Iraqi Journal of Science, 2020, Vol. 61, No. 12, pp: 3408-3416
DOI: 10.24996/ijs.2020.61.12.29

ISSN: 0067-2904

DNA Encoding for Misuse Intrusion Detection System based on UNSW-NB15 Data Set

Omar Fitian Rashid
Department of Computer Technology Engineering, Al-Hikma University College, Baghdad, Iraq

Received: 29/3/2020 Accepted: 29/5/2020

Abstract

Recent researches showed that DNA encoding and pattern matching can be used for the intrusion-detection system (IDS), with results of high rate of attack detection. The evaluation of these intrusion detection systems is based on datasets that are generated decades ago. However, numerous studies outlined that these datasets neither inclusively reflect the network traffic, nor the modern low footprint attacks, and do not cover the current network threat environment. In this paper, a new DNA encoding for misuse IDS based on UNSW-NB15 dataset is proposed. The proposed system is performed by building a DNA encoding for all values of 49 attributes. Then attack keys (based on attack signatures) are extracted and, finally, Raita algorithm is applied to classify records, either attacks or normal, based on the extracted keys. The results of the current experiment showed that the proposed system achieved good detection rates for all of attacks, which included the Analysis, Backdoor, DoS, Exploits, Fuzzers, Generic, Reconnaissance, Shellcode, and Worms, with values of 82.56%, 92.68%, 75.59%, 75.42%, 67%, 99.28%, 81.02%, 73.6%, 85%, and 90.91%, respectively. The values of false alarm rate and accuracy were equal to 24% and 89.05%, respectively. Also, the execution time for the proposed system was found to be short, where the values of the encoding time and matching time for one record were 0.45 and 0.002 second, respectively.

Keywords: Intrusion detection system, DNA Encoding, Pattern Matching Algorithm, Raita Algorithm

ترميز الحمض النووي لنظام الكشف عن التطفل بالاعتماد على مجموعة بيانات UNSW-NB15

عمر فتيان رشيد
قسم هندسة تقنيات الحاسوب ، كلية الحكمة الجامعة ، بغداد ، العراق

الخلاصة

أظهرت الأبحاث الحديثة أنه يمكن استخدام تشفير الحمض النووي ومطابقة الأنماط في نظام اكتشاف التسلل (التطفل) (IDS) وتم الحصول على نتائج عالية في الكشف عن الهجوم. يتم تقييم أنظمة كشف التسلل هذه استنادًا إلى مجموعات بيانات تم إنشاؤها منذ عقود سابقة. ومع ذلك فان العديد من الدراسات أوضحت أن مجموعات البيانات هذه لا تعكس بشكل شامل حركة مرور الشبكة والهجمات الحديثة ذات التأثير المنخفض ولا تغطي تهديد بيئة الشبكة الحالية. يقترح هذا البحث ترميز حمض نووي جديد لنظام الكشف عن التسلل استنادًا إلى مجموعة بيانات UNSW-NB15. وتم تنفيذ النظام المقترح من خلال بناء ترميز الحمض النووي

Email: omaralrawi08@yahoo.com

لجميع قيم السمات الـ 49 و من ثم استخرجت مفاتيح الهجمات (استنادًا إلى تواقيع الهجمات) وأخيراً طبقت خوارزمية Raita لتصنيف السجلات إما هجوم أو حالة عادية بناءً على المفاتيح المستخرجة. أظهرت النتائج الحالية أن النظام المقترح قد حصل على نتائج جيدة للكشف عن الهجمات المختلفة المتمثلة بـ Analysis و Backdoor و DoS و Exploits و Fuzzers و Generic و Reconnaissance و Shellcode و Worms والمجموع الكلي لهذه الهجمات وأن هذه النتائج تساوي 82.56٪ و 92.68٪ و 75.59٪ و 75.42٪ و 67٪ و 99.28٪ و 81.02٪ و 73.6٪ و 85٪ و 90.91٪ على التوالي. بينما نتائج الإنذارات الكاذبة والدقة تساوي 24٪ و 89.05٪ على التوالي. أيضا ، وكان وقت تنفيذ النظام المقترح سريعا حيث كانت مدة التشفير والمطابقة للسجل الواحد تساوي 0.45 ثانية و 0.002 ثانية على التوالي.

