International Journal of Computer Applications Technology and Research
Volume 3– Issue 9, 536 - 540, 2014, ISSN: 2319–8656

Security improvements Zone Routing Protocol in Mobile Ad Hoc Network

Mahsa Seyyedtaj
Department of computer, Shabestar branch,
Islamic Azad University, Shabestar,
Iran

Mohammad Ali Jabraeil Jamali
Department of computer, Shabestar branch,
Islamic Azad University, Shabestar,
Iran

Abstract: The attractive features of ad-hoc networks such as dynamic topology, absence of central authorities and distributed cooperation hold the promise of revolutionizing the ad-hoc networks across a range of civil, scientific, military and industrial applications. However, these characteristics make ad-hoc networks vulnerable to different types of attacks and make implementing security in ad-hoc network a challenging task. Many secure routing protocols proposed for secure routing either active or reactive, however, both of these protocols have some limitations. Zone Routing Protocol (ZRP) combines the advantages of both proactive and reactive routing protocols. In this paper we analyze the ZRP security improvements. Considering the delivery rate of packets, routing overhead, network delay, Simulation results show that Protocols operate under different constraints and none of the protocols are not able to provide security for all purposes.

Keywords: ad-hoc networks; secure routing; secure neighbor discovery; digital signature; zone routing protocol; secure zone routing protocol

1. INTRODUCTION

Mobile ad hoc networks (MANETs) consist of a collection of wireless mobile nodes which dynamically exchange data among them-selves without the reliance on a fixed base station or a wired back-bone network. MANET nodes are typically distinguished by their limited power, processing, and memory resources as well as high degree of mobility. MANET is very useful to apply in different applications such as battlefield communication, emergency relief scenario etc. In MANET nodes are mobile in nature, due to the mobility, topology changes dynamically. Due to its basic Ad-Hoc nature, MANET is venerable to various kinds of security attacks [1].

Researchers have proposed a large range of routing protocols for ad hoc networks. The basic goals of these protocols are the same: maximize throughput while minimizing packet loss, control overhead and energy usage. However, the relative priorities of these criteria differ among application areas. In addition, in some applications, ad hoc networking is really the only feasible solution, while in other applications, ad hoc networking competes with other technologies. Thus, the performance expectations of the ad hoc networks differ from application to application and the architecture of the ad hoc network, thus each application area and ad hoc network type must be evaluated against a different set of metrics. The routing protocols have organized into nine categories based on their underlying architectural framework as follows [2].

- Source-initiated (Reactive or on-demand)
- Table-driven (Pro-active)
- Hybrid
- Location-aware (Geographical)
- Multipath
- Hierarchical
- Multicast
- Geographical Multicast
- Power-aware

Among these protocols, refer to the first three:

Reactive Routing protocols: Whenever there is a need of a path from any source to destination then a type of query reply dialog does the work.Therefore, the latency is high; however, no unnecessary control messages are required.

Proactive routing protocols: In it, all the nodes continuously search for routing information with in a network, so that when a route is needed, the route is already known. If any node wants to send any information to another node, path is known, therefore, latency is low. However, when there is a lot of node movement then the cost of maintaining all topology information is very high.

Hybrid routing protocols: These protocols incorporates the merits of proactive as well as reactive routing protocols. A hybrid routing protocol should use a mixture of both proactive and reactive e approaches. Hence, in the recent years, several hybrid routing protocols are proposed like ZRP [5].

1.1 ZRP

Zone routing protocol is a hybrid protocol. It combines the advantages of both proactive and reactive routing protocols. A routing zone is defined for every node. Each node specifies a zone radius in terms of hops. Zones can be overlapped and size of a zone affects the network performance. The large routing zones are appropriate in situations where route demand is high and /or the network consists of many slowly moving nodes. On the other hand, the smaller routing zones are preferred where demand for routes is less and /or the network consists of a small number of nodes that move fast relative to one another. Proactive routing protocol works with in the zone whereas; reactive routing protocol works between the zones. ZRP consists of three components:

1) the proactive Intra zone routing protocol (IARP)

2) the reactive Inter zone routing protocol (IERP)

3) Bordercast resolution protocol (BRP).

