Radha Damodaram, Dr.M.L.Valarmathi / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 1, Jan-Feb 2012, pp. 870-876

Phishing website detection and optimization using Modified bat algorithm

Radha Damodaram,M.C.A,M.Phil.*, Dr.M.L.Valarmathi **
Asst. Professor, Department of BCA, SS & IT,CMS College of Science & Commerce, Coimbatore.
Associate Professor, Dept. of Computer Science & Engg, Government College of Technology, Coimbatore.

Abstract:
This paper presents an approach to overcome the difficulty and complexity in detecting and predicting phishing websites. Existing system is an intelligent resilient and effective model that is based on using association and classification Data Mining algorithms. These algorithms were used to characterize and identify all the factors and rules in order to classify the phishing website and the relationship that correlate them with each other also compared their performances, accuracy, number of rules generated and speed. Even though the rules generated from the associative classification model showed the relationship between some important characteristics like URL and Domain Identity, and Security and Encryption criteria in the final phishing detection rate, there is no optimal solution. In proposed system we introduced MBAT a metaheuristic algorithm to get an optimal solution for the search of fake websites. We also compare the proposed algorithm with other existing algorithms, including Ant Colony Optimization and Particle Swarm Optimization.

Keywords: ***Modified BAT, Echolocation, Velocity, loudness, frequency.***

1. INTRODUCTION:

1.1 Phishing

Phishing is a way of attempting to acquire sensitive information such as usernames, passwords and credit card details by masquerading as a trustworthy entity in an electronic communication. This is similar to *Fishing*, where the fisherman puts bait at the hook, thus, pretending to be a genuine food for fish. But the hook inside it takes the complete fish out of the lake. Communications purporting to be from popular social web sites, auction sites, online payment processors or IT administrators are commonly used to lure the unsuspecting public[1]. Phishing is typically carried out by e-mail spoofing or instant messaging and it often directs users to enter details at a fake website whose look and feel are almost identical to the legitimate one. Phishing is an example of social engineering techniques used to deceive users,[2] and exploits the poor usability of current web security technologies.[3] Attempts to deal with the growing number of reported phishing incidents include legislation, user training, public awareness, and technical security measures.

Social networking sites are now a prime target of phishing, since the personal details in such sites can be used in identity theft; in late 2006 a computer worm took over pages on MySpace and altered links to direct surfers to websites designed to steal login details.[21] Experiments show a success rate of over 70% for phishing attacks on social networks. There are anti-phishing websites which publish exact messages that have been recently circulating the internet, such as FraudWatch International and Millersmiles. Such sites often provide specific details about the particular messages [1].

1.2 Website forgery

Once a victim visits the phishing website the deception is not over. Some phishing scams use JavaScript commands in order to alter the address bar. This is done either by placing a picture of a legitimate URL over the address bar, or by closing the original address bar and opening a new one with the legitimate URL. An attacker can even use flaws in a trusted website's own scripts against the victim. These types of attacks (known as cross-site scripting) are particularly problematic, because they direct the user to sign in at their bank or service's own web page, where everything from the web address to the security certificates appears correct. In reality, the link to the website is crafted to carry out the attack, making it very difficult to spot without specialist knowledge. Just such a flaw was used in 2006 against PayPal..

A Universal Man-in-the-middle (MITM) Phishing Kit, discovered in 2007, provides a simple-to-use interface that allows a phisher to convincingly reproduce websites and capture log-in details entered at the fake site.

To avoid anti-phishing techniques that scan websites for phishing-related text, phishers have begun to use Flash-based websites [2]. These look much like the real website, but hide the text in a multimedia object. To foil these problems various anti-phishing techniques introduced. Some of them are discussed in the following.

2. ANTI-PHISHING SOLUTIONS

2.1 Digitally Signed Email

Digitally signed emails allow the recipient to verify that the sender information is genuine. This also lets the recipient know that the message has not been modified in transit. These assurances are extremely useful in the context of phishing, as they prevent individuals from impersonating established organizations in phishing emails. Popular digital signature standards include OpenPGP and S/MIME, although they are incompatible with each other. These facilities can be used with mail clients such as Outlook, Navigator and Eudora. At first glance, digitally signed emails appear well suited to combating the phishing problem. However, to date very few organizations with on-line banking or e-commerce facilities use this technology. Companies frequently targeted by phishing attacks such as Citibank, eBay and US Bank do not use digital signatures at all. This has been attributed to the difficulty end-users have in using digital signature technology (Tally, Get al, 2004)[2].

