International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 12, December-2014

Detecting Spam Zombies with Semantic Matching Based NMF Hierarchical Clustering By Monitoring Outgoing Messages

C. Daniel Nesa Kumar *
Asst Professor,
Department Of MCA,
Hindusthan College Of Arts And Science,

J. Jeyaboopathi Raja
Asst Professor
Dept Of BCA
Hindusthan College Of Arts And Science,

R. Aruna
Asst Professor
Dept Of CSE
Sasurie Engineering

***Abstract*—Due to the growth of the computer technologies the usage of the internet also increased in now days. The increases of the web usage the security of the each and every user data also play major significant roles to each user, to avoid the loss of the data. Several types of the attacks have been occurred during this process those attacks are spamming, Dos, DDoS, Worms, Viruses, and individuality theft. These types of spamming motivates the attackers to hacks the secret information of the people ,in order to conquer these problem in this paper focus on the spam detection problem for outgoing the messages in the larger organization applications such as educational data, university data ,hospital sharing information data and individual mail communication .The proposed method automatically monitor the activities of the user through the post messages, these types of activities performed by spammer is known as the spam zombies. In earlier work the detection of the spam zombie's methods is known as the SPOT, through the automatic examination of the outgoing messages. It is designed based on the probabilistic test is known as SPRT. SPRT based probabilistic measurement analysis the negative and positive results that corresponds to false positive (FP) ,False negative (FN),True Positive (TP) and true negative (TN) error rates through user defined thresholds. The methods doesn't measure semantic measurement based matching to detect worm attacks ,in order to solve this issue additionally propose an semantic matching based Nonnegative matrix factorization clustering method in hierarchical structure . The spam and non-spam clustered are formed based on the degree value of the fuzzy membership function; it tries to divide the email message into two clusters through the non negative matrix and a nonnegative coefficient matrix. In addition, the proposed NMF-Hierarchical Clustering with existing SPOT Methods respectively, and show that proposed NMF- Hierarchical Clustering SPOT outperforms than existing methods.**

Keywords— Compromised Machines, Spam Zombies, Nonnegative Matrix Factorization (NMF) ,Compromised Machine Detection Algorithms, Sequential Probability Ratio Test(SPRT).

I. INTRODUCTION

Due to the growth of the computer technologies the usage of the internet also increased in now days. The increases of the web usage the security of the each and every user data also play major significant roles to each user, to avoid the loss of the data. Since the WWW not only provides possible information in always so the security of the internet for each people also important, because many of the factors which easily affects the system and hacker the information of the data through the posting of the unwanted messages, these types of the messages posted by user is known as the spam messages .These types of the messages posted by people through the sending of email to people. These types of messages posted by user are known as the attacks it may be manual or automatic category of the types. The types of the attacks are Dos, DDOS [1-3], E-mail Worms, Viruses, Worms etc. It may stop the working principle of the machines, it is known as the zombies. If it is in the form of the spamming type of the category is specified as the spam zombies [4-6]. A Spam message becomes the one of the major important security problem in the email, facebook, twitter and other types of the social network communication in the internet.

These types of the spam messages in the internet are generally stored in the form of the bots, it is controlled by the term is called as the botnet. The Botnet controllers make use of the technologies such as the IRC channels to handle and manage these bots. Botnets contain manifold wrong uses: growing DDoS attacks, theft password of the each user and individuality, create tick fraud [7], and basis spam email [8]. This technology source damages the spam email ,everywhere spam is summarize kindly to include earliest public relations email messages, similarly as email messages through viruses, and unrelated superfluous email messages.

To detect the spam users several number of the messages are locally generated by the user to detect easily but these methods used in the earlier work may not provide the exact spam user results for outgoing the messages .In order to overcome these problem of the outgoing messages detection for spam user ,in this paper presents an novel spam detection method with semantic analysis of the each word through the creation of the hierarchical structure through fuzzy membership function .During this process the email messages the information of the individual user are separated into the data privacy and truthfulness. It boosts the spam detection results for email messages through the collection of the spam messages in the university campus at any of the country. It is also the unusual specialize in the defense guarantee troubles.