Each component works independently of the other and they may use different technologies in order to maximize efficiency in their particular area. The main role of IARP is to ensure that every node with in the zone has a consistent updated routing table that has the information of route to all the destination nodes with in the network. The work of IERP gets started when destination is not available with in the zone. It relies on bordercast resolution protocol in the sense that border nodes will perform on-demand routing to search for routing information to nodes residing outside the source node zone [6]. The architectural of ZRP is shown in Figure 1.

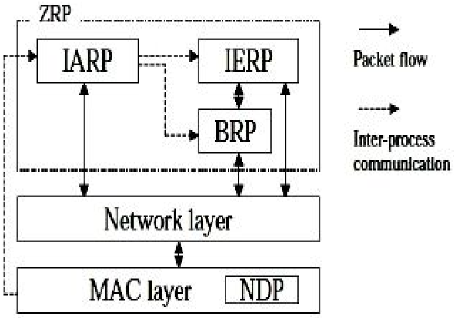


Figure 1. Architecture of ZRP [6].

2. PREVIOUS WORKS

In this section security improvements ZRP have examined.

2.1 SZRP1

The architectural design of SZRP1 is shown in Figure 2. The proposed architecture is a modification of ZRP [4]. It is designed to support both secure routing (intrazone and interzone) and effective key management. There are dedicated and independent components in SZRP1 to carry out these tasks. The functionality of each component and their interrelationship is explained below.

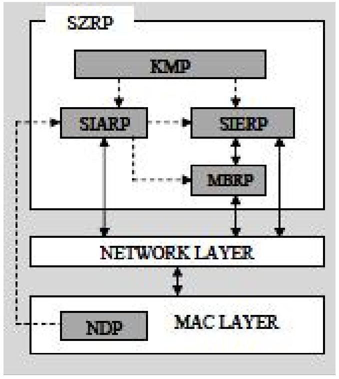


Figure 2. Architecture of SZRP1[4].

The key management protocol (KMP) is responsible for public key certification process. It fetches the public keys for each CN by certifying them with the nearest CA. The secure intrazone routing protocol (SIARP) and secure interzone routing protocol (SIERP) uses these keys to perform secure intrazone and interzone routing respectively.

SIARP is a limited depth proactive link-state routing protocol with inbuilt security features. It periodically computes the route to all intrazone nodes (nodes that are within the routing zone of a node) and maintains this information in a data structure called SIARP routing table. This process is called proactive route computation. The route information to all intrazone nodes collected in proactive route computation phase is used by SIARP to perform secure intrazone routing.

SIERP is a family of reactive routing protocols with added security features like ARAN. It offers on demand secure route discovery and route maintenance services based on local connectivity information monitored by SIARP.

In order to detect the neighbor nodes and possible link failures, SZRP relies on the neighborhood discovery protocol (NDP) similar to that of ZRP. NDP does this by periodically transmitting a HELLO beckon (a small packet) to the neighbors at each node and updating the neighbor table on receiving similar HELLO beckons from the neighbors. NDP gives the information about the neighbors to SIARP and also notifies SIARP when the neighbor table updates. We have assumed that NDP is implemented as a MAC layer protocol. A number of security mechanisms suggested in for MAC layer can be employed to secure NDP.

To minimize the delay during interzone route discovery, SIERP uses bordercasting technique similar to ZRP, which is implemented here by the modified border resolution protocol (MBRP). MBRP is a modification of the bordercast technique adopted in ZRP. It not only forwards SIERP's secure route discovery packets to the peripheral nodes of the bordercasting node but also sets up a reverse path back to the neighbour by recording its IP address. MBRP uses the routing table of SIARP to guide these route queries. Since, all security measures are taken by SIERP during interzone routing; no additional security mechanism is adopted by MBRP during bordercasting.

2.1.1 Simulation Environment

The simulation of Secure Zone Routing Protocol (SZRP) was conducted in NS-allinone-2.1b6a, on an Intel Pentium IV processor (2.4 GHz) and 512 MB of RAM running Ubuntu 7.2.

2.1.2 Performance Metrics

four performance metrics evaluated to compare the proposed protocol with ZRP under a trusted environment where all the nodes in the network are assumed to be benign. They are discussed below:

Average packet delivery fraction: This is the fraction of the data packets generated by the CBR sources that are delivered to the destination. This metric is important as it evaluates the ability of the protocol to discover routes.

