries per second.

References

[1] R. Baeza-Yates, A. Gionis, A. F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri, "The impact of caching on search engines," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 183–190, 2007.

[2] J. Cho, and G. Hector, "The evolution of the Web and implications for a gradual crawler," in *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB'00)*, Cairo, Egypt, pp. 200–209, 2000.

[3] J. Cho, G. M. Hector, "Parallel crawlers," in *ACM WWW 2002*, Honolulu, Hawaii, USA, May 7-11, 2002.

[4] T. AH Chung, F. Daniele, G. Marco, H. Markus and S. Franco, "A simple focused crawler," in *Proceeding 12th International WWW Conference*, pp. 1, 2003.

[5] J. Dean and G. Sanjay, "MapReduce: simplified data processing on large clusters," in *'OSDI'04: Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation*, USENIX Association, Berkeley, CA, USA, pp. 10, 2004.

[6] M. M. El-gayar, N. Mekky and A. Atwan, "Efficient proposed framework for semantic search engine using new semantic ranking algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 8, 2015.

[7] T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data," *ACM Transactions on Information Systems*, vol. 24, no. 1, pp. 51–78, 2006.

[8] Q. Gan, and S. Torsten, "Improved techniques for result caching in web search engines," in *Proceedings of ACM the 18th International Conference on World Wide Web*, pp. 431–440, 2009.

[9] M. Gordon, S. Michael, R. Igor, A. Dan and K. Michele, "Introduction to heritrix," in *Proceedings of the 4th International Web Archiving Workshop (IWAW'04)*, Bath, UK, 16 Sept. 2004.

[10] D. S. B. Hati, and K. Amritesh, "Adaptive focused crawling based on link analysis," in *2nd International Conference on Education Technology and Computer (ICETC'10)*, 2010.

[11] A. Heydon and N. Marc, "Mercator: A Scalable, Extensible Web Crawler," *World Wide Web*, vol. 2, no. 4, pp. 219–229, 1999.

[12] L. Hsin, L. Derek, W. Xiaoming and L. Demtri, "Scaling to 6 Billion Pages and Beyond," *ACM Transactions on Web*, vol. 3, no. 8, pp. 8–34, 2009.

[13] Md. A. Kausar, V. S. Dhaka, S. K. Singh, "Web crawler: A review," *International Journal of Computer Applications*, vol. 63, no. 2, 2013.

[14] E. P. Markatos, "On caching search engine query results," *Computer Communications*, vol. 24, no. 2, pp. 137–143, 2001.
[15] M. Najork, and H. Allan, *High-Performance Web Crawling. In Handbook of Massive Data Sets*, Kluwer Academic Publishers: Norwell, MA, USA, pp. 25-45, 2002.
[16] A. Ntoulas, C. Junghoo and O. Christopher, "What's New on the Web? The Evolution of the Web from a Search Engine Perspective," in *Proceedings of the 13th International Conference on World Wide Web (WWW'04)*, New York, NY, USA, pp. 1-12, 2004.
[17] C. Olston and P. Sandeep, "Recrawl scheduling based on information longevity," in *Proceedings of ACM 17th International Conference on World Wide Web (WWW'08)*, Beijing, China, pp. 437–446, 2008.
[18] R. Ozcan, I. S. Altingovde, and U. Özgür, "Static query result caching revisited," in *Proceedings of ACM the 17th International Conference on World Wide Web*, pp. 1169–1170, 2008.
[19] S. D. K. Pandey and O. Christopher, "A general-purpose algorithm for monitoring Web information sources," in *Proceedings of the 30th International Conference on Very Large Data Bases, (VLDB'04)*, Toronto, Canada, pp. 360–371, 2004.
[20] B. Paolo, C. Bruno, S. Massimo and V. Sebastiano, "UbiCrawler: A scalable fully distributed Web crawler," *Software: Practice and Experience*, vol. 34, no. 8, pp. 711–726, 2004
[21] S. M. Pavalam, S. V. K. Raja, M. Jawahar, and F. K. Akorli, "Web crawler in mobile systems," *International Journal of Machine Learning and Computing*, vol. 2, no. 4, Aug. 2012.
[22] M. Pirnau, "Considerations on the functions and importance of a web crawler," *7th IEEE International Conference on Electronics, Computers and Artificial Intelligence (ECAI'15)*, 2015.
[23] S. Podlipnig and B. Laszlo, "A survey of web cache replacement strategies," *ACM Computing Surveys*, vol. 35, no. 4, pp. 374–398, 2003.
[24] A. K. Sharma and D. Ashutosh, "Self adjusting refresh time based architecture for gradual Web crawler," *International Journal of Computer Science and Network Security*, vol. 8, no. 12, pp. 349–354, 2008.
[25] S. Shruti, A. K. Sharma, and J. P. Gupta, "A novel architecture of a parallel Web crawler," *International Journal of Computer Applications*, vol. 14, no. 4, Jan. 2011.
[26] N. Singhal, D. Ashutosh and A. K. Sharma, "Design of a priority based frequency regulated gradual crawler," *International Journal of Computer Applications*, vol. 1, no. 1, pp. 42–47, 2010.
[27] Y. Xie and O. H. David, "Locality in search engine queries and its implications for caching," *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, 2002.

Mohammed Khaleel Hussein was born in Baghdad - Iraq in 11 Nov 1977. He has received B.Sc. in information system of computer, AL-Rafidain University Collage, Iraq (2000-2001). M.Sc. degree in Computer Science, Iraqi Commission for Computers & Informatics/ Informatics Institute for Postgraduate Studies, Iraq 2004-2005. PhD Student in Information Systems, Faculty of Computers and Information, Mansoura University, Egypt (2012- till now). A Lecturer in AL-Mustansyria Institute for Computer Science (2001-2003). A Lecturer in AL-Mansour Bureau for Industrial and Management Consultation - Mansour University Collage (2003-2005). He joined in June 2007 Scholarships and Cultural Relations Directorate, Ministry of Higher Education and Scientific Research, Scholarships and Cultural Relations Directorate, Iraq. Fields of interest: Information Technology, Information System, Cloud Computing, Computer Education.

Hazem M. El-Bakry (Mansoura, EGYPT 20-9-1970) received B.Sc. degree in Electronics Engineering, and M.Sc. in Electrical Communication Engineering from the Faculty of Engineering, Mansoura University - Egypt, in 1992 and 1995 respectively. Dr. El-Bakry received Ph. D degree from University of Aizu - Japan in 2007. Currently, he is associate professor at the Faculty of Computer Science and Information Systems - Mansoura University - Egypt. His research interests include neural networks, pattern recognition, image processing, biometrics, cooperative intelligent systems and electronic circuits. In these areas, he has published many papers in major international journals and refereed international conferences. According to academic measurements, now the total number of citations for his publications is 2757. The H-index of his publications is 28. Dr. El-Bakry has the United States Patent No. 20060098887, 2006. Furthermore, he is associate editor for journal of computer science and network security (IJCSNS) and journal of convergence in information technology (JCIT). In addition, is a referee for IEEE Transactions on Signal Processing, Journal of Applied Soft Computing, the International Journal of Machine Graphics & Vision, the International Journal of Computer Science and Network Security, Enformatika Journals, WSEAS Journals and many different international conferences organized by IEEE. Moreover, he has been awarded the Japanese Computer & Communication prize in April 2006 and the best paper prize in two conferences cited by ACM. He has also been awarded Mansoura university prize for scientific publication in 2010 and 2011. Dr. El-Bakry has been selected in who Asia 2006 and BIC 100 educators in Africa 2008.

Ahmed A. Saleh is full professor at the Faculty of Computer Science and Information Systems - Mansoura University - Egypt. He is the vice dean for higher studies, research, and cultural affairs. His research interests include E-Business, GIS, and Information Systems.

Cloud Based Technique for Blog Search Optimization

Jitendra Singh
IT Department & Asia Pacific Institute of Management
Plot No 3-4, Institutional Area, Jasola 110025, New Delhi
emaild:jitendra.singh0705@gmail.com
(Received Oct. 15, 2015; revised and accepted Nov. 28 & Dec. 17, 2015)

Abstract

Blogs have witnessed unprecedented growth in last decade. Many new blogs are being generated on a daily basis that is leading to generate large valuable information. Searching the needed information from the available information in minimum efforts and latency is a great challenge. Present work is an attempt to explore the effort made to expedite the blog specific search, to identify and highlight the gaps in the blog specific search engines. In addition, this work has proposed a new algorithm to expedite the searching using the search engine that can be hosted on cloud. Correspondingly, additional resource requirement to meet the peak load demand can be fulfilled by provisioning them instead of procuring the resources. Proposed method uses the hierarchical based technique couples with intelligence factor to search and index the blog pages. Proposed method would greatly help in locating the information available on the blogs.

Keywords: Blog search, information searching in blogs, social media search, trend in blogs

1 Introduction

Blog is the powerful means in the hand of internet users with wide coverage of information across the world. Many people perceive blogs as an analogous with the website, however blogs are different with the website, since blogs are changed frequently [28]. It is also termed as computer mediated communication system by [2, 23, 28, 33]. Computer mediated communication signifies that users are involved in communicating by utilizing the computer technology.

Blogs history is around three decades long, when in the year 1994, the Links.net was created by the Justin Hall, a Swarthmore college student. Blogs at that point in time were not offering the rich functionality as the one they offer now. As a result, blogs could only represent the basic information related to his profile. [5]Jorn Barger in 1997 have used the term Bogging the web that eventually transformed as web-log [5]. In the year 1999, Web log was shortened to blog by the programmer Peter Merhols [5].

Blogs have witnessed slow growth initially, and up-to 1999 there were only 23 blogs on Internet [5]. The key reasons were lack of resources such as Internet connection, costly computers, lack of IT literacy, lack of popularity, etc. Post 1999, there was remarkable increase in the total number of blogs and reached around 50 million by the year 2006[2]. At that time only the IT experts were able to create the blogs. However, since then there is drastic shift in trend and Internet savvy students can create the blog. College students have also started posting their solutions on the internet in order to help their juniors. Freelancers are also leveraging the blogs to promote themselves and demonstrating skill in their specific area of expertise.

Blogs and the social media are considered as powerful tools in the hands of all those aiming to represent their views with greater coverage. Recently, it is attracting huge attention from all the stakeholders and gaining the huge popularity on various fronts. Accessibility using wide variety of devices such as mobile, tablets, PC, etc. has further enhanced the significance of blogs. Worldwide tablet sell has increased by 68% in the year 2013 relative to its preceding year 2012. Increase in the number of portable devices will further lead to the growth in the blogs. Other than the tablets, mobiles have also emerged as great source for accessing internet. Mobile users devote half of their time of mobile usage to access internet and posting information on social media. As a result, usage of mobiles and tablet for internet usage will further grow. Eventually, it will motivate the tablet and mobile users to access the information available on the blog and create as well as maintain their own blogs that will lead to huge information generation.

With the increase in broad network access [21] and coverage, information posted on blogs may go viral. Consequently, user's experience of accessibility of such blog may be poor. In order to offer rich experience, new model

named as cloud has emerged. Cloud is a utility based model where the users have to pay as they go [4, 32]. This model is equally capable to support the spike that is generated due to the sudden increase in work load [30]. Additional work load results in poor response time in the cloud environment. Resources can be increased or decreased instantly with no or minimum effort of the users. Cloud model is a promising model for subscribing the world class IT resources and paying based on the usage. Although, cloud computing model is a promising model, yet it has to overcome of many challenges including security, legal, technical, etc. before to emerge as a undisputed leader in IT resources utilization [29].

Rest of the paper is organized as follows. Section 2 describes the related work that has similarity to our work and contributed considering the blogs. Section 3 describes about the current usage of blogs and factors driving the blog usage. Proposed method is discussed in Section 4. Eventually, conclusion is highlighted in Section 5.

2 Related Work

During the search of related work, we have referred the leading journals such as Springer, Elsevier, EBSCOHOST, etc.; despite of that, many papers relevant to the subject could not be traced. However, articles describing the blog search or blog efficiency would be worth discussing here. Previous works were focused on generalized work related to the blogs . Computer mediated communication based analytical techniques was discussed, authors have suggested adequacy of rules and procedural rules for blogs analytics. Reasons to maintain the blog were highlighted by [12]. Authors had identified that a creative expressions and documentation of personal experiences are the key motivation behind the maintaining blogs. Similar findings were also reported by [25] meeting with other people was also cited as the other motivation of maintaining the blog.