In earlier work the detection of the spam zombie's methods is known as the SPOT, through the automatic examination of the outgoing messages. It is designed based on the probabilistic test is known as SPRT. The semantic measurement of the terms

are not supported in this work so some the messages are unwontedly filtered as spam and some of the spam messages are not correctly filtered in the email communication, to overcome the problem of the SPOT methods .In this paper presents an novel spam detection methods with fuzzy based semantic measurement and thus measure the semantic meaning of the each words in the spam messages by clustering the each one of the similar messages in the top down manner structure is known as the similarity tree ,it factorizes the tree values in the form of matrix is known as NMF used in the several number of the applications [9-10]. In NMF the column represents the total number of the words presented in the each and every spam messages which belongs to common word for each user and the row belongs to the single user spam messages information through the specification of the k clusters, it exactly identifies the spam zombies and detects the worm packets.

This template, modified in MS Word 2007 and saved as a "Word 97-2003 Document" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. BACKGROUND STUDY

In earlier it becomes more important to analysis the results of the unauthorized spam messages in email at the server side of the current study [11-12] investigate the aggregate universal individuality in the spamming botnets with the intention which includes the size of botnets as well as spamming patterns of the botnets. This learning method provide significance approaching interested collective entirety characteristics of spamming botnets as a result of clustering spam messages expected in the provider addicted to spam operation use the embedded URLs as well as near-duplicate satisfied clustering, in the same way.

BotMiner [13] is primary botnet identification scheme incorporate protocols. In these methods the flows are categorized into the groups based on the communication schemas; it is also equivalent malicious activity pattern, in the same way. Inter-section of two groups is calculated to the cooperation technology. The results of the botmined is also compared with the existing methods such as BotHunter, BotSniffer and BotMiner, SPOT for each and every one of the monetary motivation premeditated for attackers draft huge number into cooperation machines.

It is simple also powerful statistical technique, SPRT method effectively functional throughout many areas. In the part of the networking security, the methods detect the port scan actions [14], proxy-based spamming activities [15], anomaly-based botnet discovery [16], and MAC protocol acting up nodes in wireless networks.

In the zombie detection, first converse regarding spam zombie detection associated work [17] identify the electronic mail spam bots through the aggregating spam proceedings in addition to clustering all the way through to contented of the electronic mail messages as a result of shingling [18] algorithm. Yu.et.al [19] proposed an efficient construction called as AutoRE works through permitting for URLs classify botnets. Now AutoRE consider URLs because signatures as well as normal expressions that are used to generate spam operation.

In universal based spam botnet detection methods are discussed here. Choi et.al proposed a method use to detect bots frequently based on DNS queries created through means of them [20]. Based on correspondence in the direction of grouping activity DNS travel of bots frequently detected in this work. Database management based spam detection also proposed in earlier work for each network depending on the property of the symmetry activities [21].

III. PROPOSED SEMANTIC MATCHING BASED NMF HIERARCHICAL CLUSTERING METHODOLOGY

Anomaly detection of the email messages through the email communication is known as BotSniffer .It detects the botnet through discover through measuring the spatial –semantic based relationship measurement between the terms is known as the botnets. The Botnet controllers make use of the technologies such as the IRC channels to handle and manage these bots, flows of the controllers is categorized into the different groups based on the usual behavior of he each and every server connected to common user if it exists within the specified group of the spam it is named as the spam nodes otherwise is considered as the non-spam user.

To detect the spam users several number of the messages are locally generated by the user to detect easily but these methods used in the earlier work may not provide the exact spam user results for outgoing the messages .In order to overcome these problem of the outgoing messages detection for spam user, in this paper presents an novel spam detection method with semantic analysis of the each word through the creation of the hierarchical structure through fuzzy membership function.