Average routing load in bytes: This is the ratio of overhead control bytes to delivered data bytes. Secure Zone Routing Protocol (SZRP) has larger control overhead due to the certificate and signature embedded in the packets. For the calculation of this metric, the transmission at each hop along the route was counted as one transmission.

Average routing load in terms of packets: This metric is similar to the above, but here the ratio of control packet overhead to data packet overhead is calculated.

Average route acquisition latency: This is the average delay between the sending of a secure route discovery packet by a source for discovering a route to a destination and the receipt of the first corresponding route reply. This includes all the delays caused during the route discovery and route reply phases for signature verification and their replacement, in addition to the normal processing of the packets. If a route request timed out and needed to be retransmitted, the sending time of the first transmission was used for calculating the latency.

2.1.3 Simulation Environment

To evaluate proposed SZRP in a non-adversarial environment, the Network Simulator 2 (NS-2) have used. NS-2 is a discrete event simulator written in C++ and OTcl. At the link layer, the simulator implements the complete IEEE 802.11 standard Medium Access Control (MAC) protocol.

2.1.4 Simulation Results

In this section, The obtained results analyzed for each of the performance metric discussed. The resulting data were plotted using Gnuplot. Each data point in the resulting graphs is an average of 5 simulation runs with identical configuration but different randomly generated mobility patterns.

2.1.4.1 Average Packet Delivery Fraction

obtained results for average packet delivery fraction for both the 10 and 20 node networks. The packet delivery fraction obtained using SZRP is above 96% in all scenarios and almost identical to that obtained using ZRP. This suggests that SZRP is highly effective in discovering and maintaining routes for delivery of data packets, even with relatively high node mobility.

2.1.4.2 Average Routing Load in Bytes

The routing load measurements for both the protocols in terms of number of control bytes per data bytes delivered. The byte routing load of Secure Zone Routing Protocol (SZRP) is higher compared to that of ZRP. For example, it is nearly 40% for 20 nodes moving at 5 m/s, as compared to 22% for ZRP with identical topology and mobility pattern. With further increase in node mobility to 10 m/s, it increases to 75%, compared 45% for ZRP. This overhead is due to the certificate and signature embedded in the packets. The RSA digital signature is of 16 bytes and the certificate is 512 bytes long. Though these extra bytes are pure overhead they are necessary for security provisioning. Additionally, since ZRP has the advantage of smaller sized packets, the packet size of SZRP is not that much larger compared to other secure routing protocols even after inserting the security data.

2.1.4.3 Average Routing Load in Terms of Packets

While the number of control bytes transmitted by SZRP is larger than that of ZRP, the number of control packets transmitted by the two protocols is roughly equivalent. Figure 5.5 shows the average number of control packet transmitted per delivered data packet. Except for the scenario of 20 nodes moving at 1 m/s, where they exhibit some difference, the packet routing load for both the protocols are nearly the same for other scenarios. This is due to the fact that SZRP did not employ any extra control packets compared to ZRP for secure routing, except for the case of intrazone routing, which requires two additional control packets SKREQ and SKREP. However, with high node mobility, for example, when the nodes move with the speed of 5 m/s or 10 m/s, the number of times interzone routing carried out was significantly higher than intrazone routing. In this respect, the two protocols demonstrate nearly the same amount of packet overhead.

2.1.4.4 Average Route Acquisition Latency

The average route acquisition latency for Secure Zone Routing Protocol (SZRP) is approximately 1.7 times as that of ZRP. For example, for 10 nodes moving at 5 m/s, it is 60ms as compared to 100ms for ZRP, while for 20 nodes moving at 10 m/s, it is nearly 135ms as compared to 75ms as in the case of ZRP. While processing SZRP routing control packets, each node has to verify the digital signature of the previous node, and then replace this with its own digital signature, in addition to the normal processing of the packet as done by ZRP. This signature generation and verification causes additional delays at each hop, and so the route acquisition latency increases [4].

2.2 SZRP2

The architectural design of SZRP2 is shown in Figure 3 that modified it by using four stages. First, an efficient key management mechanism used that is considered as a prerequisite for any security mechanism. Then, a secure neighbor detection scheme provided that relies on neighbor discovery, time and location based protocols. Securing routing packets is considered as the third stage which depends on verifying the authenticity of the sender and the integrity of the packets received. Finally, detection of malicious nodes mechanism is used to identify misbehaving nodes and isolate them using blacklist. Once these goals are achieved, providing confidentiality of transferred data becomes an easy task which can be implemented using any cryptography system [3].