Several work related to content search including web mining and support vector machine (SVM) also exist. For instance, Mishne and Rijke have conducted study on the log of blogs to analyze the query intent and the user's sessions [24]. In another work authors have proposed the SVM method in order to collect the information from the blog, URL and analyzed the content. Chen, Tsai, and Chan have proposed the probabilistic models for blog search and mining that has utilized latent semantic analysis (LSA) and probabilistic latent semantic analysis (PLSA) were introduced to analyze the business data posted on the blog [7]. Melville, Gryc, and Lawrence have analyzed the blog sentiments by combining lexical knowledge with text classification method [22]. Off-line on-line social media (O2SM) and efficient way to render the social media content was suggested by [39]. In this work, authors had developed an app to rank the social media streams and to invest in limited resources by restricting pre-fetching of the multimedia content that is frequently needed. Schmidt have authored the article with wider perspective, he had propose a general model [28]. Work had proposed a framework that can be a basis for systematic comparison and longitudinal studies. Glance, Hurst, and Tomokiyo have attempted to capture the trend in weblogs and used the data mining tools to discover the trend in weblogs [13]. Upon discovering of name, key phrases and key paragraph were posted on blogpulse.com. Authored have also maintained the search-able index in order to expedite the search operation.

Based on the aforementioned studies, it is concluded that the majority of blogs research was concentrated on analytics of blog content that the blogs comprising. Methods such as indexing and mining were proposed for the efficient searching and to gain insight from the data maintained by various blogs.

3 Blogs Current Trend

Blogs are used in the variety of domains and their new usage is set to grow further. Less or no cost, wider coverage, simplicity, etc. are cited as the principal reasons for utilizing the blogs. Blogs are one of the cheapest methods to disseminate the information and is by product of web 2.0 technology. Other than the blogger(s), who represent the information, reader can also contribute the information by way of writing comments. As a result, user receives the feedback of his views [23]. Blogs are widely appreciated since users across the world can review the post and comment on it, provided the owner has allowed that feature. Information representation in blogs is not limited to one or two areas, instead, covering many domains and verticals of the business. Correspondingly, information available on internet can be categorized as political, social, technological, etc. Usage of blog in the variety of areas has been described in the upcoming sub-section.

3.1 Blog and Social Media

Nearly, 6.7 million people blog on blogging sites and 12 million people via social network [blogonomy]. Blogs have deep similarity with the social media [34]. Blogs are also considered as social media, since they allow posting the content that can be commented by the users across the world based on the restrictions imposed by the creator [19]. Several blogs are used as a social media platform that includes buffer social grow, Jon loomer, etc. In addition

Table 1: Social media facts [14, 16]

Social media	users	facts
Pinterest	70 million	69% are female
Instagram	130 million	16 billion photos are already uploaded. Average 40 photos are uploaded per second.1000 comments per second.
Linkedin	238+	Total groups 1.5 million. 27 percent access with the help of mobile.50 percent users are graduate
Google	500+	393 million users
Twitter	500+	
Facebook	1.15 billion +	

to the aforementioned prominent cited examples of blogs, many others are also leveraging the blog for the social media [15, 19]. For instance, [35] have highlighted that how blogs can be used to manage the crisis. At that time, blogs are acting as a tool for public relations.

Social media platforms are also helping in promoting the information posted on a blog. Blogs coupled with social media can have reasonable coverage relative to the one don't exploit the social media [10]. Number of users in various social network platforms has been illustrated in the figure 1. According to Mack Collier, mere good content is not enough to generate the traffic to ones blog instead a blogger needs a meticulous planning and blending with the social media in order to attract the enough traffic to one blog [10].

To promote the blog, bloggers need to select the appropriate social media platform to post their comment. Social network have varying users base and content liking. For instance, Facebook is the most popular social network that is followed by Google plus and twitter and same is illustrated in Figure 1.

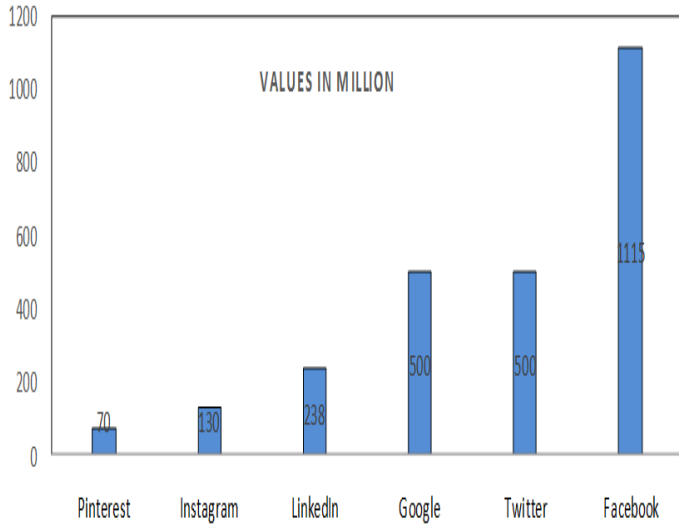

Figure 1: User base of popular social Network

During selection of the social media for promotion, factors such as gender, devices, literacy, etc. should be considered meticulously since they may have fairly significance to the content popularity. Therefore, blogger need to conduct a detailed survey on the appropriateness of the social media in order to gain the popularity of their blog in minimum efforts. For instance, LinkedIn is the social network with the users with extremely high education background have subscribed. LinkedIn is widely accessed with the help of mobile and same has been depicted in Table 1. Therefore, content related to mobile and higher education need to be posted on LinkedIn.

3.2 Blogs as an Alternative to Research Paper

In several Universities, students are expected to write the end term paper. During their presentation of the end term paper, students contribution is discussed and debated with other peer students, beside the faculty members [9, 26]. Therefore, coverage is restricted to the classroom where topic is discussed. In order to gain the wide coverage, blogs are also considered by academia for the variety of purposes including the blogging, collaboration, advertisement, etc. However, majority of academician appreciate the research usage of blogs [38]. According to [8], blogs are not only facilitating the research but also broadening the boundaries of research activities.

Although, blogs were started by the hobbyist to express their opinion, however, now writing blogs have emerged as a part of curriculum. Many universities including Longwood University have introduced the blogging as a replacement to an end term research paper [6]. In the Longwood University, students have written a blog titling Stress and crisis in military families in which they have highlighted the stress level borne by the military personnel and their family during the military services rendered by military personals [1]. Story was written along with the videos which are also acted as support document for the disclosure made in on the blog. It was widely appreciated across the world for raising a cause and the way entire story was written by the students. This resulted in a great learning experience for the students besides change in students vision toward the various aspects of the life. Blog writing can develop the sense of responsibility among the students as a result they start thinking differently [26]. At the same time it improves the student's writing skill and the art of representations; consequently they are market ready for the employer. However, blog is not equally utilized in all the subjects for instance; it is widely adopted by the subjects such as sociology, communication skill, media, etc whereas less used in science subjects.

[8] have conducted the study on the academic usage of blog in research writing. Authors have described the three cases of students of 3rd year of music class and have leveraged the blog for their research activities. [8] have also expressed their concern over the cut and pasted culture prevailing among students during the research writing. In addition, researchers are utilizing the blog to represent the ideas after attending the conference. Ideas generated are reflected in the weblogs writing. Blog also helps in improving the profile of a researcher maintaining the blog, correspondingly, improves the chances of hiring [36].

3.3 Business Usage of Blog

Blogs are widely utilized by the business groups in order to promote themselves and their product. Blogs are the widely used method to promote the product and to gain the business benefit. Organizations using blogs have witnessed huge growth in their business relative to the one not maintaining blogs. Success due to utilization of blog is primarily attributed to the large user base that can be covered with the help of blogs. At the same time, marketing using blog is less expensive relative to the other methods of marketing [14].

Realizing the huge business benefit, organizations that may be big or small are leveraging the blogs in order to gain the business benefits. For instance, in IT sector Microsoft, IBM, Oracle, etc are maintaining blogs beside their own website. In FMCG sector, the company like Procter and Gamble, Hindustan lever, etc. are also maintaining their blogs beside own website. In such cases blogs are not only reducing the website complexity instead offering large space to represent the other information or achievement of the company in terms of good services and quality of product offered, customer satisfaction view collected, etc.

Blog can be created either subscribing to the paid services or to the free services offered by various blog sites. A number of blog site such as world press, blogspot, blogger, etc. are existing. Although, blogger/blogspot are first who have offered the bloggers free space to maintain their blogs, however, they were surpassed by world press due to the ease of usage and greater flexibility offered by the world press. Blogs are especially useful for the new entrepreneur and enterprises who cannot afford the promotional cost. For such needs, blog offers an alternative solution with extremely low cost or at no cost.

3.4 Blogs for Revenue Generation

Blogs are acting as a great source of income. Many prominent bloggers have initially started with basic blogs and subsequently switched over to the complete website, upon the demand and traffic to the website increased. Earlier fewer companies were interested to advertisement on the blog. Ad sense was the first advertising platform that was launched in the year 2003 that can display the advertisement on the blog [14, 16]. Correspondingly, advertising has offered the huge opportunity to earn money from the blogs. Darren Rowse of problogger.net and John cow were the leading bloggers who have made sizable amount of earning from their blogs that they have started.

Blogs are equally maintained by the individual bloggers as well as the corporate. In the individual category, it needs great motivation in the beginning, in order to keep the blog going when the revenue is low. However, with efforts and planning blog may emerge as a great source of earning that may even touch the unimaginable value. The Huffington Post, Mashable, Tech crunch, etc, are the prominent blogs.Per day income of these prominent blogs have

already touched the $1000 marks. Pay per clicks, advertising banner, Membership area, CPM advertising, Affiliate sales, Private advertising, etc. are some of the methodologies employed by them for the earning (Smith, 2015).

Although, consistent increase in terms of blogger as well as users accessing the blogs have been witnessed over the couple of years. Despite of that all the blogs created are not emerging successful, instead popularity of any blog(s) is governed by various factors; some of them have been enumerated as follows:

- Content of the blog.
- Blogger capability to promote the blog by way of face book, twitter, LinkedIn, etc.
- Relevance of the topic.

Correspondingly, mere maintaining the blog is not enough, instead the content posted, relevance of the topic in the contemporary world along with the promotional activities are the key for the success of the blog. Absence of these attributes may lead to the failure of the blog.

3.5 Challenges to Blog

Recently, new paradigm on internet have also emerged and acting as an alternative to the blog. Emerged methods allowed posting the content in audio or video formats therefore, equally liked by the bloggers and content readers. Blogger appreciate it since it needs fewer efforts, whereas readers like it since, it needs less time to understand the topic and great amount of text can be covered in minimum time.

Beside blogs, podcast are widely used to represent the information. Primarily blogs are used to represent the personal view of the blogger, whereas podcast are used to post the video and audio contents. Micro blog, video content, podcast, etc. are posing the great challenge to blogs; however, it is fairly sought medium of representing the information. On a positive side blogs are free; as a result, blogger is not incurring any cost in maintaining the information.

In addition, to several good usages, mis-utilization of social media has also been noticed. [37] have highlighted the social media as a promising tool for marketing and to find out the potential client for the drug selling. Authors have also quoted the various instances where social media has been utilized for drug distribution. Correspondingly, authors has also suggested the analytical method for detecting the mis-utilization of the social media for illicit drug distribution [37]. Despite aforementioned challenges and issues, blogs are still considerable popular and same can be determined by the fact that in every 06 seconds one new blog is created across the world.

4 The Proposed Method

As the time passes, a successful model needs to be suitably modified in order to meet the contemporary business requirement. In the same line, in order to optimize the blog search, no dedicated engine is available; as a result, users are searching the information with the help of available general search engine. Although in 2008, Google the IT giant has envisioned the search engine that was dedicated for the blog search, however, it was rolled back in the year 2014. Due to lack of blog specific search engine, users are unable to search the blog specific information effectively.

4.1 Need for Cloud Based Research Technique

Although the search engines such as blogpulse, Google etc. were envisioned to determine the blog specific trend, analytics, real time buzz, etc. However, these research engines could not last for the long time. Blogpulse was ended in the year 2012 [8] whereas Google ended the blog specific search engine in the year 2014. The key reasons for the failure were the investment and the lack of business benefit that can be realized with the help of blog search engine.