Introduction

Many researches on DNA proposed various methods for intrusion detection systems. These systems are constructed through applying several steps, the main two of which are, firstly, proposing and applying a DNA encoding method that is used to convert plaintext (intrusion detection dataset records) to DNA sequence and, secondly, applying pattern matching algorithms to classify these records as normal or attacks. An anomaly intrusion detection system was introduced by Mahdy and Saeb [1], where the gene idea was used Then, suspicious actions. were detected based on "normal behaviour" and the use of string matching. This system was performed by using two steps; first, the creation of DNA sequence and, second, the control of the network, where a monitoring phase is adopted, the hardware is implemented, and DNA pattern matching is performed. An anomaly IDS was established by Al-Ibaisi *et al*. [2]. They adopted a new signature sequence with the use of a threshold value that is produced in the system training, converting these network connections to DNA sequences. Finally, DNA sequence was arranged in rows to find the similarity degree and to classify the network traffic as either an attack or normal. This method led to new generations to choose the signature that has the best alignment value for normal sequences. This system was carried out by applying two steps; the first step included DNA sequence encoding, where the main idea is using less space and having available nucleotides for any new value of new attacks attributes values. Also, the system divided the network traffic values into either static or dynamic parameters, where the static parameters included flag, protocol, and services, while the dynamic parameters included integer, real, and Boolean. In the second step, the genetic algorithm was used to enhance the selection of the target solution. A misuse IDS based on DNA sequence was built by Hameed and Rashid [3]. Three steps were followed to perform this system. In the first step, the enhanced DNA encoding method proposed by Al-Ibaisi *et al*. [2] is used the same number values to represent both integer and Boolean attributes. Then two characters are used to represent these values instead of the three characters that were used previously. This encoding procedure converts these records to DNA sequences. In the second step, the attack signature keys and their positions are extracted by using Teiresias algorithm. In the last step, these records are categorized into two groups, attacks or normal, based on attack signature keys and their positions by using Horspool algorithm. Rashid *et al*. [4] proposed a novel IDS depending on the concept of DNA sequence. This system is built firstly by converting the network traffic to DNA sequence by using Cryptography encoding idea. Secondly, keys are extracted, and thirdly the Horspool algorithm for matching process is applied. Rashid *et al*. [5] created a DNA encoding idea to convert all record values, by using KDDCup dataset. After that, Teiresias algorithm is applied in order to find the STR sequence. In the testing phase, the same DNA encoding is used for converting; finally, Brute-force algorithm is applied to group the records as normal or attacks.

However, the above techniques are applied based on KDDCUP99, NSLKDD, or both datasets that were produced over twenty years. Many studies showed that these datasets cannot cover the existing network attacks environment [6]. Therefore, the current paper proposed a new DNA encoding method for IDS, based on UNSW-NB15 dataset.

Materials and Methods

Most organisms contain DNA molecules that normally have four chemical bases, namely adenine (A), cytosine (C), guanine (G), and thymine (T) [7]. Many algorithms can search for DNA strings. One algorithm which is very popular is the Raita algorithm. This algorithm works quickly in searching string parts in short or long strings. The searched characters are based on patterns and sequences in a

parent string. Raita algorithm works in reverse, starting from the last part of the character. If the searched character is found, Raita algorithm will search the character from the character in the middle. When searching in the middle of a block, if the character searched has been found, the Raita algorithm will move to other characters. The search will continue from the second character to the character before the last character. Then the search will return from starting the middle character. Raita algorithm has several stages in carrying out its algorithmic process. Precisely, two stages must be performed during the search, which are the pre-processing and searching [8].

The UNSW-NB15 dataset [6] was firstly published in 2015, which contains different modern attacks and real normal activities. This dataset contains nine types of attacks which are Reconnaissance, Shellcode, Exploit, Fuzzers, Worm, DoS, Backdoor, Analysis and Generic, the description of which is shown in Table-1. Also, the network traffic record in this dataset includes 49 features [9]. The UNSW-NB15 dataset has training and testing datasets, both containing normal records and nine types of attack records. These records have 45 features [6].