In the base work use methods for spam zombie detection is named as the SPOT through the continuous monitoring the behavior of the user based on the outgoing messages of a network. It is designed based on the SPRT based probabilistic measurement test to analysis the negative and positive results that corresponds to false positive (FP) ,False negative (FN),True Positive (TP) and true negative (TN) error rates through user defined thresholds. For user defined threshold values is structured based on the threshold count values such as the percentage and the count based threshold measurement for each spam messages for university or individual campus of the organization it consists of the major properties such as the IP address ,source and destination address ,completion time ,date of the message send by user and received date also .It determines the percentage threshold through the forwarding of the spam messages by internal machines. Second determine the whether the calculated PT user is spam or non spam user based on the semantic matching results from the fuzzy relation

based hierarchical tree structure clustering model with the NMF factorization methods and also communicate to the Hotmail server also . The log file of the each user consists of several number of the sender and received email counts of every server on a usual.

The relation of the each and every spam magnitude is counted based on the filter is removes the unwanted email messages and stores the history of the spam email messages communication history for every and every one of the spam users ,non-spam user history . In conclusion, they're interested concerning judgment of the spam whether or not based on the set of the email messages stored in the filter which is properly grouped into two categories in the NMF based hierarchical clustering tree structure clustering methods thus takes an SPRT test. For each and every set of the spam messages they study the relationship among the equivalent destination sites.

In the NMF based hierarchical tree structure in this work presents a Rank-2 NMF based clustering methods which divides the spam messages into several parts based on the general properties collected from spam messages in the email communication. The tree is constructed in the top down the most important node in the tree is considered as the root nodes and the subsequent properties in the nodes is considered as the child nodes ,the tree is balanced tree until all the attributes are completed in the collected spam messages. The splitting criteria for each and every node in the tree are constructed by referring the earlier work [22] based on the label defined by user.

To split the spam message and non spam message data as the left or right child node in the tree through the measurement of the semantic meaning with score value of the individual node based on the ranking score value rank-2NMF ,then formation of the messages into two groups either spam or non spam user messages. In the context of NMF, column represents the total number of the words presented in the each and every spam messages which belongs to common word for each user and the row belongs to the single user spam messages information through the specification of the k clusters for a topic [21] . Our strategy is to calculate the ranking score value from rank-2 NMF amongst the columns W. Nodes with highest probability is chosen for next splitting process for user. For representation of the general tree structure for spam and non-spam messages are shown in Fig.1.

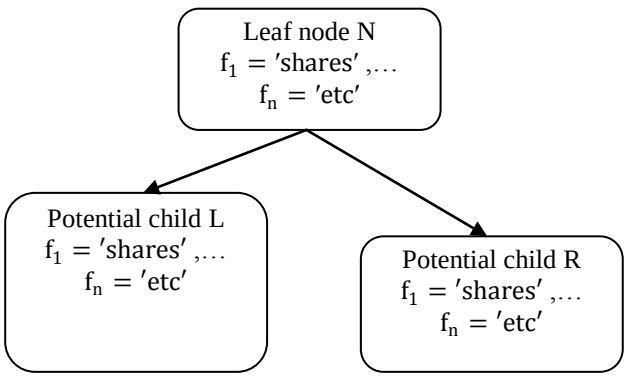

Fig. 1. NMF- hierarchical clustering

Initially the root node collected messages split the data as spam or non-spam user if the threshold of the user is exceeds in the spam threshold then it is splited as the left node L or else it is right node is considered as non spam user message .It is divided based on the rank score value of the each messages, it is also refers the concept of the normalized discounted cumulative gain (NDCG) for each and every spam messages. The ranked scored value of each and every word is converted into the normalized values between zero to one through the calculation of the term or word distribution w_N, in each column of the rank NMF through the rank node N in the tree structure, it descending order of the values is denoted by f_N, for children, L and R, denoted by f_L,and f_R. Suppose the completely prearranged terms equivalent to f_N is

$$f_1, f_2, \ldots f_m \quad (1)$$

and the mix up orderings in f_L,and f_R are correspondingly

$$f_{l_1}, \ldots . f_{l_m} \quad (2)$$
$$f_{r_1}, \ldots . f_{r_m} \quad (3)$$

First describe a location reduction factor $p(f_i)$ and a gain $g(f_i)$ for every one term f_i

$$p(f_i) = log(m - max\{i_1, i_2\} + 1) \quad (4)$$
$$g(f_i) = \frac{log(m - i + 1)}{p(f_i)} \quad (5)$$