In our system, we have proposed the cloud based technique for the search engine optimization because searching to the blog specific information may rise remarkably upon specific information goes viral. Therefore, there is a need for a technique that should able to scale-up in order to meet the growing users demand and should be capable to scale-down once the spike is over. Maintaining the search engine on the cloud would be cheaper at the same time cost effective as it won't require maintaining the additional resources to meet the peak need of users.

4.2 Motivation

Blogs are the cheapest mode to represent the information, as a result, growing with the tremendous pace. According to insight, one new blog is generated in every 06 second. It means in near future, huge information is set to be available from the internet. Information seekers may be lost in the ocean of information. Therefore, a method is

needed in order to represent the information in a manner that can be searched and retrieved in minimum time. Presently, no such viable solution exist. Although, Google has created the search engine that was entirely dedicated to the blog, unfortunately, they have to stop their operation in the year 2014. Since then, there is no tools exist to search the information and expedite the searching of blog based information. Therefore, it is worth to evolve a technique that should expedite the blog specific search. At the same time, the suggested technique should be cheap and cost effective in order to mitigate the risk of failure.

4.3 Intelligence Based Hierarchical Based Model

In blogs, wide variety of information is posted. In order to optimize the blog search, Blog need to be categorized based on the discipline of the post, sub-group titles, etc.In order to facilitate the searching, information database of blog provider need to be created. Search engine maintains the information of all the blogs in its own database by categorizing them as suggested in Figure 2.

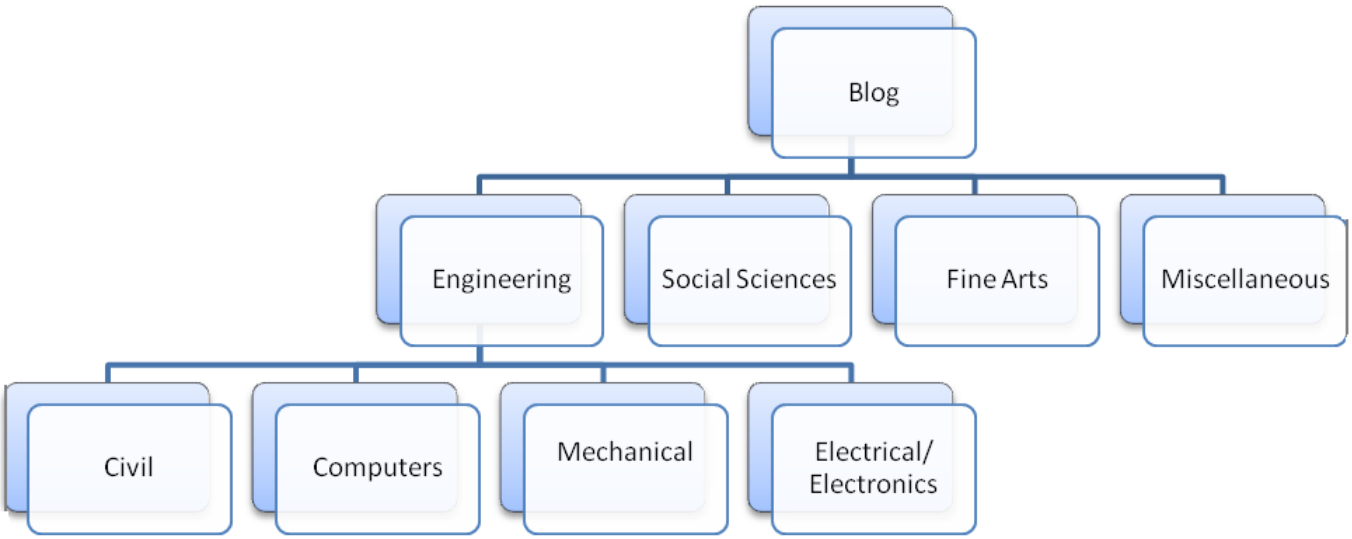

Figure 2: Hierarchical based approach

During creation of a blog, the blogger needs to categorize blogs in the appropriate category and sub-category in order to be placed at the appropriate location in the search engine database. Upon creation, the blog would be position based on index. However, only hierarchical based approach won't be sufficient for the efficient data search. Therefore, intelligence is needed in order to minimize the latency. Therefore, two additional parameters total count and day count are also included. Upon each access, total count and the day count of the blog will increase by one. Positioning of the blog would be based on the day count followed by the total count a web page is searched. The principal reason behind this is that on a specific day if the information posted becomes viral in such cases only those information is accessed rest of the pages/ blogs are less accessed [2]. However, there is the information that is always needed for instance, blog offering the information related to the basics or the fundamental knowledge of a subject.

5 Conclusion

Blogs are increasingly sought by the variety of users and have gained in popularity due to their simplicity and cost effectiveness. Contents in blog is growing at tremendous rate. Although, blogs are increasing on a daily basis yet no search mechanism exist that should be entirely dedicated to the blog search. Searching mechanism should be effective enough to obtain the needed information in minimum time. Therefore, a web based tool for searching the blog is the need of an hour. Proposed method would not only improve the searching on the web instead would improve the potential of expansion to the cloud platform. Proposed method would greatly help in improving the search on the internet pertaining to the blog without needing any additional expenses.

Acknowledgments

The author(s) gratefully acknowledge the anonymous reviewers for their valuable comments with the help of which the paper's quality can be improved.

References

[1] L. Bidwell, "Stress and crisis in military families," 15 May 2015. [Online]. Available: http://blogs.longwood.edu/socl306f12/

[2] R. Blood, "How blogging software reshapes the online community," *Communications of the ACM*, vol. 47, no. 12, pp. 53–55, 2004.

[3] D. Bosomworth, "Statistics on mobile usage and adoption to inform your mobile marketing strategy," Jul 22, 2015. [Online]. Available: http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/

[4] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, 2009.

[5] C. Chapman, "A brief history of blogging," MAR 14, 2011. [Online]. Available: http://www.webdesignerdepot.com/2011/03/a-brief-history-of-blogging/

[6] M. Chau and J. Xu, "Mining communities and their relationships in blogs: A study of online hate groups," *International Journal of Human-Computer Studies*, vol. 65, no. 1, pp. 57–70, 2007.

[7] Y. Chen, F. S. Tsai, and K. L. Chan, "Machine learning techniques for business blog search and mining," *Expert Systems with Applications*, vol. 35, no. 3, pp. 581–590, 2008.

[8] E. K. Chong, "Using blogging to enhance the initiation of students into academic research," *Computers & Education*, vol. 55, no. 2, pp. 798–807, 2010.

[9] D. Churchill, "Educational applications of web 2.0: Using blogs to support teaching and learning," *British Journal of Educational Technology*, vol. 40, no. 1, pp. 179–183, 2009.

[10] M. Collier, "The idea that 'content is king' in blogging is total bullshit," June 04, 2009. (http://moblogsmoproblems.blogspot.in/2009/06/idea-that-content-is-king-in-blogging.html)

[11] D. Drezner and H. Farrell, "The power and politics of blogs." American Political Science Association, 2004.

[12] S. Fox and A. Lenhart, "Bloggers: A portrait of the internets new storytellers," *Pew Internet & American Life Project*, vol. 19, 2006.

[13] N. Glance, M. Hurst, and T. Tomokiyo, "Blogpulse: Automated trend discovery for weblogs," in *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, vol. 2004. New York, 2004.

[14] E. Hood, "The #1 small business marketing idea [infographic]," Aug 07, 2013. [Online]. Available: http://blog.ignitespot.com/blog/small-business-marketing-idea

[15] N. Hookway, "Entering the blogosphere': some strategies for using blogs in social research," *Qualitative Research*, vol. 8, no. 1, pp. 91–113, 2008.

[16] D. Karr, "The blogconomy blogging statistics [infographic]," August 26, 2013. [Online]. Available: http://blog.ignitespot.com/blog/small-business-marketing-idea

[17] A. Keen, *The Cult of the Amateur: How blogs, MySpace, YouTube, and the rest of today's user-generated media are destroying our economy, our culture, and our values.* Broadway Business, 2008.

[18] T. Kelleher and B. M. Miller, "Organizational blogs and the human voice: Relational strategies and relational outcomes," *Journal of Computer-Mediated Communication*, vol. 11, no. 2, pp. 395–414, 2006.

[19] C. King, "Top 10 social media blogs: The 2015 winners!" February 5, 2015. [Online]. Available: http://www.socialmediaexaminer.com/top-10-social-media-blogs-2015-winners/

[20] C. Macdonald and I. Ounis, "The trec blogs06 collection: Creating and analysing a blog test collection," *Department of Computer Science, University of Glasgow Tech Report TR-2006-224*, vol. 1, pp. 3–1, 2006.

[21] P. Mel and T. Grance, "The nist definition of cloud computing. national institute of standards and technology," *Information Technology Laboratory, Version*, vol. 15, no. 10.07, p. 2009, 2009.

[22] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, pp. 1275–1284, 2009.

[23] C. R. Miller and D. Shepherd, "Blogging as social action: A genre analysis of the weblog," *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*, vol. 18, no. 1, pp. 1–24, 2004.

[24] G. Mishne and M. De Rijke, "A study of blog search," in *Advances in Information Retrieval.* Springer, pp. 289–301, 2006.

[25] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz, "Why we blog," *Communications of the ACM*, vol. 47, no. 12, pp. 41–46, 2004.

[26] P. release, "Blogs becoming alternative to research paper in some longwood classes," 30 Nov 2012. [Online]. Available: http://www.longwood.edu/2012releases_45769.htm
[27] J. Schmidt, "Blogging practices in the german-speaking blogosphere. findings from the "wie ich blogge?! working paper 07-02," June 13, 2007. [Online]. Available: http://www.fonk-bamberg.de/pdf/fonkpaper0702.pdf
[28] J. Schmidt, "Blogging practices in the german-speaking blogosphere," *Bamberg: Universität Bamberg, at http://nbn-resolving. de/urn: nbn: de*, 2007.
[29] J. Singh, "Cyber-attacks in cloud computing: A case study," 2015.
[30] J. Singh and V. Kumar, "Implementation of user-end broker policy to improve the reliability of cloud services," *International Journal of Cloud Applications and Computing (IJCAC)*, vol. 3, no. 4, pp. 13–27, 2013.
[31] ——, "Multi-disciplinary research issues in cloud computing," *Journal of Information Technology Research (JITR)*, vol. 7, no. 3, pp. 32–53, 2014.
[32] J. Singh, K. S. Mathur, and V. Kumar, "Enhancing security in mobile cloud computing," *Proceedings of M4D 2012 28-29 February 2012 New Delhi, India*, pp. 460–464, 2012.
[33] S. Stewart, "Blogging and research," Dec 2008. [Online]. Available: http://sarah-stewart.blogspot.com/2008/11/blogging-and-research.html
[34] H. Sullivan, "Why are blogs considered social media?" June, 2009. [Online]. Available: fromhttp://www.cision.com/us/2009/06/why-are-blogs-considered-social-media/
[35] K. D. Sweetser and E. Metzgar, "Communicating during crisis: Use of blogs as a relationship management tool," *Public Relations Review*, vol. 33, no. 3, pp. 340–342, 2007.
[36] M. Thelwall and L. Hasler, "Blog search engines," *Online Information Review*, vol. 31, no. 4, pp. 467–479, 2007.
[37] P. A. Watters and N. Phair, "Detecting illicit drugs on social media using automated social media intelligence analysis (asmia)," in *Cyberspace Safety and Security*. Springer, pp. 66–76, 2012.
[38] J. B. Williams and J. Jacobs, "Exploring the use of blogs as learning spaces in the higher education sector," *Australasian Journal of Educational Technology*, vol. 20, no. 2, 2004.
[39] Y. Zhao, N. Do, S.-T. Wang, C.-H. Hsu, and N. Venkatasubramanian, "O 2 sm: Enabling efficient offline access to online social media and social networks," in *Middleware 2013*. Springer, pp. 445–465, 2013.