Table 1- UNSW-NB15 dataset of attack types and their characteristics

Attack	Description
Fuzzers	Aims to stop a program by sending a lot of data.
Analysis	Includes port scan and spam.
Backdoors	Aims to pass system security in order to access the computer.
DoS	Aims to cut host services to make network server unavailable to users.
Exploits	Aims to exploit about the vulnerability in operating system or part of software.
Generic	Is a mechanism that is utilized with all block ciphers.
Reconnaissance	Aims to gather information.
Shellcode	Aims to use a part of code to exploit software vulnerability.
Worms	Aims to replicate itself, then move to different computers.

The proposed system is first constructed by building DNA encoding tables for all attribute values. These attributes are divided into four groups:

1- Protocol attributes: which contain 131 different values, all of which are nominal.
2- Service attributes: which contain 13 different values, all of which are nominal.
3- State attributes: which contain 7 different values, all of which are nominal.
4- Digit attributes: which represent the remaining of the attributes, having numerical values that can be either integer, binary, or float.

Therefore, for nominal attributes with values that are equal to 151 values (the total number of protocol, service, and state attributes), four DNA characters are used that can handle all these values. While for numerical attributes with 11 values (from 0 to 9 and a fraction point), two DNA characters are used that can handle all these values to represent each digit separately. The values of attributes and their equivalent DNA sequences are provided by the current work and are shown in Tables-(2, 3, 4 and 5).

Table 2- Protocol attribute values and their equivalent DNA sequence

Protocol	DNA Seq.	Protocol	DNA Seq.
3pc	AAAC	merit-inp	CGCT
a/n	CCTG	mfe-nsp	CAAC
aes-sp3-d	TTGG	mhrp	GATC
any	GGAT	micp	ATGC
argus	GCTA	mobile	GCAA
aris	ACGC	mtp	AGAC
arp	GGGG	mux	AAGA
ax.25	TGTC	narp	GGAC
bbn-rcc	GCAC	netblt	AGCT
bna	ACAA	nsfnet-igp	ATGT
br-sat-mon	CGTT	nvp	ATAG
cbt	GACA	ospf	TCCT
cftp	TCAC	pgm	TTAT

chaos	AGTG	pim	CTGA
compaq-peer	GTTG	pipe	AGCA
cphb	CTTC	pnni	TAGG
cpnx	CAGA	pri-enc	ATCA
crtp	GACG	prm	ACGT
crudp	TACT	ptp	GGTT
dcn	CCCC	pup	TTTT
ddp	AATA	pvp	CCGG
ddx	ACCA	qnx	GCCG
dgp	GTGA	rdp	GCGG
egp	ATTA	rsvp	GAGT
eigrp	CGAT	rvd	AGCG
emcon	GATT	sat-expak	GTGT
encap	GCAG	sat-mon	GATA
etherip	TAGC	sccopmce	CCCA
fc	CGCG	scps	AGAG
fire	GACC	sctp	ATCT
ggp	GGGT	sdrp	ACCC
gmtp	TTAC	secure-vmtp	TCCG
gre	TTTG	sep	CATT
hmp	GGTC	skip	TGCT
iatp	CTAA	sm	TGAC
ib	GCTT	smp	TACA
idpr	GCCT	snp	TATG
idpr-cmtp	CCAT	sprite-rpc	TTCA
idrp	TTTA	sps	CACC
ifmp	AGAT	srp	TGCG
igmp	TGTA	st2	TTAG
igp	TCTG	stp	CCGA
il	ATGA	sun-nd	TCCA
i-nlsp	CTCT	swipe	TTCC
ip	GCTG	tcf	AAGT
ipcomp	TGAA	Tcp	CGTG
ipcv	CTTT	Tlsp	GCTC
ipip	GTAC	tp++	TACG
iplt	CGGC	trunk-1	GAGC
ipnip	GGCC	trunk-2	CCAA
ippc	ACAC	Ttp	TGGC
ipv6	GTCG	Udp	GTAA
ipv6-frag	CGGG	Unas	CATA
ipv6-no	CTTA	Uti	GTGC
ipv6-opts	CCCT	vines	AAGG
ipv6-route	AGCC	Visa	GGTA
ipx-n-ip	TATT	Vmtp	CCAC
irtp	GTTA	Vrrp	GGCG
isis	GTGG	wb-expak	GAAA
iso-ip	TTAA	wb-mon	GGTG
iso-tp4	TCGC	Wsn	CACG
kryptolan	CAAT	Xnet	GTTT
l2tp	TCCC	xns-idp	ATTG
larp	CGCA	Xtp	CGTC
leaf-1	TCTC	Zero	GAAG
leaf-2	GTCA		