Where $l_{i_1} = r_{i_2} = i$ .In other words, for each and every terms of the spam messages f_i ,discover the location i_1, i_2 in the mix up orderings in f_L,and f_R are correspondingly and place a location reduction factor $p(f_i)$ and a gain $g(f_i)$ for every one term f_i $\{g(f_i)\}_{i=1}^{m}$orderings. The highest ranking gain score value of the terms $\{\hat{g}(f_i)\}_{i=1}^{m}$ is arrange in downward order, resultant in one more sequence. Then, for mix up ordering f_s ($f_s = f_L$ or f_R), mNDCG is defined as:

$$mDCG(f_s) = g(f_{n_1}) + \sum_{i=2}^{m} g(f_{n_1})/log_2(i) \quad (6)$$

$$mIDCG = \hat{g} + \sum_{i=2}^{m} \hat{g}/log_2(i) \quad (7)$$

$$mNDCG(f_s) = \frac{mDCG(f_s)}{mIDCG} \quad (8)$$

Lastly, the attain of the leaf node N is calculate as:

$$score(N) = mNDCG(f_L) \times mNDCG(f_R) \quad (9)$$

In the above step if the left node mNDCG(f_L)of the message relates to spam messages and mNDCG(f_R) will be relates to the non-spam message results mNDCG(f_R)is small, and score (N) is small.

The overall procedure for spam detection with hierarchical clustering workflow based NMF methods for each and every one of the spam user files is defined in algorithm 1 .In step 8-15 detects the spam user mails in email communication until the maximum number of the iterations T trials with rank 2-NMF to split nodes into two cluster either spam or non-spam. At each and every one of the iteration the two clustered such as spam and non-spam messages are filtered to the potential children nodes N_1, N_2 are created. If none of the messages found as outlier in the number of iterations T ,then it is considered as non-spam outgoing messages or those messages are removed entirely from dataset .Some of the other procedure

also used in this type of situation through specifying the threshold value $\sigma = 0$ until all the leaf nodes turn into the lasting leaf nodes.

Input :A term –spam message matrix $X \in R_{+}^{m \times n}$ (often sparse) maximum no of leaf nodes k ,parameter $\beta > 1$ and $T \in N$ for outlier detection

Create a root node R containing all the n number of the messages

Score(R) $\leftarrow \infty$
Repeat
M $\leftarrow$ a current leaf node with the highest score
Trail index i $\leftarrow 0$
Outlier set Z $\leftarrow \emptyset$
While i $< T$ do
Run rank-2 NMF on M and create two potential children N_1, N_2 where $|N_1| \geq |N_2|$
If $|N_1| \geq \beta |N_2|$ and score (N_2) is smaller than every position score of the current leaf nodes then
$Z \leftarrow Z \cup N_2 M \leftarrow M - Z$,i $\leftarrow i + 1$
Else
Break
End if
End while
If i<T
then split M into $N_1 \& N_2$
compute score (N_1) and score (N_2)
else
$M \leftarrow M - Z$
score(M) $\leftarrow -1$
End if Until ≠leaf nodes =k

IV. Experimentation Results

In order to evaluate the performance of the spam detection NMF-HC SPOT method and the existing SPOT spam detection methods first we collect or gather the email trace files from large US campus network for continuous two months, the spam server is runned automatically in Asian countries. The collected email messages consists of the following attributes such as the local arrival time of the each user ,IP address of the user ,then choose the messages as either spam or non spam user based on the message from email trace files .

Initially the files are collected by continuous monitoring of the user behavior from one user machine to another user machine in the US campus university then calculate the count and percentage threshold value to each spam user in the outgoing messages in the US campus network to detect spam zombies user between the methods NMF-HC SPOT and SPOT spam detection methods. In this evaluation the identified spam messages are forwarded to the FSU university campus network and intended to an FSU account. Then this set of the messages in FSU e-mails and perform continuous spam detection for each and every one of the FSU e-mails. The performance comparison results of proposed NMF-HC SPOT and the existing SPOT for confirmed and missed spam messages detected results are tabulated in Table 1 and illustrated in Fig.2, it shows that proposed system have less false results (0.01)