Jitendra Singh has Pursued his PhD(Computer Science) in the area of cloud computing. He is having over 12 years of experience in the various domains of Information technology. Prior to doctorate, he has pursued the master in computer applications as well as masters in computer technology. In addition, he has qualified the prestigious UGC-NET examination conducted by UGC India in the year 2006. As a faculty engaged in teaching to the students of Bachelors and masters Degree of several premier Universities of India that includes University of Delhi, IP University, etc. as well as foreign University including the Stratford University, USA, etc. to name a few. He is also contributing as a member of several committees which are monitoring and reviewing the curriculum implementation. He authored more than 20 research articles on cloud computing security, performance, etc. in which several of them are published in the leading research journals and conferences. Besides, he has authored two books pertaining to the computer science, one on the cloud computing titling cloud computing for beginner and researcher and the other one on data structure titling Data structure simplified: Implementation using C^{++}.

Cryptanalysis of Multi-prime RSA With Two Decryption Exponents

Kumar R. Santosh[1], Challa Narasimham[2], and Pallam ShettyS[3]
(Corresponding author: Kumar R. Santosh)
IT Department& MVGR College of Engineering[1]
Chintalavalasa, Vizianagarm, Andhra Pradesh, India
CSE Department & Vignan Engineering College[2]
Vishakapatnam, Andhra Pradesh, India
CS SE Department & Andhra University[3]
Vishakapatnam, Andhra pradesh, India
(Email: rsantumvgr@gmail.com)
(Received Dec. 6, 2015; revised and accepted Jan. 20, 2016)

Abstract

Multi-prime RSA is an extended version of RSA in which the modulus is a product of three or more primes. In this paper, we cryptanalyze the Multi-prime RSA if user generates two instances with the same modulus. We use lattice basis reduction and our attack improves the bound of Howgrave-Graham.

Keywords: Lattices, Lattice reduction, Multi-prime RSA

1 Introduction

Multi-prime RSA is an extension of RSA, where the modulus N is a product of three or more large distinct primes [13]. The advantage of Multi-prime RSA over standard RSA lies with the decryption process. The efficiency of decryption process can be improved if one uses the Chinese Remaindering Theorem. The encryption process is same as the standard RSA. In this paper, we assume that the primes in the modulus are balanced. So we have if $p_i < p_{i+1}$ then $4 < \frac{1}{2}N^{\frac{1}{r}} < p_1 < N^{\frac{1}{r}} < p_r < 2N^{\frac{1}{r}}$, where r is the number of primes in the modulus N. The public and private exponents are inverses to each other modulo $\phi(n)$. So we have $\phi(n) = N - s$ where $s = \Sigma_{i=1}^{r}\frac{N}{p_i} - \Sigma_{i,j=1}^{r}\frac{N}{p_i p_j} + \Sigma_{i,j,k=1}^{r}\frac{N}{p_i p_j p_k} + \cdots$. After simplification, one can get the upper bound for s as $N^{1-\frac{1}{r}}$.

Multi-prime RSA can be broken if one can find the prime factors of the modulus. For the standard RSA, if one knows the secret exponent d or $\phi(N)$, then there are deterministic algorithm to factor the modulus. But in Multi-prime RSA, there exist only probabilistic algorithms to find the factors of the modulus if one knows the multiple of $\phi(N)$.

In this paper, we investigate the Multi-prime RSA cryptosystem if user generates two instances with the same modulus. That is, user generates two public exponents e_1, e_2 and two private exponents $d_1, d_2$with the same modulus N. Hinek and Lam [8]investigated and showed that if $d_1, d_2 < N^{\delta}$ then modulus can be factored probabilistically if $\delta < \min\{\frac{3+r}{7r}, \frac{1}{r}\}$ in time polynomial in bit size of N.We improve this bound by applying Lattice reduction techniques.

The remainder of this paper is organized as follows. In Section 2, some mathematical preliminaries are introduced, like the foundations of lattices, and the concept of lattice reduction technique and its application of solving integer equation for small solutions. In Section 3, we present our attack followed by comparing with earlier attacks. in Section 4, we present experiment details of this attack. In Section 5, we analyze the strengths of the proposed attack. Finally, we draw our conclusions in Section 6.

2 Preliminaries

In this section, we introduce the concept of the lattices, lattice reduction algorithms and Coron's approach for solving integer equations.

2.1 Lattices

Let $v_1, v_2, \cdots, v_n$ be linear independent vectors of Z^m, where n, m are two positive integers satisfying $n \leq m$. The linear combination of these vectors using integer coefficients is called a lattice. That is $L = \{\Sigma_{i=1}^{n} : \lambda_i v_i | \lambda_i \in Z\}$. The vectors $v_1, v_2, \cdots, v_n$ is called a lattice basis. The dimension of the lattice is n and it is called the full rank if $m = n$. The determinant of L is defined as the $\Pi_{i=1}^{n} v_i{}^*$, where the v_i^* are orthonormal vectors obtained from the Gram-Schimdt orthogonalization process. If lattice is full rank, the determinant of the lattice is equal to the determinant of the matrix by considering the basis vectors as rows.

2.2 Lattice Reduction

For given lattice, there are infinitely many lattice bases. Among all these lattice bases the one which has small norm other than zero is an interesting one. The basis with small norm is called the reduced basis. Finding the optimal reduced basis for a given lattice is a hard problem if the dimension of the lattice > 2. So by defining the reduced basis such that it can be computed in polynomial time is an interesting problem. In 1982, Lenstra, Lenstra, and Lovasz introduced the notion of LLL-reduced basis and they introduced the polynomial time algorithm too, called LLL Algorithm. The notion of LLL reduced basis and LLL algorithm refer [2, 9].

2.3 Multi-prime RSA

Multi-prime RSA is an extended version of RSA introduced by the Rivest et.al in their seminal paper [13]. As for RSA, there exists several attacks on Multi-prime RSA. The first attack on Multi-prime RSA is given by Hinek citeHinek08 and it is the extension of Wiener's attack on RSA [16]. He showed that the secret exponent d satisfies $d \leq \frac{N}{2sk} \leq \frac{N^{1r}}{2k(2r-1)}$, then one can recover the prime factors of the modulus probabilistically.

Later, Ciet et.al [4] extended Boneh-Durfee sub lattice attack [3] on RSA to Multi-prime RSA and he showed that given the public informaton $(N, e = N^{\alpha})$, if the secret exponent $d \leq N^{\delta}$ satisfies $\delta \leq \frac{r-\sqrt{\alpha(r-1)}}{r}$, then the prime factors of the modulus can be factored efficiently. The same result also can be applied by using unravelled linearization technique [12].There also some attacks exist if Multi-prime RSA with small prime difference among the primes.For the results in direction please refer [1, 17, 18].

Navneet Ojha et.al [10] also analyzed the Multi-prime RSA if the secret exponent is greater than the public exponent In this paper, we investigate the Multi-prime RSA cryptosystem if user generates two instances with the same modulus. That is, user generates two public exponents e_1, e_2 and two private exponents $d_1, d_2$with the same modulus N. Stusying the cryptosystem with two or more decryption exponents has been started by analyzig the Wiener's attack on RSA by Howgrave-Graham [6]. For the multiprime RSA the first attack in this direction given by Hinek and Lam [7] and showed that if $d_1, d_2 < N^{\delta}$ then modulus can be factored probabilistically if $\delta < \min\{\frac{3+r}{7r}, \frac{1}{r}\}$ in time polynomial in bit size of N. For attacks on other variants by considering two decryption exponents please refer [11, 15, 19].

2.4 Solving Equations for Small Integer Roots

In this section, we introduce the coron's approach of solving equations for small integer roots with the ideas of Jochemsz and May čiteJochemay2006 . Suppose one wants the small integer root $x_1^0, \cdots, x_n^0$ of an irreducible polynomial f. Assume that root is small so let $|x_j^0| < X_j$ for all $j = 1, 2, \cdots, n$. Let ϵ be arbtrarily small constant. Let m be fixed integer depending on $\frac{1}{\epsilon}$. Let maximum degree of x_j in f be d_j and the maximal coefficient of $f(x_1X_1, x_2X_2, \cdots, x_nX_n)$ be W, i.e $W = \|f(x_1X_1, x_2X_2, \cdots, x_nX_n)\|_{\infty}$.Define $R = W\Pi_{j=1}^{n} X_j^{d_j(m-1)}$. Also let $f' = a_0^{-1} f mod R$, where a_0 is the constant term of f. Denote f' as f.

Basic strategy:

Define the sets S and M contains the monomials of $f^(m-1)$ and f^m respectively and set $l_j = d_j(m-1)$, which is the largest exponent of x_j that appear as a monomial in S. Next define the shift polynomials

$$g: \quad x_1^{i_1} x_2^{i_2} x_3^{i_3} \ldots x_n^{i_n} f(x_1, x_2, \cdots, x_n) \Pi_{j=1}^{n} X_j^{l_j - i_j} \text{for} x_1^{i_1} x_2^{i_2} x_3^{i_3} x_n^{i_n} \in S \tag{1}$$

$$g': \quad x_1^{i_1} x_2^{i_2} x_3^{i_3} \cdots x_n^{i_n} R for x_1^{i_1} x_2^{i_2} x_3^{i_3} \cdots x_n^{i_n} \in M \setminus S. \tag{2}$$

It is trivial that g and g' have the common root $x_1^(0), \cdots, x_n^(0)$ modulo R. The lattice is constructing by considering the coefficient vectors of $g(x_1X_1, x_2X_2, \cdots, x_nX_n)$. Using lexicographic ordering of the monomials one can get the triangular matrix. The diagonal entries in the row are constant terms in f. So the diagonal

entries of the matrix are $\Pi_{j=1}^{n} X_j^{d_j(m-1)}$ for the polynomials g and $W\Pi_{j=1}^{n} X_j^{d_j(m-1)+i_j}$ for the polynomials g'. After applying LLL algorithm to the above matrix it outputs $n-1$ polynomials having the root of the given polynomial. The condition $\det L < R^{\omega+2-n}$ reduces to $\Pi_{j=1}^{n} X_j^{s_j} < W^{s_w}$, for $s_j = \Sigma_{x_1^{i_1}x_2^{i_2}x_3^{i_3}\ldots x_n^{i_n} \in M_S^{\S}} i_j$ and $s_W = |S|$. The choice of R ensures that h_i are independent of f. But it does not prevent algebraic dependency from the polynomials h_i. So again by assumption only one can reveal the root.

Extended strategy:

The above process can be extended by considering the extra shifts of the certain variables. If one wants the extra shift for the variable x_1, one may define the new sets S and M as follows:

$$S = \cup_{0\le j\le t} x_1^{i_1+j} x_2^{i_2} x_3^{i_3} \cdots x_n^{i_n} | x_1^{i_1} x_2^{i_2} x_3^{i_3} \cdots x_n^{i_n}.$$

The above equation is a monomial of f^{m-1} and $M = \text{Monomials of} x_1^{i_1} x_2^{i_2} x_3^{i_3} \cdots x_n^{i_n} f | x_1^{i_1} x_2^{i_2} x_3^{i_3} \cdots x_n^{i_n} \in S$.

With These definitions the rest of the strategy follows as basic strategy. In this case the R value is changed to $R = W\Pi_{j=1}^{n} X_j^{l_j}$, where l_j as before.

3 Main Result

In this we introduce the main result and its justification. We follow the approach of Jocemsz and May.