2.2 Online Brand Monitoring

Companies such as Cyveillance, NameProtect and Netcraft offer on-line brand monitoring services. This
Entails monitoring domain name registrations, web pages, spam emails and other on-line content for illegal use of clients' brand names. If illegal use of a client's brand name is detected, for example on a phishing web site, then the client is notified and can take remedial to close the website. Brands and Legitimate Entities Hijacked by Email Phishing Attacks – 2nd Half 2010 shown in Fig.1 [4].

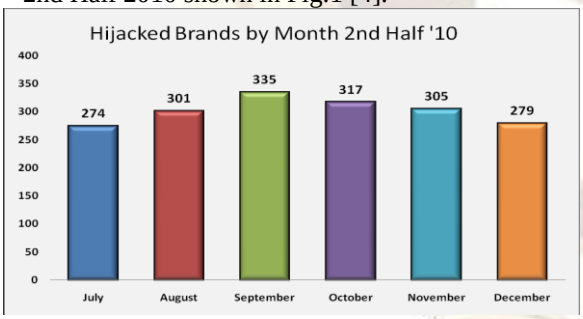

Fig.1 Online Brand Monitoring, 2010

2.3 Spam Filters

Phishing emails are transmitted using normal spam mechanisms and often contain similar characteristics to commercial spam emails. Consequently, current spam filters can be used to defend end-users against
phishing attacks. Spam filters classify incoming mail as either spam or non-spam. Where the classification takes place depends on the type of spam filter employed. Gateway spam filtering is normally used by large organizations and ISPs. This type of filter adjudges email messages arriving at the mail gateway. Desktop spam filters are also available and may be integrated or run in combination with a user's mail program.

2.4 Web browser extensions

Since phishing relies largely on deceptive Web sites, Web browsers are a natural focus for anti-phishing measures. An early means of adding anti-phishing capabilities to Internet Explorer was the EarthLink Toolbar. Whenever the user browses to a known phishing web site, the tool alerts them to this fact and the user is redirected to a warning page hosted by EarthLink. Similar strategies for user alerting are now appearing within mainstream Web browsers. Mozilla Firefox has a facility enabled by default that also works by checking visited Web sites against a list of known phishing sites. In this case, the phishing site list is automatically downloaded and regularly updated within Firefox. Since new phishing attacks may arise at any time, an additional option allows users to check sites against an online service for more up-to-date protection. Users may also report 'Web Forgery' in cases where a suspect site is not detected by the antiphishing system. In similar vein, Microsoft have added comparable anti-phishing features to version 7 of Internet Explorer[3].

3. PHISHING CHARACTERISTICS AND INDICATORS

There are many characteristics and indicators that can distinguish the original legitimate e-banking website from the phishing one. This system managed to gather 27 phishing features and indicators and clustered them into six Criteria (URL & Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar and Social Human Factor), and each criteria has its own phishing components. The full list is shown in Table 1 which is used later on our analysis and methodology study.

3.1 Why use Data Mining?

DM is the process of searching through large amounts of data and picking out relevant information. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from large data sets . Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions[5].

3.2 Existing system:

The approach described here is to apply data mining algorithms to assess e-banking phishing website risk on the 27 characteristics and factors which stamp the forged website. We utilized data mining classification and association rule approaches in our new e-banking phishing

website detection model to find significant patterns of phishing characteristic or factors in the e-banking phishing website archive data. Particularly, we used a number of different existing data mining association and classification techniques[6].

Criterion	Phishing Indicators
URL & Domain Identity	Using IP address
	Abnormal request URL
	Abnormal URL of anchor
	Abnormal DNS record
	Abnormal URL
Security & Encryption	Using SSL Certificate
	Certificate authority
	Abnormal cookie
	Distinguished names certificate
Source Code & Java script	Redirect pages
	Straddling attack
	Pharming attack
	On Mouse over to hide the Link
	Server Form Handler (SFH)
Page Style & Contents	Spelling Errors
	Copying website
	Using form s with Submit button
	Using pop-ups windows
	Disabling right-click
Web Address Bar	Long URL address
	Replacing similar char for URL
	Adding a prefix or suffix
	Using the @ Symbols to confuse
	Using hexadecimal char codes
Social Human Factor	Emphasis on security
	Public generic salutation
	Buying time to access accounts

Table 1: Phishing Indicators and their criteria

1. The approach described here is to apply data mining algorithms to assess website phishing risk on the 27 characteristics and factors which stamp the forged website [4].
2. Associative and classification algorithms can be very useful in predicting Phishing websites.
3. It can give us answers about what are the most important e-banking phishing website characteristics and indicators and how they relate with each other.
4. The choice of PART algorithm is based on the fact that it combines both approaches to generate a set of rules.
5. Associative classifiers produce more accurate classification models and rules than traditional classification algorithms.