TABLE 1: PERFORMANCE COMPARISON RESULTS

Methods	Total number of the IP	Number of the spam detected	Confirmed	Missed
SPOT	410	147	138(0.971)	9(0.29)
NMF- HC SPOT	410	157	151(0.99)	3(0.01)

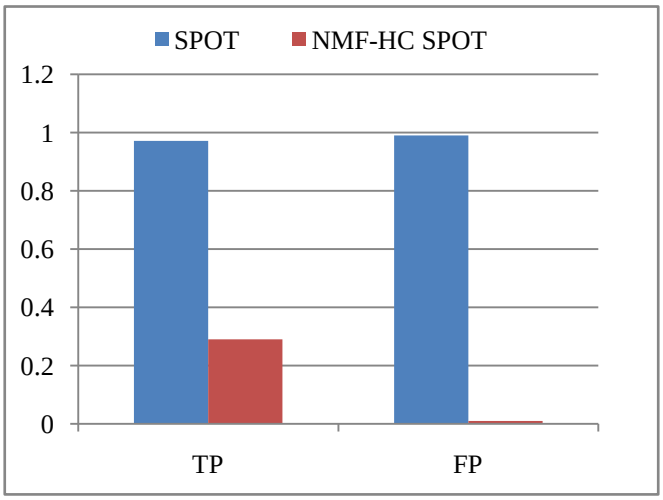

Fig.2. Performance comparison results

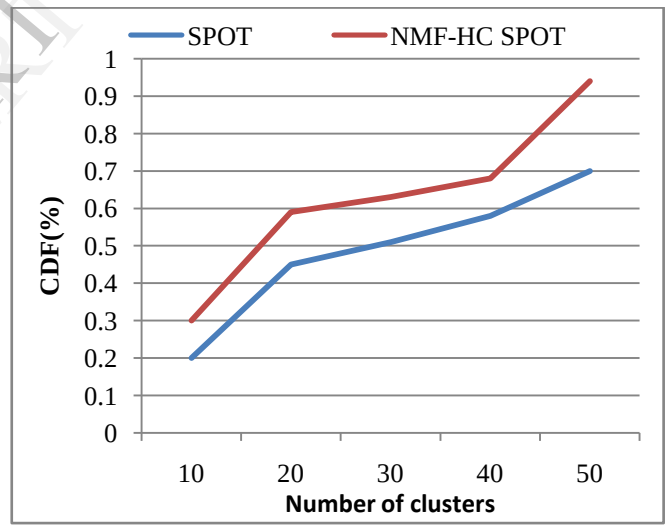

Fig.3. Distribution of spam messages in each cluster

Fig. 3 shows the cumulative distribution function (CDF) results for each cluster returned from the NMF-HC SPOT and the existing methods. Upto 90 percentage of the cluster value there are only less than 10 spam messages are present and it becomes less when the percentage of the cluster increases ,it shows that the number of the spam user are correctly filtered in the NMF-HC SPOT than the SPOT methods.

V. Conclusion and Future Work

Spam zombie becomes one of the most important internet security problems in nowadays, it ever-increasing daytime extremely quickly. So the detection of the spam messages also plays major important to detect spam zombies for email communication ,it becomes important to classify the messages into spam and non-spam messages In order to perform this

process in this paper presents an novel spam zombies detection methods based on the measurement of the semantic matching through NMF hierarchical Clustering .Proposed NMF-HC method cluster the message into two categories and filet the unwanted messages it is not under the both categories ,it also detects virus ,worm attack files separately .The evaluation of the proposed NMF-HC algorithm based semantic measurement is evaluated based on the email messages gathered from SFSU campus network ,it shows that the proposed NMF-HC based SPOT is an effective system and detect more number of the spam messages when compare to existing SPOT methods.

In future we detect the other types of the attack and apply different semantic matching based matching methods to detect the spam zombies, apply this work to other types of the email traces files in the different application at different campus network.

REFERENCES

[1] Bacher, T.Holz, M. Kotter, and G. Wicherski, "Know Your Enemy: Tracking Botnets," http://www.honeynet.org/papers/ bots, 2011.