Theorem 1. *Let N be a modulus of r-Multi-prime RSA. Let e_1, e_2 be two public exponents and d_1, d_2 be private exponents modulo $\phi(N)$. Let $d_1, d_2 < N^{\delta}$ and $|d_1 - d_2| < N^{\beta}$. Then the prime factors of the modulus N can be recovered in poly$(LogN)$ time when $2r\beta - r - 3 + 4r\delta < 0$.*

Justification. From two RSA equations

$$\begin{aligned} e_1 d_1 - k_1\phi(N) &= 1, \\ e_2 d_2 - k_2\phi(N) &= 1. \end{aligned}$$

From above two equations:

$$e_1 e_2 (d_1 - d_2) - e_2 k_1 \phi(N) + e_1 k_2 \phi(N) = e_2 - e_1.$$

So

$$f(x_1, x_2, x_3, x_4) = a_1 x_1 + a_2 x_2 + a_3 x_2 x_4 + a_4 x_3 + a_5 x_3 x_4 + a_6.$$

Then $(d_1 - d_2, k_1, k_2, v)$ is a root of $f(x_1, x_2, x_3, x_4)$. Since the root is small it can be recovered by using Coppersmiths methods [5]. The approach of Jochemsz and May for choosing the polynomials is used. Let $e_1, e_2 \approx N^{\gamma}$, $d_1, d_2 \approx N^{\delta}$, $|d_1 - d_2| < X_1 = N^{\beta}$, $k_1 = (e_1 d_1 - 1)/(\phi(N)) < X_2 = N^{\gamma+\delta-1}$, $k_2 = (e_2 d_2)/(\phi(N)) < X_3 = N^{\gamma+\delta-1}$, $\phi N = N - S$ with $|S| < (2r-1)N^{1-\frac{1}{r}} = X_4$. After applying the above process of finding small integer roots, we get

$$\begin{aligned} s_1 &= \frac{1}{12} m(m+1)(m+2)(m+2t+1), \\ s_2 &= \frac{1}{24} m(m+1)(m+2)(3m+4t+5), \\ s_3 &= \frac{1}{24} m(m+1)(m+2)(3m+4t+5), \\ s_4 &= \frac{1}{24} (m+1)(m+2)(3m^2+5m+8tm+6t+6t^2) \\ |S| &= \frac{1}{12} m(m+1)(m+2)(m+2t+1). \end{aligned}$$

Plug $t = \tau m$, then we have

$$\begin{aligned} s_1 &= \frac{1}{12}(2\tau+1)m^4 + o(m^4), \\ s_2 &= \frac{1}{24}(4\tau+3)m^4 + o(m^4), \\ s_3 &= \frac{1}{24}(4\tau+3)m^4 + o(m^4), \\ s_4 &= \frac{1}{12}(6\tau^2+8\tau+3)m^4 + o(m^4), \\ |S| &= \frac{1}{12}(2\tau+1)m^4 + o(m^4). \end{aligned}$$

After cancelled the terms of higher order terms and substitute in $\Pi_{j=1}^{4} X_j^{s_j} < W^{|S|}$. we get $X_1^{\frac{1}{12}(2\tau+1)}\ X_2^{\frac{1}{24}(4\tau+3)}\ X_3^{\frac{1}{24}(4\tau+3)}\ X_4^{\frac{1}{24}(6\tau^2+8\tau+3)} < W^{\frac{1}{12}(2\tau+1)}$. Plug the X_1, X_2, X_3, X_4, W values into above inequation we get $\frac{\beta}{12} + \frac{1}{8}(2\delta + 1 - \frac{1}{r}) < \frac{1}{12}(2+\delta)$. This further, reduced into $2r\beta - r + 4r\delta < 0$. The result of the LLL algorithm with above condition produces four polynomials with the root. We Apply the resultant technique repeatedly to get the desired result. If $r = 2$, the bound is equal sarkars attack [14].

Corollary 1. *Let N be a modulus of the Multi-prime RSA, Let e_1, e_2 be two public exponents and d_1, d_2 be private exponents modulo $\phi(N)$. Let $d_1, d_2 < N^{\delta}$. Then the prime factors of the modulus N can be recovered (probabilistically) in poly$(LogN)$ time when $\delta < \frac{1}{6} + \frac{1}{2r}$.*

Put $\beta = \delta$ in above result, the desired result is obtained.

The bound in the above attack is extended compare to the Hibek-Lam attack. The following table illustrates the comparisons.

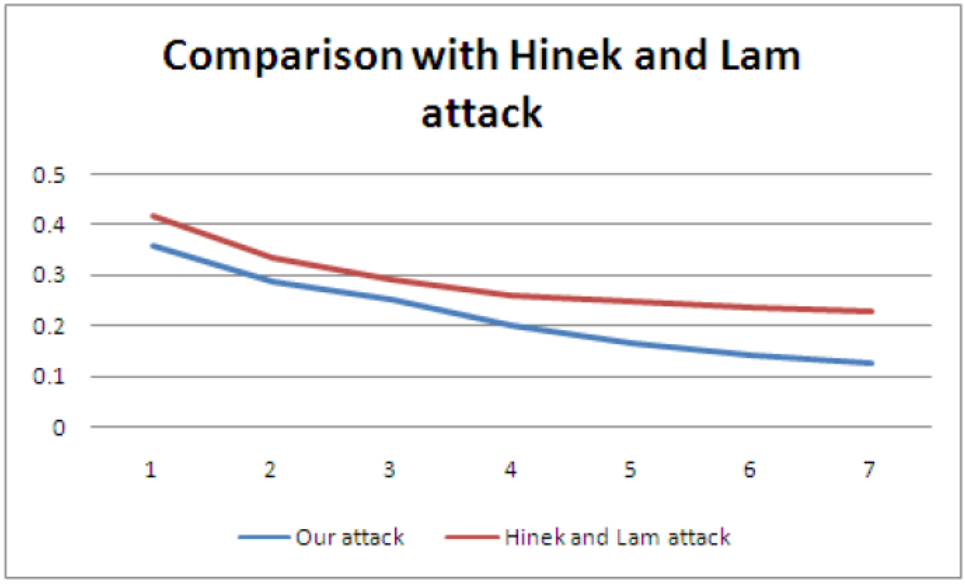

Figure 1: Hinek-Lam vs our attack

It is trivial that Multi-prime RSA system can be broken for the secret exponent values in our attack, where as Hinek- Lam cannot.

4 Experiment Details

We implemented the attack program in SAGE 2.1 over windows 7 on a laptop with intel i5 CPU 1.83GHz, 4 GB RAM and 2MB Cache. We considered the results for $m = 3 and t = 0$, $m = 3, t = 1$, and $m = 3 and t = 2$ all takes the few hours of time due to constraint on the lattice dimension. For almost all the inputs, we successfully retrieve the root. From the lattice, we apply LLL algorithm and considered the first four rows of LLL reduction bases. Then we repeatedly apply resultant technique for desired output.

5 Conclusions

- Multi-prime RSA is one of the RSA varaint can be used in modern cryptography. But there may some flaws in the Multi-prime RSA crypto system. In this paper we analyzed this system if user generates two instances with same modulus.
- We improved the previous bound by using lattice reduction algorithm namely LLL.
- This attack can be extended, if user generates n instances with the same modulus.

References

[1] H. M. Bahig, A. Bhery, and D. I. Nassr, "Cryptanalysis of **Multiprime RSA** with small prime difference," in *Informtion and Communications Security*, LNCS 7618, pp. 33–44, Springer, 2012.

[2] D. Boneh, "Twenty years of attacks on the **RSA** cryptosystem," *Notices of the AMS*, vol. 46, no. 2, pp. 203–213, 1999.

[3] D. Boneh and G. Durfee, "Crypanalysis of **RSA** with private key d is less than $n^{0.292}$," *IEEE Transactions of Information Theory*, vol. 46, no. 4, pp. 1339–1349, 2000.

[4] M. Ciet, F. Koeune, F. Laguillaumie, and J.-J. Quisquater, "Short private exponent attacks on fast variants of **RSA**," Technical Report UCL Crypto Group Technical Report Series CG-2002/4, 2002.

[5] D. Coppersmith, "Small solutions to polynomial equations and low exponent vulnerabilities," *Journal of Cryptology*, vol. 10, no. 4, pp. 223–260, 1997.

[6] N. H. Graham and J. P. Seifert, "Extending wiener's attack in the presence of many decryption exponents," in *CQRE*, LNCS 1740, pp. 153–166, Springer, 1999.

[7] M. J. Hinek, "On the security of multiprime **RSA**," *Journal of Mathematical Cryptology*, vol. Vol 2, no. 2, pp. 117–147, 2008.

[8] M. J. Hinek and C. C. Y. Yam, "Common modulus attacks on small private exponent **RSA** and some fast varaints (in practice)," Technical Report Cryptology ePrint Archive 2009/037, 2009.

[9] A. K. Lenstra, H. W. Lenstra, and L. Lovasz, "Factoring polynomials with rational coeffients," *Mathematische Annalen*, vol. 261, pp. 267–282, 1982.

[10] N. Ojha and S.Padhye, "Cryptanalysis of **Multiprime RSA** with secret key greater than public key," *IEICE Transactions on Fundamentals of Electronics*, vol. 16, no. 1, pp. 53–57, 2014.

[11] L. Peng, L. Hu, Y. Lu, S. Sarkar, J. Xu, and Z. Huang, "Cryptanalysis of variants of **RSA** with multiple small secret exponents.," in *INDOCRYPT'15*, LNCS 9462, pp. 105–123, Springer, 2015.

[12] S. K. Ravva, Ch Narasimham, and S. S. Pallam, "Short secret exponent attack on **Multiprime RSA**," *International Journal of Soft Computing and Engineering*, vol. 3, no. 2, pp. 132–134, 2013.

[13] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public key cryptosystems," *Communications of ACM*, vol. 21, no. 2, pp. 158–164, 1978.

[14] S Sarakar and S Maitra, "Cryptanalysis of **RSA** with two decryption exponents," *Information Procesing Letters*, vol. 110, pp. 178–181, 2010.

[15] A. Takayasu and N. Kunihiro, "Cryptanalysis of **RSA** with multiple small secret exponents," in *Information Security and Privacy*, LNCS 8544, pp. 176–191, Springer, 2014.

[16] M. Wiener, "Cryptanalysis of short **RSA** secret exponents," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 553–558, 1990.

[17] H. Zhang and T. Takagi, "Improved attacks on **Multiprime RSA** with small prime difference," *IEICE Transactions on Fundamentals of Electronics*, vol. E.97-A, no. 7, pp. 1533–1541.

[18] H. Zhang and T. Takagi, "Attacks on **Multiprime RSA** wit small prime difference," in *Informtion and Security*, LNCS 7959, pp. 41–56, Springer, 2013.

[19] M. Zheng and H. Hu, "Cryptanalysis of **Prime Power RSA** with two private exponents," in *Science China Information Sciences*, vol. 58, pp. 1–8, 2015.

Kumar R. Santosh is a research scholar in the Department of Information of Technology of MVGR College of Engineering. He published several 10 articles in various International Journals. His research interests are cryptology, computational number theory, algorithms.

Challa Narasimham is a Professor in the Department of Computer science and Engineering of Vignan Engineering College, Warangal. His research interests include grid computing, cryptology, cloud computing. He has published more than 50 publications in Various National, International Journals.

Pallam Shetty is a Professor in the Department of Computer Science and Systems Engineering of AU College of Engineering, Andhra University. He has guided 3 PhDs and right now, he is guiding 18 PhD scholars. He has guided more than 100 MTech projects. He received many honors, received Gold Medal for securing first rank in MSc. He has been the Resource person for various organizations. He visited Berkley University, USA. He is a Life Member of IETE. He is the Editorial Board Member of various journals like IJCA, IJCIIS, and IJSAT etc..

Classification of Breast Cancer Using Softcomputing Techniques

Ibrahim M. El-Hasnony, Hazem M. El-Bakry, Ahmed A. Saleh
(Corresponding author: Hazem M. El-Bakry)

Information Systems Department, Faculty of Computers and Information Sciences, Mansoura University
Mansoura, Egypt
(Email: helbakry5@yahoo.com)
(Received Dec. 25, 2015; revised and accepted Jan. 10 & Jan. 20, 2016)

Abstract

Cancer is a major issue that has a great deal with the entire world. It is a fatal disease that affects the lives of numerous people and will keep on influencing the lives of some more. It is important to develop reliable and precise system for diagnosing the benign or malignant breast cancer. In this paper an effective hybrid system for breast cancer classification is presented. The proposed system combines K-means clustering algorithm, fuzzy rough feature selection (FRFS), and discernibility nearest neighbor (D-KNN) classifier. Compared to other studies in the breast cancer literature, it is proven that the proposed model outperforms other techniques with accuracy up to 98.9%.

Keywords: Data mining, discernibility K-nearest neighbor, fuzzy rough feature selection, K-means algorithm

1 Introduction

Today, breast cancer [5] is considered the second essential cause for women deaths that suffer from cancer. Early diagnosis for breast cancer is considered an important issue as it is one of the essential factors for its treatment. So there is a need to develop modern and effective systems to deal with this disease and the speed of its diagnosis. Data mining [8] is an important stage of knowledge discovery for finding and extracting patterns from numerous, noisy and incomplete data, Figure 1. Handling data mining tasks is a very difficult job for humans especially when they have to deal with a huge amount of data. Machine learning [13] concept is a branch of artificial intelligence that is interested in automating the system components. Machine learning algorithms make use of and learn from data given to it. They are very helpful in dealing with massive data like the case in medical fields.