Objective:

The motivation behind this study is to create a resilient and effective method that uses Data Mining algorithms and tools to detect e-banking phishing websites in an Artificial Intelligent technique. Associative and classification algorithms with bat algorithm can be very useful in predicting Phishing websites[6].

3.3 Proposed system:

Optimization is nothing but selection of a best element from some set of available deternatives. An optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function. In the proposed system, we implement the IBAT (Modified Bat) which is a metaheuristic algorithm that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. The Modified Bat Algorithm is based on the echolocation behavior of micro-bats with varying pulse rates of emission and loudness with Doppler Effect.

4. THEORY AND METHODOLOGY

4.1 Rule Generation using Associative Classification Algorithms

To derive a set of class association rules from the training data set, it must satisfy certain user-constraints, i.e support and confidence thresholds. Generally, in association rule mining, any item that passes MinSupp is known as a frequent item. We recorded the prediction accuracy and the number of rules generated by the classification algorithms and a new associative classification algorithm. Error rate comparative having specified the risk of e-banking phishing website and its key phishing characteristic indicators, the next step is to specify how the e-banking phishing website probability varies. Experts provide fuzzy rules in the form of **if…then** statements that relate e-banking phishing website probability to various levels of key phishing characteristic indicators based on their knowledge and experience.

On that matter and instead of employing an expert system, we utilized data mining classification and association rule approaches in our new e-banking phishing website risk assessment model which automatically finds significant patterns of phishing characteristic or factors in the e-banking phishing website archive data.

Association Classification is a special case of association rule mining in which only the class attribute is considered in the rule's right-hand side, ie A,B -) Y, then A. B must be input items attributes and Y must be the output class attribute. The output class attribute of our phishing website rate is one of these values (Very Legitimate, Legitimate, Suspicious, Phishy or Very Phishy). Example of the training phishing data sets to be classified is shown in the Table 2(G –Genuine, D-Doubtful & F –Fraud)[7].

ID	URL	Security	Java	Style	Address	Social	Class/Phishing Rate
1	G	G	D	G	G	G	Very Legitimate
2	D	G	G	D	G	D	Legitimate
3	D	D	G	F	D	G	Suspicious
4	F	D	G	D	F	D	Phishy
5	D	F	F	D	F	F	Very Phishy

Table 2 Example of Training Data Set

4.2 Modified Bat algorithm (enhancement)

By idealizing some of the echolocation characteristics of micro-bats, we can develop various bat-inspired algorithms or bat algorithms. Here we developed Modified Bat Algorithm with Doppler Effect. For simplicity, here some of the approximate or idealized rules:

1. All bats use echolocation to sense distance, and they also 'know' the difference between food/prey and background barriers in some magical way;

2. Bats fly randomly with velocity vi at position xi with a fixed frequency f_{min}, varying wavelength λ and loudness A0 to search for prey. They can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission r € [0, 1], depending on the proximity of their target;

3. Doppler Effect is the change in frequency of a wave for an observer moving relative to the source of the wave. The received frequency is higher (compared to the emitted frequency) during the approach, it is identical at the instant of passing by, and it is lower during the recession.

   (i) where v_s is positive if the source is moving away from the observer, and negative if the source is moving towards the observer.

$$f = \left(\frac{c}{c + v_s}\right) f_0 \quad \text{----(1)}$$

(ii) (or) where the similar convention applies: v_r is positive if the observer is moving towards the source, and negative if the

$$f = \left(\frac{c + v_r}{c}\right) f_0 \quad \text{----(1)}$$

(iii) (or) Single equation with both the source and receiver moving.

$$f = \left(\frac{c + v_r}{c + v_s}\right) f_0 \quad \text{-----(1)}$$

where

C is the velocity of waves in the medium

v_r is the velocity of the receiver relative to the medium; positive if the receiver is moving towards the source.

v_s is the velocity of the source relative to the medium; positive if the source is moving away from the receiver.