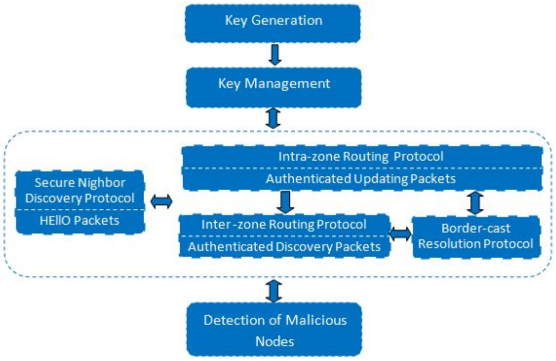

Key Generation
Key Management
Intra-zone Routing Protocol
Authenticated Updating Packets
Secure Nighbor Discovery Protocol
HEllO Packets
Inter-zone Routing Protocol
Authenticated Discovery Packets
Border-cast Resolution Protocol
Detection of Malicious Nodes

Figure 3. Architecture of SZRP2[3].

2.2.1 Performance Metrics

proposed protocol evaluated by comparing it with the current version of ZRP. Both protocols are run on identical movements and communication scenarios; the primary metrics used for evaluating the performance of SZRP are packet delivery ratio, routing overhead in bytes, routing overhead in packets, and end-to-end latency. These metrics are obtained from enhancing the trace files.

Packet delivery ratio: This is the fraction of the data packets generated by the CBR sources to those delivered to the destination. This evaluates the ability of the protocol to discover routes.

Routing overhead (bytes): This is the ratio of overhead bytes to the delivered data bytes. The transmission at each hop along the route is counted as one transmission in the calculation of this metric. The routing overhead of a simulation run is calculated as the number of routing bytes generated by the routing agent of all the nodes in the simulation run. This metric has a high value in secure protocols due to the hash value or signature stored in the packet.

Routing overhead (packets): This is the ratio of control packet overhead to data packet overhead over all hops. It differs from the routing overhead in bytes since in MANETs if the messages are too large, they will be split into several packets. This metric is always high even in unsecure routing protocols due to control packets used to discover or maintain routes such as IARP and IERP packets.

Average End-to-End latency: This is the average delay between the sending of data packet by the CBR source and its receipt at the corresponding CBR receiver. This includes all the delays caused during route acquisition, buffering and processing at intermediate nodes [3].

2.2.2 Simulation Results

proposed SZRP simulated over four scenarios to evaluate it through different movement patterns, network size, transmission rate, and radius of the zone.

2.2.2.1 Performance against Different Mobility Networks

In this scenario, The SZRP and ZRP compared over different values of the pause time. The pause time was changed from 100 s to 500 s to simulate high and low mobility networks. Concerning the packet delivery ratio as a function of pause time, the result shows that the packet delivery ratio obtained using SZRP is above 90% in all scenarios and almost similar to the performance of ZRP. This indicates that the SZRP is highly effective in discovering and maintaining routes for the delivery of data packets, even with relatively high mobility network (low pause time). A network with high mobility nodes has a lower packet delivery ratio because nodes change their location through transmitting data packets that have the predetermined path. For this reason, a high mobility network has a high number of dropped packets due to TTL expiration or link break. For the extra routing overhead introduced by both SZRP and ZRP, where the routing overhead is measured in bytes for both protocols, the results show that the routing overhead of SZRP is significantly higher and increased to nearly 42% for a high mobility network and 27% for a low mobility network. This is due to the increase in size of each packet from the addition of the digest and the signature stored in the packets to verify the integrity and authentication. This routing overhead decreases as the mobility decreases due to increase of the number of updating packets required to keep track of the changes in the topology in order to maintain routing table up-to-date. These packets include both IARP and IERP packets as well as the error messages.