Data Medical mining is one of the most critically important issues that need attention and better performance with respect to accuracy and speed factors. Therefore, there were many studies in this field and it still need lots of research because of the explosive growth in data and its diversity. The research area in the medical field is covered in large targets such as the diagnosis of some endemic disease, which afflicts various parts of the world. Also, it addressed how to predict occurrence of a particular disease based on a set of characteristics that may be present in the medical data. Moreover, suggesting a drug or therapeutic treatment for some diseases according to similar treated cases from that disease.

Data mining provides various techniques for efficient and effective data analysis for medical data sets. This paper is interested in preparing data or putting data in a format that facilitates the process of classification of medical data with efficient and more accurate processing with simple and faster classification algorithm. Data pre-processing is one of the most important tasks of data mining. It overcomes the data problems that result from human errors, noise and missing data values that may cause obstacles affecting the analysis process.

The rest of the paper is organized as follows: Section 2 is a review on the most related previous studies in using data mining techniques and machine learning techniques for medical applications and classification of breast cancer data set. Section 3 describes the techniques and methodologies used in this study such as data discernibility K-nearest neighbor classifier (D-KNN), K-means clustering algorithm and fuzzy rough feature selection algorithm (FRFS). Section 4 presents an overview on the proposed model and its components. The output of the experimental results and conclusion will present in sections 5 and 6 respectively.

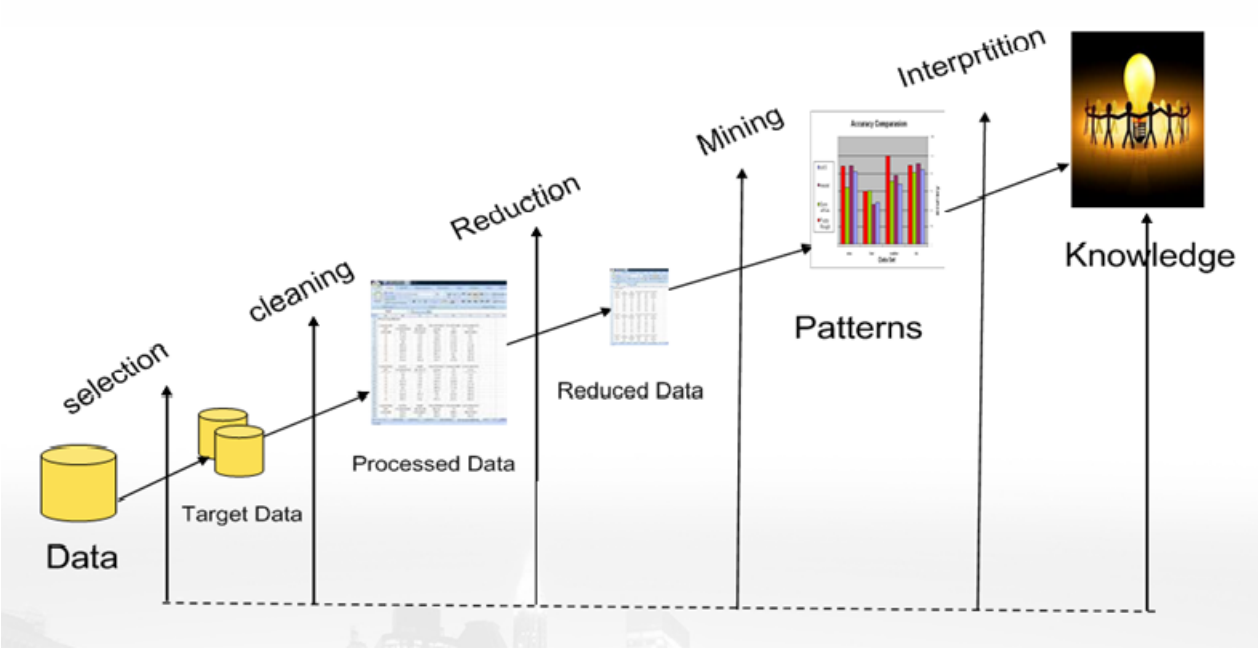

Figure 1: Knowledge discovery process

2 Related Work

There are many studies that worked on breast cancer classification and related techniques used in this paper.

Joshi et al. [11] proposed a model for classifying breast cancer data and used K-means and FF algorithm for faster diagnosis. The proposed model achieved better results with given data.

Banu and Gomathy [1] made a prediction system that depends on data pre-processing to make medical mining more efficient. After pre-processing, data clustered using k-means and then they begin to produce maximal frequent patterns in heart disease dataset using maximal frequent item sets algorithm (MAFIA (. Finally made classification to produced frequent patterns using C4.5 algorithm with information entropy concept

Bichen et al. [21] developed a hybrid system composed of K-means and support vector machine for breast cancer diagnosis. Data grouped from UCI to Breast Cancer. The hybrid system achieved good classification with accuracy up to 97.38%.

Nguyen et al. [16] designed a system for medical diagnosis by combining fuzzy standard additive model (SAM) with wavelet features. Comparing results from applying fuzzy SAM with wavelet for data reduction on breast cancer and heart disease with probabilities neural network (PNN), support vector machine (SVM), fuzzy Adaptive resonance theory (ARTMAP) and adaptive neuro-fuzzy inference system (ANFIS). The system proved higher accuracy in medical diagnosis compared to other techniques.

Ibrahim et al. [6] proposed a comparative study among different feature reduction techniques for evaluating the most reasonable for medical data. The study included the comparison among fuzzy rough feature selection (FRFS), rough set attribute reduction (RSAR), gain ratio, principal components analysis (PCA), and correlation feature selection (CFS). The results showed that FRFS outperforms other techniques over classification algorithms accuracy.

3 Materials and Methods

3.1 Methodology

In this paper an efficient and effective hybrid classifier for breast cancer data is proposed. The proposed model consists of two main steps. The first step is the preprocessing that combines data cleaning, data clustering and feature reduction. The second step uses D-KNN classifier for the resulted data in the first step.

3.2 Data Pre-processing

Data preprocessing [8] transforms data from low-quality to an acceptable one that will lead to high-quality mining results. The proposed model shows how the efficient preprocessing stage for medical data can be developed to achieve high-quality of data. Hence, it leads to good classification results and improves mining process efficiency. What meant by data quality is to achieve accuracy, consistency, completeness, timeliness, believability, and to be understandable. There are many forms of data preprocessing tasks such as data cleaning, data reduction and data integration. Data cleaning [2] handle missing values, outliers' recognition and deletion, noise smoothing, and determining inconsistencies. Data reduction produces a small version of data set that achieves almost equivalent analytical results. Data integration is important and useful when the data merged from different and multiple data sources.

3.2.1 Data Clustering

Data clustering [9, 20] is the task of data mining for dividing data into groups that called clusters. Data in the same cluster are very similar while data in different cluster are dissimilar depending on the values of attributes. The object's similarities are depicted frequently using distance measures. The k-means algorithm is an unsupervised mining clustering technique. K-means is widely applied in bioinformatics and related fields that need to determine the number of clusters that are appropriate for specific problems [8].

3.2.2 Fuzzy Rough Feature Selection

Fuzzy Rough Feature Selection (FRFS) is a hybrid feature selection and data reduction method for combining rough set for calculating and determining the attributes dependencies with no loss in its corresponding data sets [10, 7].

FRQuickReduct [7] used for FRFS implementation and used Equation (1),(dependency equation), for calculating the degree of membership dependency. This membership dependency is between the equivalence classes and its fuzzy attributes. For each attribute that expands the total dependency is appended to the core attribute set.

$$r'_A(D) = \frac{\sum_{\chi \in U} \mu_{POS_A(D)}(\chi)}{|U|}, \tag{1}$$

where D is for the equivalent classes set on the universe U, A is a specific attribute in the set of fuzzy rough attributes.

The equation of dependency takes the degree of the membership of an object x to the fuzzy positive region of the fuzzy rough attribute A as a parameter which is:

$$\mu_{POS_A(D)}(\chi) = \sup_{\chi \in U|D} \min(\mu_{F_i}(\chi), \mu_{POS_A(F_i)}(\chi)), \tag{2}$$

where F_i is the fuzzy attribute subsets in D that is the equivalent classes; $\mu_{F_i}(\chi)$ is the degree of the membership of object x in F_i; $\mu_{POS_A}(F_i)$ indicates the fuzzy positive region corresponding to a fuzzy equivalence class $F_i \in U/A$ which can be displayed as:

$$\mu_{POS_A}(F_i) = \sup_{\chi \in U|D} (\mu_\chi(F_i)). \tag{3}$$

$\mu(\chi)$ is label for the degree of the membership for the subset fuzzy attributes, (fuzzy membership function). The corresponding process is iterative for every attribute and the attribute that has the largest dependency is appended to the core set. This repeats until no more attributes expanding the dependency is preceded.

3.3 Data Classification

Classification [17] is the data mining task that allocates every record in the data set to one of the few predefined classes. Data set is divided into training and test sets. Training set has a known class labels while the test set labels are unknown. Classification is likewise characterized as the task of target function (classification model) learning for mapping each attribute set to its corresponding class label. There are numerous classification algorithms such as naive Bayesian, neural networks, rule-based classifiers, decision tree, support vector machine, and K-nearest neighbor classifiers.

3.3.1 Discernibility K-nearest Neighbor

For each classification technique, there exists a learning algorithm to define the best model for the relationship between the conditional attributes and class label. Some algorithms make the decision every time they classify data by contrasting the test set with the training set. These classifiers are called instance-based learning algorithms.

The K-nearest neighbor (KNN) [14] is an instance based classifier and is utilized mainly for pattern recognition. It assumes that the class label for each instance is the same for its nearest neighbor. KNN is one of most simple and easy to use classification techniques. There are many variations for KNN because of its simplicity that help to enhance its predictive accuracy. KNN example is shown in Figure 2.

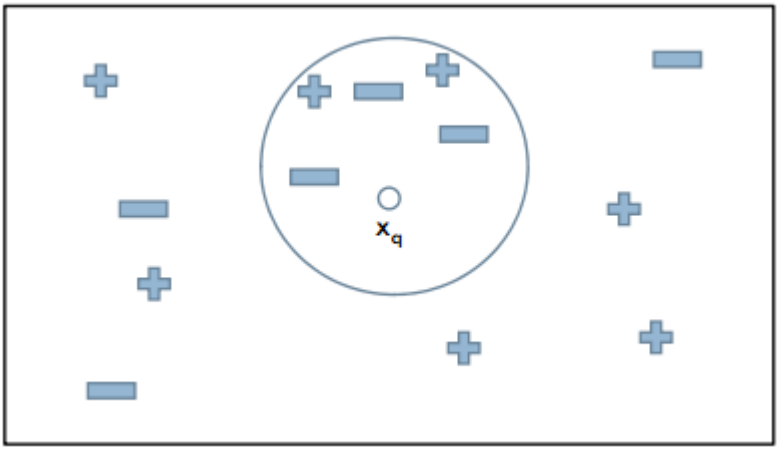

Figure 2: A KNN example

As shown in Figure 2, if K=5, then the instance xq classified as negative.

There are many variations [19] for KNN such as The Density Based kNN Classifier (DB-kNN), The Variable k Nearest Neighbor Classifier (VkNN), Class Based K-nearest neighbor (CB-kNN), The Weighted K-nearest neighbor Classifier (W-kNN), and The Discernibility kNN Classifier (D-kNN).

D-kNN has a structure that considers similar to the original kNN extension of the DB-kNN. The goal of this algorithm is to make a fast classifier with no loss in the accuracy. From the previous studies, the time for D-KNN is less than other classifiers time complexity.

D-kNN [4] considers all neighbors distances and their discernibility. Unlike the structure density, it utilizes the discernibility of each element. For every neighbor, a score is created. The instance score is the discernibility divided by the corresponding distance. Then, a unique classification score for each class is produced by taking the average of all scores for the class. The class with the highest score is chosen for classification. The discernibility for every instance is calculated by taking a radius around each instance (the average between all instances with similar class labels and that instance).

4 The Proposed Model

The proposed model consists of:

A - Data preprocessing

1) Missing values handling;
2) Clustering using K-means algorithm;
3) Feature reduction by FRFS;
4) Merging the reduced clustered data.

B - Classification using D-KNN;

C - Performance evaluation.