4. Although the loudness can vary in many ways, we assume that the loudness varies from a large (positive) A_0 to a minimum constant value A_{min}

Another obvious simplification is that no ray tracing is used in estimating the time delay and the three dimensional topography. Though this might be a good feature for the application in computational geometry, however, we will not use this as it is more computationally extensive in multidimensional cases. In addition to these simplified assumptions, we also use the following approximations, for simplicity. In general the frequency f in a range [f_{min}, f_{max}] corresponds to a range of wavelengths [λmin,λ max]. For example a frequency range of [20 kHz, 500 kHz] corresponds to a range of wavelengths from 0.7mm to 17mm.

For a given problem, we can also use any wavelength for the ease of implementation. In the actual implementation, we can adjust the range by adjusting the wavelengths (or frequencies), and the detectable range (or the largest wavelength) should be chosen such that it is comparable to the size of the domain of interest and then matching down to smaller ranges. Furthermore, we do not necessarily have to use the wavelengths themselves; instead, we can also vary the frequency while fixing the wavelength λ. This is because **λ** and f are related due to the fact **λ** f is constant. We will use this later approach in our implementation.

For simplicity, we can assume f € [0, fmax]. We know that higher frequencies have short wavelengths and travel a shorter distance. For bats, the typical ranges are a few meters. The rate of pulse can simply be in the range of [0, 1] where 0 means no pulses at all, and 1 means the maximum rate of pulse emission[8].

Movement of Virtual Bats

In simulations, we use virtual bats naturally. We have to define the rules how their positions x_i and velocities v_i in a d-dimensional search space are updated. The new solutions $\mathbf{x_i^t}$ and velocities $\mathbf{v_i^t}$ at time step t are given by

$$\mathbf{fi = fmin + (fmax - fmin)\beta,} \quad (2)$$

$$\mathbf{v_i^t = v_i^{t-1} + (x_i^t - x_*)f_i,} \quad (3)$$

$$\mathbf{x_i^t = x_i^{t-1} + v_i^t,} \quad (4)$$

Where **β** [0, 1] is a random vector drawn from a uniform distribution. Here x_* is the current global best location (solution) which is located after comparing all the solutions 4 among all the n bats.

As the product $\lambda_i f_i$ is the velocity increment, we can use either fi (or λ_i) to adjust the velocity change while fixing the other factor λ_i (or fi), depending on the type of the problem of interest. In our implementation, we will use fmin = 0 and fmax = 100, depending the domain size of the problem of interest. Initially, each bat is randomly assigned a frequency which is drawn uniformly from [fmin, fmax].

For the local search part, once a solution is selected among the current best solutions, a new solution for each bat is generated locally using random walk

$$\mathbf{x_{new} = x_{old} + EA^t,} \quad (5)$$

where E € [−1, 1] is a random number, while $\mathbf{A^t} =< \mathbf{A^t_i} >$ is the average loudness of all the bats at this time step. The update of the velocities and positions of bats have some similarity to the procedure in the standard particle swarm optimization as f_i essentially controls the pace and range of the movement of the swarming particles. To a degree, BA can be considered as a balanced combination of the standard particle swarm optimization and the intensive local search controlled by the loudness and pulse rate.

Loudness and Pulse Emission

Furthermore, the loudness Ai and the rate ri of pulse emission have to be updated accordingly as the iterations proceed. As the loudness usually decreases once a bat has found its prey, while the rate of pulse emission increases, the loudness can be chosen as any value of convenience. For example, we can use

$$\mathbf{A_0} = \mathbf{100} \text{ and } A_{min} = 1.$$

For simplicity, we can also use $\mathbf{A_0} = 1$ and $A_{min} = \mathbf{0}$, assuming $A_{min} = \mathbf{0}$ means that a bat has just found the prey and temporarily stop emitting any sound[9] Now we have

$$\mathbf{A_i^{t+1} = \alpha + A_i^t}$$

$$\mathbf{r_i^{t+1} = r_i^0}\,[1-\exp(-\gamma t)] \quad (6)$$

Where **α** and **γ** are constants. In fact, **α** is similar to the cooling factor of a cooling schedule in the simulated annealing. For any $0 < \alpha < 1$ and $\gamma > 0$, we have

$$\mathbf{A_i^t \dashrightarrow 0,\ r_i^t \dashrightarrow r_i^0 \ as\ t \dashrightarrow \infty} \quad (7)$$