2.2.2.2 Performance against Different Data Rates and Mobility Patterns

In this scenario, The SZRP and ZRP compared over different values of data rate. These values considered since high data rate is always an imperative need in any network although it has an extreme effect in increasing the congestion in MANETs. The data rate was changed from one to nine packets per second. These scenarios are performed under high and low mobility networks, 100 s and 500 s, respectively. Fig. 4 shows the packet delivery ratio of SZRP and ZRP for both low and high mobility networks. We note that the packet delivery ratio exceeds 89% in all cases which can be considered as a good indicator that SZRP goes in the same manner as the conventional ZRP. The delivery packet ratio of low mobility networks increases as the data rate increases as expected since the discovered route to the destination will not change during transmitting the packets, and thus the success of delivering the packet to the same destination will increase. On the other hand, the packet delivery ratio decreases in high mobility networks as the data rate increases because of the high probability of congestion by both the increased data packets and the increased control messages needed to maintain the network nodes up-to-date with the changeable topology.

2.2.2.3 Performance against Different Network Sizes and Mobility Patterns

The third scenario studies the performance of SZRP and ZRP over different network sizes. The number of nodes changes from ten to forty in order to validate our secure routing protocol in different networks. The experiments are performed under high and low mobility rates with data rate of five packets per second. To be consistent, the dimension of the topology used is changed with the same ratio as the number of mobile nodes. The SZRP still performs well in low mobility network where it exceeds 99%. However, its performance degrades in a high mobility network. In both cases, the result obtained is accepted because it degrades in the same manner as the conventional ZRP. A final point observed from this figure is that the packet delivery ratio decreases in a large network which is an expected result due to the increase of the traveling time that may lead to TTL expiration.

2.2.2.4 Performance against Different Routing Zones and Mobility Patterns

The last scenario studies the performance of both protocols under different routing zones. The number of routing zone nodes can be regulated through adjustments in each node’s transmitter power. To provide adequate network reachability, it is important that a node is connected to a sufficient number of neighbors. However, more is not necessarily better. As the transmitters’ coverage areas grow larger, so do the embership of the routing zones, an excessive amount of update traffic

may result [3].

3. CONCLUSION

The paper conducted a survey on the two various security improvements suggested for ZRP. An analysis is conducted on each improvement and the applications which best suits each enhancement is suggested. All protocols in standard mode, In terms of the network performance are acceptable. But there are some security problems. To solve these security problems for each of these algorithms, an extension is proposed. The extensions of the protocol's security problems have been resolved, But in terms of network performance problems have developed. Thus presentation an algorithm for ad hoc networks, both in terms of security and in terms of network performance is acceptable, it seems necessary. In evaluating the performance of both secure protocols, The results show that by increasing the routing overhead and average delay, packet delivery rate than the standard protocol is better. Both secure protocols to thwart further attacks at the network layer are suitable. The disadvantages of these two protocols failure to detect some attacks, such as jamming attack at the physical layer and the computational overhead is high.

According to Previous studies have reached conclude That all security protocols operate under different constraints and none of the protocols are not able to provide security for all purposes. Thus the design of new secure routing protocols against multiple attacks and to reduce the processing time in the process of identifying the problem still remains challenging.

4. REFERENCES

[1] Boora, S. et. al (2011). A Survey on Security Issues in Mobile Ad-Hoc Networks, International Journal of Computer Science & Management Studies, Vol. 11, Issue 02.

[2] Boukerche, A. et. al (2011). Routing protocols in ad hoc networks: A survey, Elsevier Computer Networks Journal, Vol. 55, Issue 13.

[3] Ibrahim, S. I. et. al (2012). Securing Zone Routing Protocol in Ad-Hoc Networks. I. J. Computer Network and Information Security, 10, 24-36.

[4] Kumar Pani, N. (2009) .A Secure Zone-Based Routing Protocol For Mobile Adhoc Network, thesis.

[5] Parvathavarthini, A. et. al (2013). An Overview of Routing Protocols in Mobile Ad-Hoc Network, International Journal of Advanced Research in Computer Science and Software Engg 3(2), February - 2013, pp. 251-259.

[6] Sudarsan, D. et. al (2012). A survey on various improvements of hybrid zone routing protocol in MANET, International Conference on Advances in Computing, Communications and Informatics Pages 1261-1265.