[2] N. Ianelli and A. Hackworth, "Botnets as a Vehicle for Online Crime," Proc. First Int'l Conf. Forensic Computer Science, 2006

[3] J.Markoff, "Russian Gang Hijacking PCs in Vast Scheme," The New York Times, http://www.nytimes.com/2008/08/06/technology/ 06hack.html, 2008.

[4] A. Ramachandran and N. Feamster, "Understanding the Network-Level Behavior of Spammers," Proc. ACM SIGCOMM, pp. 291-302, 2006.

[5] P. Resnick, "Internet Message Format," IETF RFC 2822, 2001.

[6] F. Sanchez, Z. Duan, and Y. Dong, "Understanding Forgery Properties of Spam Delivery Paths," Proc. Seventh Ann. Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS '10), 2010.

[7] Daswani, N., Stoppelman, M., "The Google click quality and security teams", The anatomy of clickbot.a. In HotBots'07, 2007.

[8] Ramachandran, A., and Feamster, N," Understanding the network-level behavior of spammers", In SIGCOMM'06, 2006.

[9] W. Xu, X. Liu, and Y. Gong ,"Document clustering based on non-negative matrix factorization", In SIGIR '03: Proc. of the 26th Int. ACM Conf. on Research and Development in Information Retrieval, pp. 267-273, 2003.

[10] H. Kim and H. Park," Sparse non-negative matrix factorizations via alternating non-negativity constrained least squares for microarray data analysis", Bioinformatics, vol.23,no.12,pp.1495-1502, 2007.

[11] Y. Xie, F. Xu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov, "Spamming Botnets: Signatures and Characteristics," Proc. ACM SIGCOMM, 2008.

[12] L. Zhuang, J. Dunagan, D.R. Simon, H.J. Wang, I. Osipkov, G. Hulten, and J.D. Tygar, "Characterizing Botnets from Email Spam Records," Proc. First Usenix Workshop Large-Scale Exploits and Emergent Threats, 2008.

[13] G. Gu, R. Perdisci, J. Zhang, and W. Lee, "BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection," Proc. 17th USENIX Security Symp. 2008.

[14] J. Jung, V. Paxson, A. Berger, and H. Balakrishnan, "Fast Portscan Detection Using Sequential Hypothesis Testing," Proc. IEEE Symp. Security and Privacy, 2004.

[15] M. Xie, H. Yin, and H. Wang, "An Effective Defense against Email Spam Laundering," Proc. ACM Conf. Computer and Comm. Security, 2006.

[16] G. GU, J. Zhang, and W. Lee, "BotSniffer: Detecting Botnet Command and Control Channels in Network Traffic," Proc. 15th Ann. Network and Distributed System Security Symp. (NDSS '08), 2008.

[17] L. Zhuang, J. Dunagan, D.R. Simon, H.J. Wang, I. Osipkov, G. Hulten, and J.D. Tygar, "Characterizing Botnets from Email Spam Records," Proc. First Usenix Workshop Large-Scale Exploits and Emergent Threats, 2008.

[18] BRODER, A. Z., GLASSMAN, S. C., MANASSE, M. S., AND ZWEIG, G, "Syntactic clustering of the web", WWW'97, 1997.

[19] Y. Xie, F. Xu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov, "Spamming Botnets: Signatures and Characteristics," Proc. ACM SIGCOMM, 2008.

[20] Hyunsang Choi, Hanwoo Lee, Heejo Lee, Hyogon Kim ,"Botnet Detection by Monitoring Group Activities in DNS Traffic", IEEE pp. 715-720, 2007

[21] M. Xie, H. Yin and H. Wang ,"An effective defense against email spam laundering", In ACM Conference on Computer and Communications Security, Alexandria, VA, 2006.

[22] C. Ding and X. He ,"Cluster merging and splitting in hierarchical clustering algorithms", In ICDM '02: Proc. of the 2nd IEEE Int. Conf. on Data Mining, pp.139-146, 2002.

[23] Y. Chen, E.K. Garcia, M.R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-Based Classification: Concepts and Algorithms," J. Machine Learning Research, vol. 10, pp. 747-776, 2009.