Data preprocessing [12] is one of the most essential stages for effective data mining. After missing values are handled, the data are reduced to a set of features which the classification model entirely depends on. It is clear that the collected data are not initially equipped for the mining purposes. Some of the features or attributes are insignificant for classification and may be redundant. Classification quality has many factors and feature reduction is considered a critical issue for its improvements.

Gathering similar instances in the same pattern help the feature reduction process to retrieve the most significant attributes. Clustering [3] using K-means algorithm is a good method because of its simplicity and hidden patterns recognition from data set. K-means clustering algorithm showed in Figure 3.

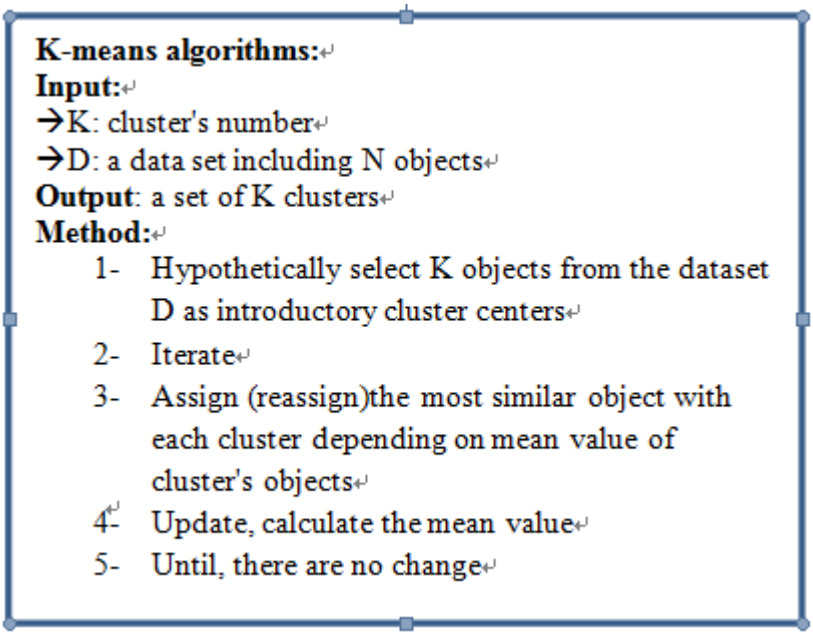

Figure 3: K-means algorithm

The clusters are then passed to the feature reduction algorithm to extract the features which the decision depends on. Fuzzy rough feature selection (FRFS) [15] is utilized for feature reduction because of its advantages to handle noisy, discrete and continuous data with no loss. FRFS quick reduct for FRFS automation displayed in Figure 4.

The significant attributes in each cluster are merged together, without repetition, to form the final reduct used later in the classification process. Figure 5 shows the main steps involved in the proposed model.

5 Experimental Results

5.1 Data Set

Data set used from UCI machine learning repository [18] for examining the proposed model is breast cancer. Table 1 shows a description for Breast cancer data sets.

Table 1: Breast cancer data set description

Dataset characteristics	Multivariate	Attributes	10
Attribute characteristics	Integer	Instances	699
Missing values	Yes	Class	2

Breast cancer data set contains 699 cases for patients who had experienced surgery for breast cancer. The yield values are either 2 or 4 demonstrating that resting tumor protuberance (benign) or risky bump (malignant). Nine different fields are esteemed from 1 to 10 in addition to ID number. The undertaking is to figure out whether the identified tumor is benign (2) or malignant (4) given estimations of nine characteristics.

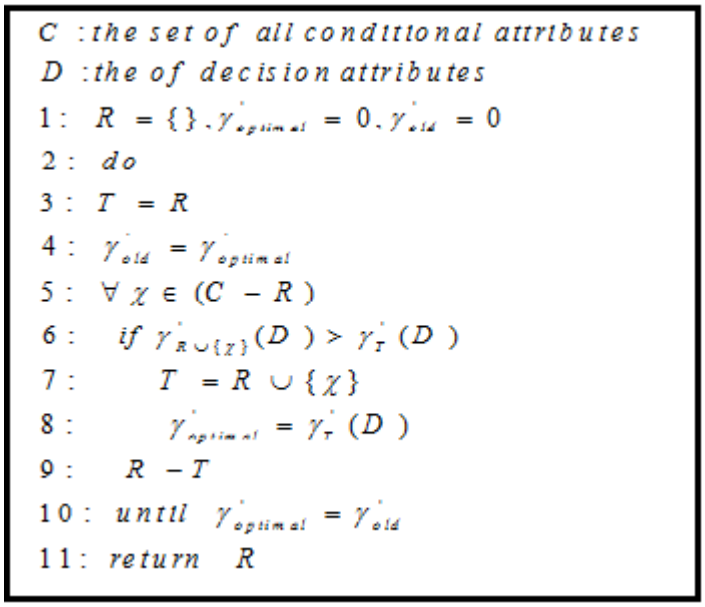

Figure 4: FRQuickReduct algorithm

5.2 Model Results

First, the data set is preprocessed for noisy and missing values using WEKA and then rapid miner used for clustering utilizing the K-means algorithms with K=2. Clustering process shown in Figure 6.

Second, the clustered data are reduced by FRFS using WEKA then the reduced features are merged to produce a new data set.

Finally, Discernibility k-nearest neighbor Classifier (D-kNN) is utilized for classification.

The proposed system classifies new instances of breast cancer test set with accuracy up to 98.9%. Compared to previous studies on the same data set (breast cancer); the proposed system approved its efficiency for increasing the accuracy. Table 2 shows the proposed model accuracy for classifying breast cancer data and other techniques used for the same purpose.

Table 2: Comparison between proposed and previous studies on breast cancer data set

Classification models	Accuracy
ANFIS	96.59
Fuzzy-MLP	96.3
SVM	95
SOFM+PGA	94.2
k-means +SVM	97.38
SOFM +Fuzzy Rough	97
K-means +FRFS+ discernibility NN	98.9

The proposed model outperforms the previous studies as displayed in Figure 7. The higher performance of the proposed model is due to the effective preprocessing step combined with simple, easy to use and fast classifier.

6 Conclusion

An effective classifier for breast cancer has been presented. The proposed model has composed of the intelligent soft-computing techniques (handling missing values + data clustering using K-means algorithm + feature reduction using