In the simplicity case, we can use **α** = **γ**, and we have used **α** = **γ** = 0.9 in our simulations. The choice of parameters requires some experimenting. Initially, each bat should have different values of loudness and pulse emission rate, and this can be achieved by randomization. For example, the initial loudness $\mathbf{A_i^0}$ can typically be [1, 2], while the initial emission rate $\mathbf{r_i^0}$ can be around zero, or any value $\mathbf{r_i^0}$ € [0, 1] if using (6). Their loudness and emission rates will be updated only if the new solutions are improved, which means that these bats are moving towards the optimal solution.[10]

5. IMPLEMENTATION

5.1 Association and Classification Rule

Input: Webpage URL

Output: Phishing website identification

Step 1: Read web phishing URL

Step 2: Extract all 27 feature

Step 3: For each feature, Assign fuzzy membership degree value and Create fuzzy data set[11].

Step 4: Apply association rule mining & generate classification rule

Step 5: Aggregate all rule above minimum confidence.

Step 6: Defuzzification of fuzzy values into original values.

Step 7: Apply rule on test data and find whether the site is phishing or not and these steps are shown in Fig.2

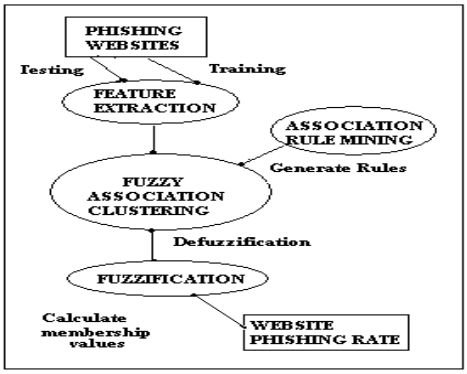

Fig. 2 Work Flow

5.2 Modified Bat Algorithm

Based on these approximations and idealization, the basic steps of the Modified Bat Algorithm can be summarized as the pseudo code shown in Fig.3[14].

Objective function f(**x**), **x**=(**x1….xd**)T
Initialize the bat population x_i = 1,2…n) and $\mathbf{V_i}$
Define Pulse frequency fi at $\mathbf{x_i}$
Initialize the rates $\mathbf{r_i}$ and the loudness $\mathbf{A_i}$
While (t < Max number of iterations)
Generate new solutions by adjusting frequency,
Apply equation (1)
And updating velocities and locations /solutions
[Equations (2) and (4)]
If (rand > $\mathbf{r_i}$)
Select a solution among the best solutions
Generate a local solution around the selected best solution
End if
Generate a new solution by flying randomly
If (rand < $\mathbf{A_i}$ & $\mathbf{f(x_i)} < \mathbf{f(x_*)}$)
Accept the new solutions
Increase ri and reduce $\mathbf{A_i}$
end if
Rank the bats and find the current best $\mathbf{x_*}$
end while

Fig.3: Pseudo code of the Modified Bat algorithm (MBA).

From the pseudo code, it is relatively straightforward to implement the Bat Algorithm in any programming language. We implemented it in Java for various test functions[11].

6. RESULTS AND DISCUSSION:

6.1 Performance Comparisons: In order to compare the performance of the new algorithm, we have tested it against other heuristic algorithms, including Ant Colony ptimization (ACO) and particle swarm optimization (PSO). There are any variants of PSO, and some variants such as the mean PSO could perform better than the standard PSO; however, the standard PSO is by far the most popularly used. Therefore, we will also use the standard PSO in our comparison. There are many ways to carry out the comparison of algorithm performance, and two obvious approaches are: to compare the numbers of function evaluations for a given tolerance or accuracy, or to compare their accuracies for a fixed number of function evaluations. Here we selected the second one. The performance analysis of the proposed system is compared with the existing system with the performance metrics mentioned[12].

Error rate: The proposed Bat algorithm will get the less error rate when compared to the existing ACO and PSO algorithms as shown in the fig.4.

Correct prediction: the proposed algorithm predicts the phishing website more accurate than the existing algorithms as shown in the fig.5.

Factors	ACO	PSO	BAT	MBAT
Training Data	1080	1080	1080	1080
Testing Data	572	572	572	572
No. of Folds	10	10	10	10
No. of Rules	27	27	27	27
Iterations	100	100	100	100
Accuracy %	88.92	92.13	94.97	97.98
Time taken (ms)	1,24,542	1,16,903	93,773	82,280
Error Rate %	10	8 6	5	3 2

Table 3 Comparison with existing algorithms

There is a significant relation between the two phishing website criteria's *(URL & Domain Identity)* and *(Security & Encryption)* for identifying phishing website. Also we found insignificant trivial influence of the *(Page Style & Content)* criteria along with *(Social Human Factor)* criteria for identifying phishing websites. Modified Bat Algorithm produces more accurate classification models than Associative classifiers [15].