Different Types of Attacks and Detection Techniques in Mobile Ad Hoc Network

Mahsa Seyyedtaj
Department of computer, Shabestar branch,
Islamic Azad University, Shabestar,
Iran

Mohammad Ali Jabraeil Jamali
Department of computer, Shabestar branch,
Islamic Azad University, Shabestar,
Iran

Abstract: A Mobile Ad-Hoc Network (MANET) is a collection of mobile nodes (stations) communicating in a multi hop way without any fixed infrastructure such as access points or base stations. MANET has not well specified defense mechanism, so malicious attacker can easily access this kind of network. In this paper we investigate different types of attacks which are happened at the different layers of MANET after that we discuss some available detection techniques for these attacks. To our best knowledge this is the first paper that studies all these attacks corresponding to different layers of MANET with some available detection techniques.

Keywords: Security; Attacks; MANET; Prevention; Routing

1. INTRODUCTION

A MANET contains mobile nodes (stations) that can communicate with each other without the use of predefined infrastructure. There is not well defined administration for MANET. MANET is self organized in nature so it has rapidly deployable capability. MANET is very useful to apply in different applications such as battlefield communication, emergency relief scenario etc. In MANET nodes are mobile in nature, due to the mobility, topology changes dynamically. Due to its basic Ad-Hoc nature, MANET is venerable to various kinds of security attacks [1].

2. SECURITY GOALS FOR MANET

The ultimate goal of the security solutions for MANET is to provide a framework covering availability, confidentially, integrity, authentication and non-repudiation to insure the services to the mobile user. A short explanation about these terms:-

2.1 Availability

ensures the survivability of network services despite denial of service attacks. The adversary can attack the service at any layer of an ad hoc network. For instance, at physical and media control layer it can employ jamming to interfere with communication on physical channels; on network layer it could disrupt the routing protocol and disconnect the network; or on higher layers it could bring down some high-level services (e.g., the key management service).

2.2 Confidentiality

ensures that certain information is never disclosed to unauthorized entities. It protects the network transmission of sensitive information such as military, routing, personal information, etc.

2.3 Integrity

guarantees that the transferred message is never corrupted. A corruption can occur as a result of transmission disturbances or because of malicious attacks on the network.

2.4 Authentication

enables a node to ensure the identity of the peer node with whom it is communicating. It allows manipulation-safe identification of entities (e.g., enables the node to ensure the identity of the peer node), and protects against an adversary gaining unauthorized access to resources and sensitive information, and interfering with the operation of other nodes.

2.5 Non-repudiation

ensures that the origin of a message cannot later deny sending the message and the receiver cannot deny the reception. It enables a unique identification of the initiator of certain actions (e.g., sending of a message) so that these completed actions can not be disputed after the fact [11].

3. TYPES OF SECURITY ATTACKS

3.1 On the basis of nature

3.1.1 Passive attacks

In passive attack there is not any alteration in the message which is transmitted. There is an attacker (intermediated node) between sender & receiver which reads the message. This intermediate attacker node is also doing the task of network monitoring to analyze which type of communication is going on.

3.1.2 Active attacks

The information which is routing through the nodes in MANET is altered by an attacker node. Attacker node also streams some false information in the network. Attacker node also do the task of RREQ (re request) though it is not an authenticated node so the other node rejecting its request due these RREQs the bandwidth is consumed and network is jammed.

3.2 On the basis of domain

3.2.1 External attacks

In external attack the attacker wants to cause congestion in the network this can be done by the propagation of fake routing information. The attacker disturbs the nodes to avail services.

3.2.2 Internal attacks

In internal attacks the attacker wants to gain the access to network & wants to participate in network activities. Attacker does this by some malicious impersonation to get the access to the network as a new node or by directly through a current node and using it as a basis to conduct the attack [12].

4. ATTACKS CORRESPONDING TO DIFFERENT LAYERS IN MANET

First of all let we explain how many layers are there in MANET stack. Basically there are five layers i.e. application layer, transport layer, network layer, Mac layer, & physical layer [3].

4.1 Attacks at application layer

4.1.1 Repudiation attack

Due to repudiation attack deny of participation is happened in whole communication, or in a part of communication [8].

4.1.2 Attack by virus & worms

Attack is done by virus, worms to infect the operating system or application software installed in mobile devices [2].

4.2 Attacks at transport layer

4.2.1 TCP SYN attack (Denial of service attack)

TCP SYN attack is DOS in nature, so the legitimate user does not get the service of network when attack is happened. TCP SYN attack is performed by creating a large no of halt in opened TCP connection with a target node [3].