Table 3- Service attribute values and their equivalent DNA sequence

Service	DNA Seq.	Service	DNA Seq.
-	CACT	pop3	AGAA
dhcp	CTGC	radius	TCAG
dns	TGGG	smtp	CGGT
ftp	TTCT	snmp	AAAG
ftp-data	TGTG	Ssh	TGGA
http	AGTA	Ssl	AATT
irc	ACGG		

Table 4- State attribute values and their equivalent DNA sequence

State	DNA Seq.	State	DNA Seq.
ACC	AATC	INT	CAGC
CLO	GAAT	REQ	AGTC
CON	AGGC	RST	ATGG
FIN	GCGT		

Table 5- Digit attributes values and their equivalent DNA sequence

Digit	DNA Seq.	Digit	DNA Seq.
0	AC	6	AG
1	GT	7	CA
2	GC	8	TG
3	CC	9	TT
4	CT	.	TC
5	CG		

An example of converting a UNSW-NB15 dataset record to DNA sequences based on the building tables is given as follows:

Record: 47933,0.000009,unas,-,INT,2,0,200,0,111111.1072,254,0,88888888,0,0,0, 0.009,0,0,0,0, 0,0,0,0,0,0,100,0,0,0,8,2,5,4,4,8,0,0,0,13,8,0

DNA sequences: CTCATTCCCCACTCACACACACACTTCATACACTCAGCGCAC GCACAC ACGTGTGTGTGTGTTCGTACCAGCGCCGCTACTGTGTGTGTGTGTGT GACACACACTCACACTTACACACACACACACACACACACGTACACACACACTGGCCGCTCTT GACACACGTCCTGAC

When the conversion is finished, the next step is to extract the keys found in the attack records and, at the same time, not found in the normal records. This is achieved by dividing training dataset records, after converting them to DNA sequences, into blocks each with a length of 5 DNA characters. Then, the blocks found in both attack and normal records are removed and, finally, the blocks with large repetitions are selected. These keys are shown in Table-6.

Table 6- The extracted keys for Misuse IDS

Number	Keys
1	CAACA
2	CATTC
3	GGCGT
4	CTTTA

After that, random records from UNSW-NB15 testing dataset are used. These records are first converted to DNA sequence based on the building tables mentioned before. Then, Raita algorithm is

applied to classify these records, as normal or attack, based on keys (i.e. if one of these keys is found then this record is an attack record, otherwise it is a normal record).
An example of the application of Raita algorithm, that is used to find the first key (CAACA) in the sequence "CTGGACAACATGGA" is shown in Table- 8. But, before that, a shift table for these keys must be built (Table-7). The search procedure is applied and repeated for all four keys.

Table 7- Shift tables for the extracted keys

Character	**A**	**C**	**G**	**T**
Key 1 - Shift	2	1	5	5
Key 2 - Shift	3	4	5	1
Key 3 - Shift	5	2	1	5
Key 4 - Shift	5	4	5	1

Table 8- Example of the application of Raita algorithm

1st													
C	T	G	G	A	C	A	A	C	A	T	G	G	A
2		3		1									
C	A	A	C	A									
Move key based on value of character (A) that is equal to 2													
2nd													
C	T	G	G	A	C	A	A	C	A	T	G	G	A
		2				1							
		C	A	A	C	A							
Move key based on value of character (C) that is equal to 1													
3rd													
C	T	G	G	A	C	A	A	C	A	T	G	G	A
			2				1						
			C	A	A	C	A						
Move key based on value of character (C) that is equal to 1													
4th													
C	T	G	G	A	C	A	A	C	A	T	G	G	A
								1					
				C	A	A	C	A					
Move key based on value of character (C) that is equal to 1													
5th													
C	T	G	G	A	C	A	A	C	A	T	G	G	A
					2	4	3	5	1				
					C	A	A	C	A				
Where the red color in the table indicates mismatch, while the green color indicates match.													