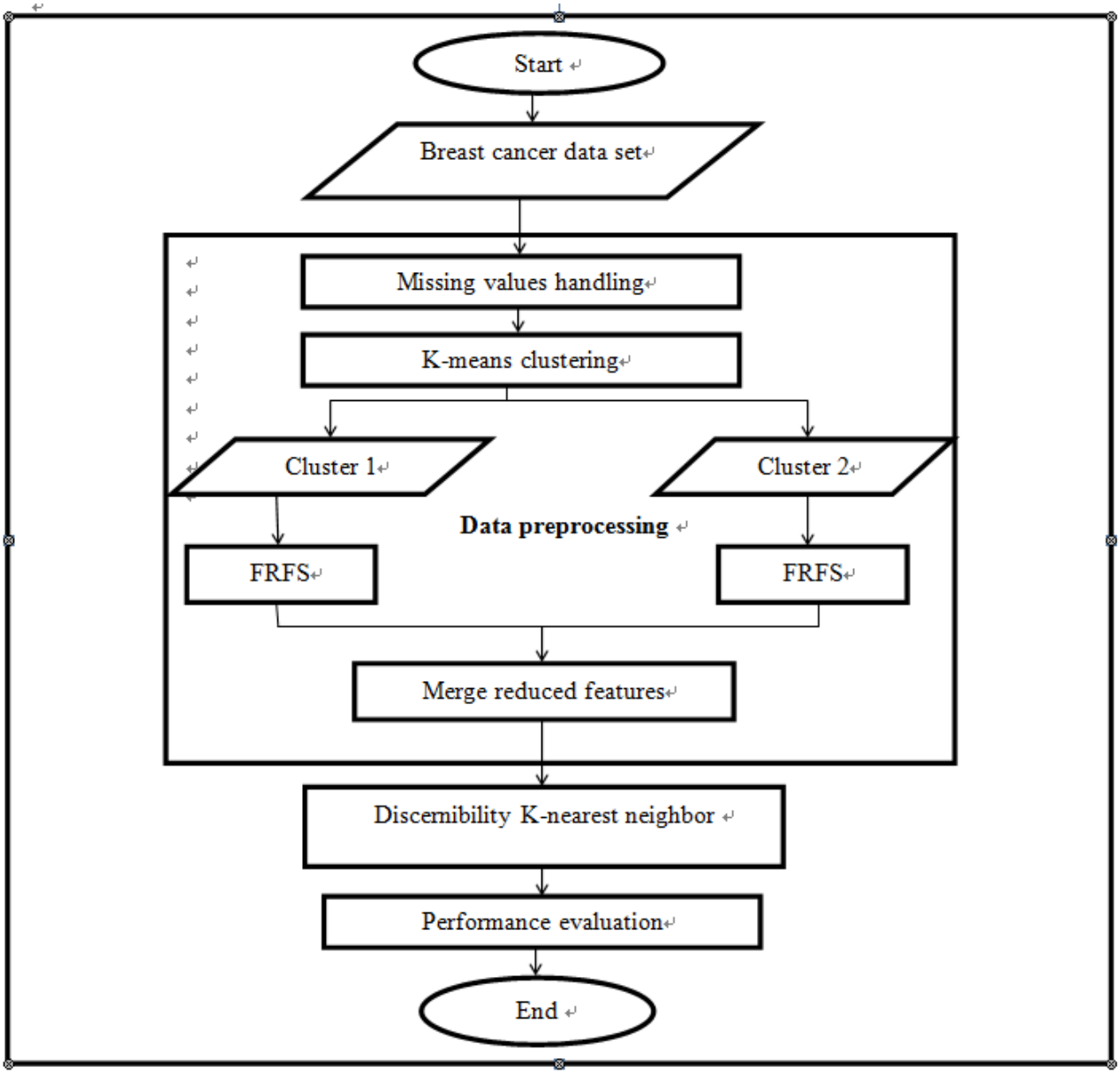

Figure 5: The proposed model architecture

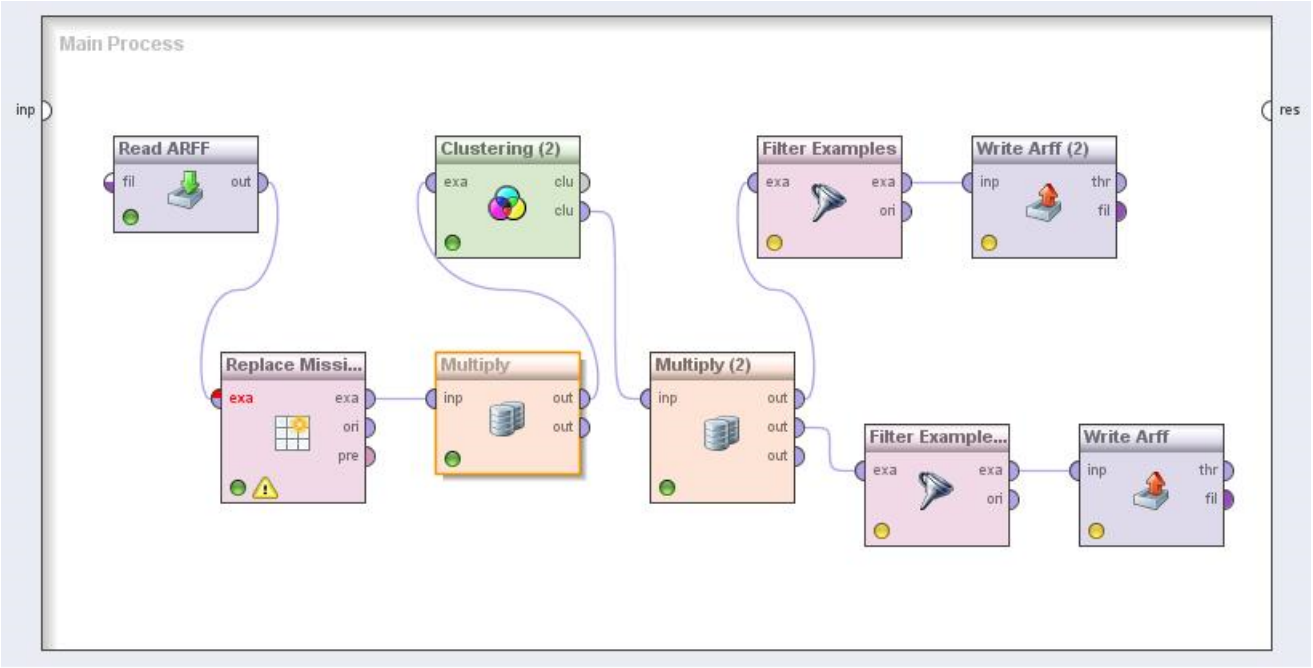

Figure 6: Clustering process

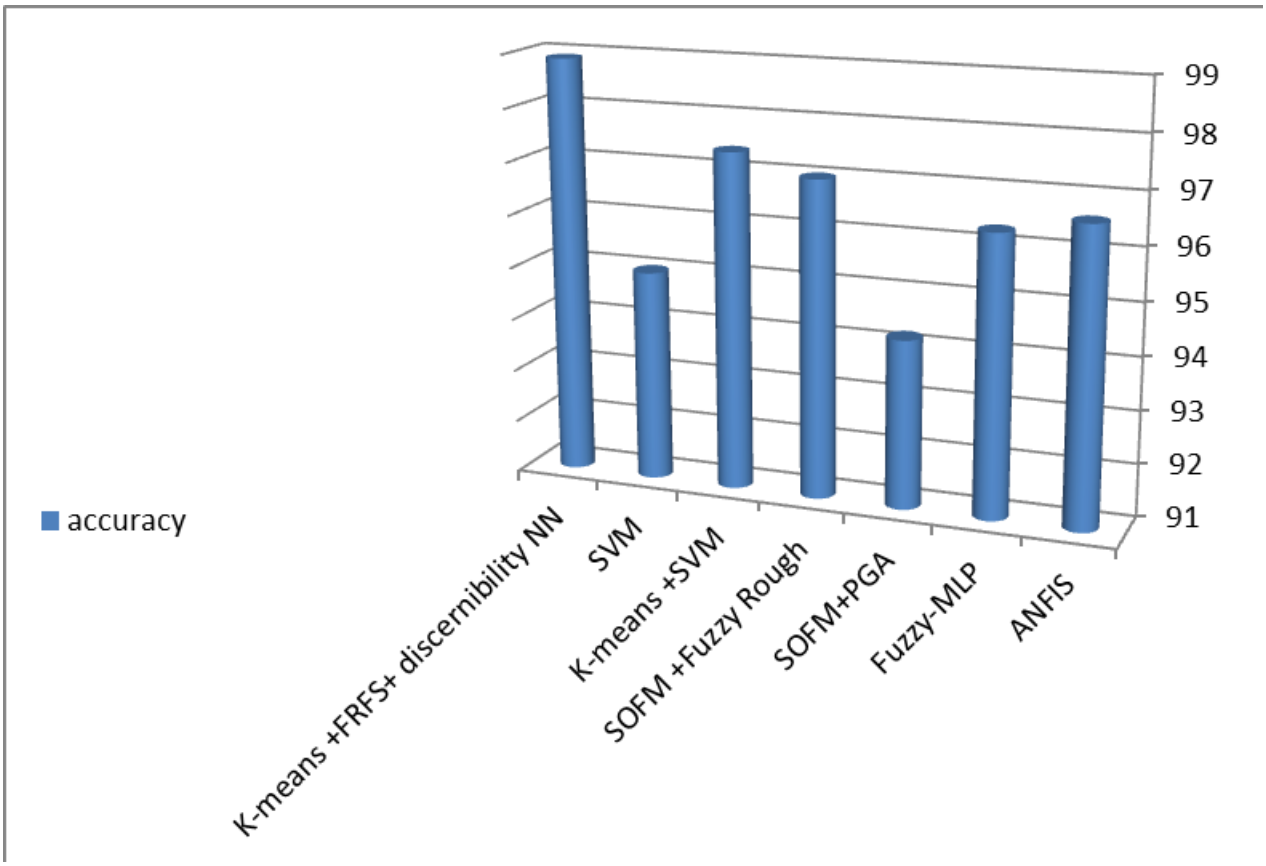

Figure 7: Proposed system vs. previous studies

FRFS + data classification using D-KNN for the merged reduced features). Such operations have been performed to achieve higher accuracy and better classification of breast cancer data. The proposed model has been compared to more than one study that is interested in classifying the same data. It has been proven that the proposed model outperforms other related techniques.

References

[1] M. A. Banu, B. Gomathy, "Disease forecasting system using data mining methods," in *IEEE International Conference on Intelligent Computing Applications (ICICA'14)*, pp. 130–133, 2014.

[2] L. Bertossi, S. Kolahi, and L. VS Lakshmanan, "Data cleaning and query answering with matching dependencies and matching functions," *Theory of Computing Systems*, vol. 52, no. 3, pp. 441–482, 2013.

[3] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert Systems with Applications*, vol. 40, no. 1, pp. 200–210, 2013.

[4] J. Derrac, N. Verbiest, S. Garc?a, C. Cornelis, F. Herrera, "On the use of evolutionary feature selection for improving fuzzy rough set based prototype selection," *Soft Computing*, vol. 17, no. 2, pp. 223–238, 2013.

[5] C. DeSantis, et al., "Breast cancer statistics, 2013," *CA: A Cancer Journal for Clinicians*, vol. 64, no. 1, pp. 52–62, 2014.

[6] I. M. El-Hasnony, H. M. El Bakry, A. A. Saleh, "Comparative study among data reduction techniques over classification accuracy," *International Journal of Computer Applications*, vol. 122, no. 2, pp. 8–15, 2015.

[7] M. Gamal, A. A. El-Fetouh, S. Barakat, "A fuzzy rough rule based system enhanced by fuzzy cellular automata," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 5, pp. 1–11, 2013.

[8] J. Han, M. Kamber, and J. Pei, *Data Mining, Second Edition: Concepts and Techniques*, The Morgan Kaufmann Series in Data Management Systems, 2006.

[9] J. Jacques, C. Preda, "Functional data clustering: A survey," in *Advances in Data Analysis and Classification*, vol. 8, no. 3, pp. 231–255, 2014.

[10] R. Jensen, Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, pp. 824–838, 2009.

[11] J. Joshi, J. P. RinalDoshi, "Diagnosis of breast cancer using clustering data mining approach," *International Journal of Computer Applications*, vol. 101, no. 10, pp. 13–17, 2014.

[12] F. Kamiran, T. Calders. , "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[13] H. C. Koh, G. Tan, "Data mining applications in healthcare," *Journal of Healthcare Information Management*, vol. 19, no. 2, pp. 56, 2011.

[14] F. Moreno-Seco, L. Mico, J. Oncina, "A modification of the LAESA algorithm for approximated k-NN classification," *Pattern Recognition Letters*, vol. 24, no. 13, pp. 47–53, 2003.

[15] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in neuroimaging," *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, 2014.

[16] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, "Medical diagnosis by fuzzy standard additive model with wavelets," in *IEEE International Conference on Fuzzy Systems*, pp. 1937–1944, 2014.

[17] L. Tao, F. Sun, and S. Yang, "A fast and robust sparse approach for hyper spectral data classification using a few labelled samples," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 6, pp. 2287–2302, 2012.

[18] UCI, Donald Bren School of Information and Computer Sciences, University of California, Irvine, July 20, 2015. (`http://www.ics.uci.edu`)

[19] H. Wang, D. Bell, "Extended k-Nearest neighbors based on evidence theory," *The Computer Journal*, vol. 47, no. 6, pp. 662–672, 2004.

[20] I. Yoo, et al. , "Data mining in healthcare and biomedicine: A survey of the literature," *Journal of Medical Systems*, vol. 36, no. 4, pp. 2431–2448, 2012.

[21] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014.

Ibrahim M. El-Hasnony is demonstrator at the Faculty of Computer Science and Information Systems - Mansoura University - Egypt. His research interests include Data mining, Artificial intelligence and Machine learning.

Hazem M. El-Bakry (Mansoura, EGYPT 20-9-1970) received B.Sc. degree in Electronics Engineering, and M.Sc. in Electrical Communication Engineering from the Faculty of Engineering, Mansoura University - Egypt, in 1992 and 1995 respectively. Dr. El-Bakry received Ph. D degree from University of Aizu - Japan in 2007. Currently, he

is associate professor at the Faculty of Computer Science and Information Systems - Mansoura University - Egypt. His research interests include neural networks, pattern recognition, image processing, biometrics, cooperative intelligent systems and electronic circuits. In these areas, he has published many papers in major international journals and refereed international conferences. According to academic measurements, now the total number of citations for his publications is 2757. The H-index of his publications is 28. Dr. El-Bakry has the United States Patent No. 20060098887, 2006. Furthermore, he is associate editor for journal of computer science and network security (IJCSNS) and journal of convergence in information technology (JCIT). In addition, is a referee for IEEE Transactions on Signal Processing, Journal of Applied Soft Computing, the International Journal of Machine Graphics & Vision, the International Journal of Computer Science and Network Security, Enformatika Journals, WSEAS Journals and many different international conferences organized by IEEE. Moreover, he has been awarded the Japanese Computer & Communication prize in April 2006 and the best paper prize in two conferences cited by ACM. He has also been awarded Mansoura university prize for scientific publication in 2010 and 2011. Dr. El-Bakry has been selected in who Asia 2006 and BIC 100 educators in Africa 2008.

Ahmed A. Saleh is full professor at the Faculty of Computer Science and Information Systems - Mansoura University - Egypt. He is the vice dean for higher studies, research, and cultural affairs. His research interests include E-Business, GIS, and Information Systems.

Guide for Authors
International Journal of Electronics and Information Engineering

International Journal of Electronics and Information Engineering (IJEIE) will be committed to the timely publication of very high-quality, peer-reviewed, original papers that advance the state-of-the art and applications of Electronics and Information Engineering. Topics will include, but not be limited to, the following: Algorithms, Bioinformatics, Computation Theory, AI, Fuzzy Systems, Embedded Systems, VLSI/EDA, Computer Networks, Information Security, Database, Data Mining, and Information Retrieval, Digital Content, Digital Life, and Human Computer Interaction, Image Processing, Computer Graphics, and Multimedia Technologies, Information Literacy, e-Learning, and Social Media, Mobile Computing, Wireless Communications, and Vehicular Technologies, Parallel, Peer-to-peer, Distributed, and Cloud Computing, Semiconductor, Software Engineering and Programming Languages, Telecommunication, etc.

1. Submission Procedure

Authors are strongly encouraged to submit their papers electronically by using online manuscript submission at http://ijeie.jalaxy.com.tw/.

2. General

Articles must be written in good English. Submission of an article implies that the work described has not been published previously, that it is not under consideration for publication elsewhere. It will not be published elsewhere in the same form, in English or in any other language, without the written consent of the Publisher.

2.1 Length Limitation:

All papers should be concisely written and be no longer than 30 double-spaced pages (12-point font, approximately 26 lines/page) including figures.

2.2 Title page

Title page should contain the article title, author(s) names and affiliations, address, an abstract not exceeding 100 words, and a list of three to five keywords.

2.3 Corresponding author

Clearly indicate who is willing to handle correspondence at all stages of refereeing and publication. Ensure that telephone and fax numbers (with country and area code) are provided in addition to the e-mail address and the complete postal address.

2.4 References

References should be listed alphabetically, in the same way as follows:
For a paper in a journal: M. S. Hwang, C. C. Chang, and K. F. Hwang, ``An ElGamal-like cryptosystem for enciphering large messages," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 2, pp. 445--446, 2002.

For a book: Dorothy E. R. Denning, *Cryptography and Data Security*. Massachusetts: Addison-Wesley, 1982.

For a paper in a proceeding: M. S. Hwang, C. C. Lee, and Y. L. Tang, ``Two simple batch verifying multiple digital signatures," in *The Third International Conference on Information and Communication Security (ICICS2001)*, pp. 13--16, Xian, China, 2001.

In text, references should be indicated by [number].

2.5 Author benefits

No page charge is made.

Subscription Information

Individual subscriptions to IJEIE are available at the annual rate of US$ 200.00 or NT 6,000 (Taiwan). The rate is US$1000.00 or NT 30,000 (Taiwan) for institutional subscriptions. Price includes surface postage, packing and handling charges worldwide. Please make your payment payable to "Jalaxy Technique Co., LTD." For detailed information, please refer to http://ijeie.jalaxy.com.tw or Email to ijeieoffice@gmail.com.