2

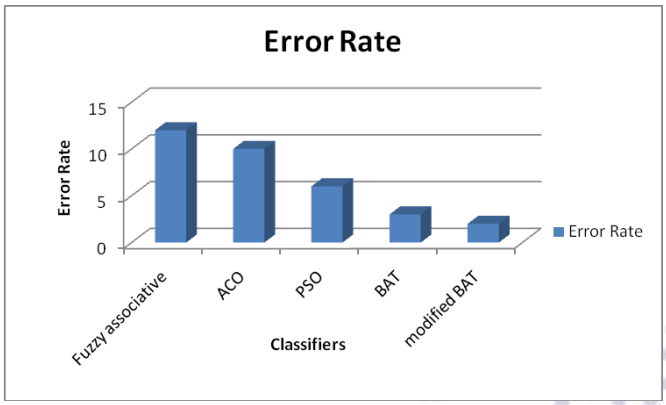

Fig.4 Error Rate Comparision

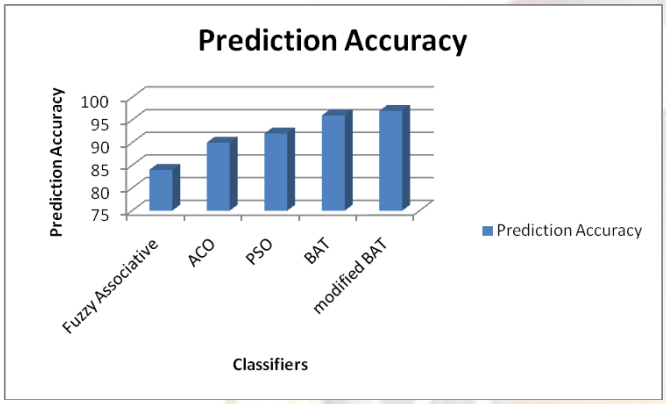

Fig.5 Accuracy comparision

8. REFERENCES

1. http://commons.wikimedia.org/wiki/File:Phishing_info_ graph.svg, http:/ /www .gartner. com /it/ page
2. STAMFORD, Conn., (April 14, 2009). "Gartner Says Number of Phishing Attacks on U.S. Consumers Increased 40 Percent in 2008". Gartner. "UK phishing fraud losses double". Finextra. March 7, 2006. http:// www. finextra. com/ fullstory asp?id=15013.
3. Richardson, Tim (May 3, 2005). "Brits fall prey to phishing".The Register. http:/ /www/ theregister .co. uk/ 2005/05/03/aol_phishing/.
4. Miller, Rich. "Bank, Customers Spar Over Phishing Losses". *Netcraft*. http://n ews.netcraft .com/ rchives/ 2006/09.
5. Associative Classification Techniques for predicting e-Banking Phishing Websites, Maher Aburrous Dept. of Computing ,Universit y of BradfordBradford, UK.
6. GARTNE R, INC. Gartner Says Number of Phishing Emails Sent to U.S. Adults Nea rly Doubles in Just Two Years, http //www .gartner. com/ it/pag e.jsp3.
7. T.Moore and R. Clayton, "An empirical analysis of the current state of phishing attack and defence", In Proceedings of the Workshop on the Economics of Information Security (WEIS2007)
8. Antiphishing Toolbars the comparison with existing toolbars, IJCA November,2009
9. A. Hossain, M. Dorigo, The Two Way Authentication Approach, http:// iridia.ulb.ac.be /mdorigo/ TWA.html
10. Ant Colony Optimization, Vittorio Maniezzo, Luca Maria Gambardella, Fabio de Luig**i.**
11. Mining Fuzzy Weighted Association Rules Proceedings of the 40th Hawaii International Conference on System Sciences – 2007.
12. WEKA - University of Waikato, New Zealand, EN,2006: "Weka -Data Mining with Open Source Machine Learning Software in Java", 2006 ,
13. Echolocation bats, www. scholarpedia. org/article/Bats theory.
14. A New Metaheuristic Bat-Inspired Algorithm, Xin-She Yang, Department of Engineering, University of Cambridge.
15. "Bats behaviour", www. Swam inteLligence.org.