Results and Discussion

The current proposed system uses UNSW-NB15 dataset as a source of information; this is a new dataset that is used to calculate the performance of the IDS and includes 9 various attacks. This dataset has two datasets: training and testing. Firstly, the attack keys are extracted based on the training dataset. Secondly, the proposed system performance is calculated based on 4000 random records of the testing dataset. The experimental environment involves the operating system of Microsoft Windows 10 Professional, a CPU which is Intel 2.50GHz, and memory of 4.00 GB.

The performance of the proposed system is determined by five measurements, the first of which is based on detection rate (DR), calculated as in the following [10]:

$$DR = \frac{TP}{TP + FN} \quad \ldots\ldots\ldots (1)$$

The second and third measurements are based on false alarm rate (FAR) and accuracy, respectively, calculated as shown below [10]:

$$FAR = \frac{FP}{TN + FP} \quad \ldots\ldots (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad \ldots\ldots\ldots (3)$$

While the fourth measurement is based on the encoding time, which is the time needed to convert the record to DNA sequences. The fifth measurement is the matching time, which is the time needed to classify the testing record. Either as attack or normal, based on the extracted keys. The equation used to calculate time is:

$$Time = Encoding\ time + Matching\ time \quad \ldots (4)$$

The DR results for various attack types and the DR result for all attacks (described previously in Table-1) are shown in Table-9 and Figure-1.

Table 9- DR results for different attack types

Attack	DR result (%)
Analysis	82.56
Backdoor	92.68
DoS	75.59
Exploits	75.42
Fuzzers	67
Generic	99.28
Reconnaissance	81.02
Shellcode	73.6
Worms	85
All attacks	90.91

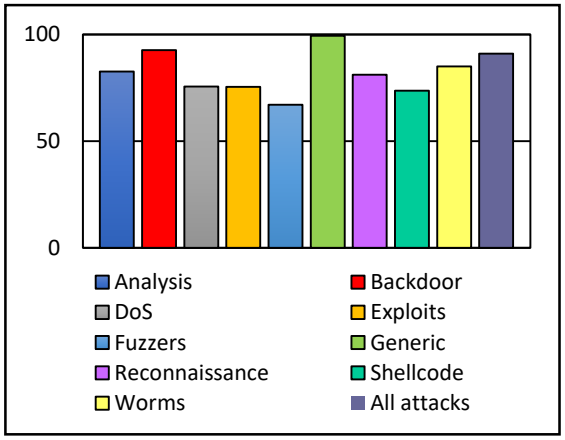

Figure 1- DR results for different attacks types

As outlined in Table-9 and Figure-1, the best achieved detection rate is obtained for the Generic attack type (99.28%).
The DR, FAR, and accuracy results for all attacks are shown in Table-10 and Figure-2.

Table 10- DR, FAR, and accuracy results that achieved by the proposed IDS

Measure	Result (%)
DR	90.91%
FAR	24%
Accuracy	89.05%

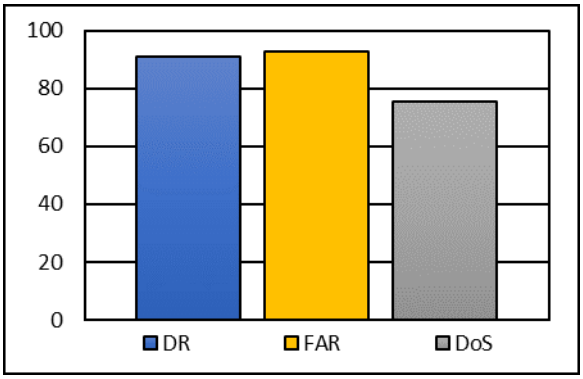

Figure 2- DR, FAR, and accuracy results that achieved by the proposed IDS

As outlined in Table-10 and Figure-2, the achieved detection rate is equal to 90.91%, while FAR value is equal to 24%, and the accuracy value is equal to 89.05%.

. From the results shown in Table-11, it is clear that the values of encoding and matching times obtained by the proposed system are equal to 0.45 and 0.002 seconds, respectively, for one record. This indicates that the time values obtained by the proposed method are very short. In addition, the values of the encoding and matching times for one record and for all of the 4000 records are illustrated in Figures-(3 and 4), respectively.

Table 11- encoding time and matching time needed for converting and classify network traffic records

	Time (seconds)
Matching time for all 4000 records	1814
Encoding time for all 4000 records	8
Total time for all 4000 records	1822
Matching time for one record	0.45
Encoding time for one record	0.002
Total time for one record	0.452

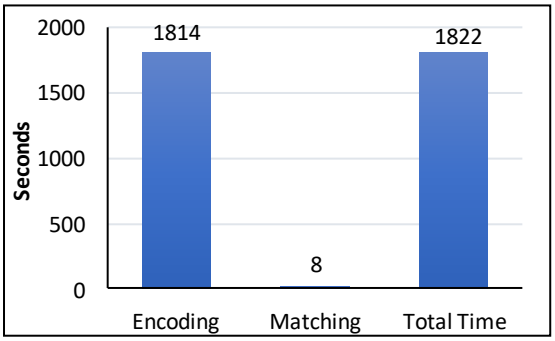

Figure 3- Time needed to converting and matching all testing records

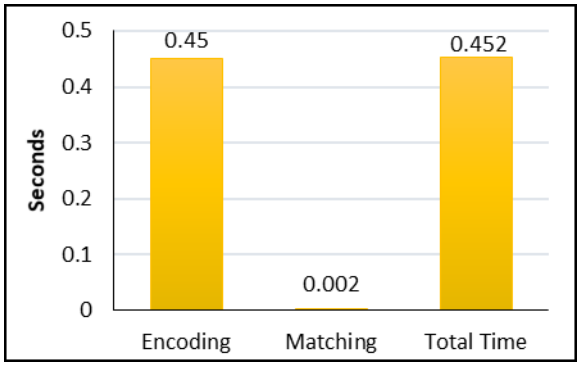

Figure 4- Time needed to converting and matching one testing record

Conclusions

The present paper exhibited a new DNA encoding method for an intrusion detection system through the application of Raita algorithm on UNSW-NB15 dataset, as a recent dataset which covers a set of attack types. High detection rate values are achieved for nine attack types. The obtained DR, FAR, and accuracy values are equal to 90.91%, 24%, and 89.05%, respectively. The values of encoding and matching times for one record are equal to 0.45 and 0.002 seconds, respectively. The findings confirm the high efficiency of the proposed method.

References

1. Mahdy, R. and Saeb, M. **2007**. Design and implementation of an anomaly-based network intrusion detection system utilizing the DNA model. Proceeding of the 9th WSEAS Int. Conference on Data Networks, Communications, and Computers.
2. Al-Ibaisi, T., Abu-Dalhoum, A., Al-Rawi, M., Alfonseca, M. and Ortega, A. **2008**. Network intrusion detection using genetic algorithm to find best DNA signature. *Wseas Transactions on Systems*, **7**(7): 589-599.
3. Hameed, S. M. and Rashid, O. F. **2014**. Intrusion detection approach based on DNA signature. *Iraqi Journal of Science*, **55**(1): 241-250.
4. Rashid, O. F., Othman, Z. A. and Zainudin, S. **2017**. A novel DNA sequence approach for network intrusion detection system based on cryptography encoding method. *International Journal on Advanced Science Engineering and Information Technology*, **7**(1): 183-189.
5. Rashid, O. F., Othman, Z. A. and Zainudin, S. **2019**. Four Char DNA Encoding for Anomaly Intrusion Detection System. Proceedings of the 2019 5th International Conference on Computer and Technology Applications: 86–92.
6. Moustafa, N. & Slay, J. **2015**. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). 2015 Military Communications and Information Systems Conference (MilCIS).
7. Soram, R. and Khomdram, M. **2010**. "Biometric DNA and ECDLP Based Personal Authentication System: A Superior Posse of Security", IJCSNS *International Journal of Computer Science and Network Security*, **10**(1), January 2010.
8. Hudaa, S., Nguyen, P. H., Lestari, S. P., Gunawan, G. and Supiyandi, S. **2020**. Data Search using Raita Algorithm. *Journal of Critical Reviews*, **7**(1): 72-75.
9. Janarthanan, T. and Zargari, S. **2017**. Feature Selection in UNSW-NB15 and KDDCUP'99 datasets. 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE).
10. Wu, S. and Benzhaf, W. **2010**. The use of computation intelligence in intrusion detection systems. *Applied Soft Computing*, **10**(1): 1-35.