4.2.2 TCP Session Hijacking

TCP session hijacking is done by the spoofing of IP address of a victim node after that attacker steals sensitive information which is being communicated. Thus the attacker captures the characteristics of a victim node and continues the session with target [6].

4.2.3 Jelly Fish attack

Similar to the blackhole attack, a jellyfish attacker first needs to intrude into the forwarding group and then it delays data packets unnecessarily for some amount of time before forwarding them. This results in significantly high end-to-end delay and delay jitter, and thus degrades the performance of real-time applications. [9].

4.3 Attacks at network layer

4.3.1 Flooding attack (Denial of service attack)

Attacker exhausts the network resources, i.e. bandwidth and also consumes a node's resources, i.e. battery power to disrupt the routing operation to degrade network performance. A malicious node can send a large no. of RREQ (re request) in short duration of time to a destination node that dose not exist in the network. Because no one will replay to these RREQ so they will flood in the whole network. Due to flooding the battery power of all nodes as well as network bandwidth will be consumed and could lead to denial of service [7].

4.3.2 Route tracking

This kind of attack is done to obtain sensitive information which is routed through different intermediate nodes [8].

4.3.3 Message Fabricate, modification

In this kind of attack false stream of messages is added into information which is communicated or some kind of change is done in information [13].

4.3.4 Blackhole attack

In a blackhole attack a attacker node sends fake routing information in the network to claims that it has an optimum route and causes other good nodes to route data packets through the malicious one. For example in an Ad-Hoc on demand distance vector routing (AODV), attacker can send fake RREQs including a fake destination sequence number that is fabricated to be equal or higher than the one contain in the RREQ to source node, claiming that it has a sufficient fresh route to the destination node. This causes the source node to select the route that passes through the attacker node. Therefore all the traffic will be routed through the attacker and therefore, the attacker can misuse the information or sometime discard the traffic [1].

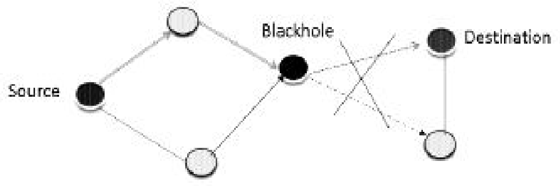


Figure 1. Blackhole attack

4.3.5 Wormhole attack

It is the dangerous one among the all attacks. In this attack, a pair of colluding attackers recodes packets at one location and replays them at another location using a private high speed network [5]. The seriousness of this attack is that it can be launched in all communication that provides authenticity & confidentiality.

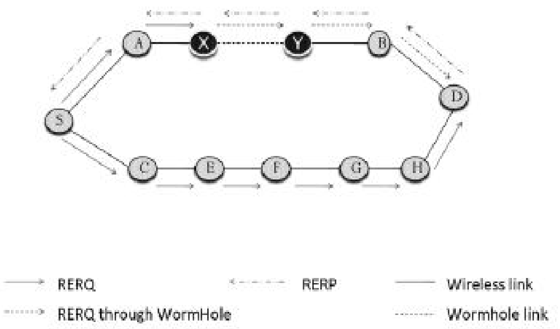


Figure 2. Wormhole attack

4.3.6 Grayhole attack

A variation of black hole attack is the gray hole attack, in which the nodes will drop the packets selectively. Selective forward attack is of two types they are

- Dropping all UDP packets while forwarding TCP packets.
- Dropping 50% of the packets or dropping them with a probabilistic distribution. These are the attacks that seek to disrupt the network without being detected by the security measures [8].

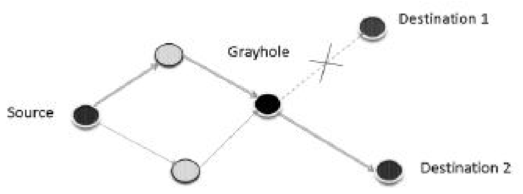


Figure 3. Grayhole attack

4.3.7 Rushing attack

Many demand-driven protocols such as ODMRP, MAODV, and ADMR, which use the duplicate suppression mechanism in their operations, are vulnerable to rushing attacks. When source nodes flood the network with route discovery packets in order to find routes to the destinations, each intermediate node processes only the first non-duplicate packet and discards any duplicate packets that arrive at a later time. Rushing attackers, by skipping some of the routing processes, can quickly forward these packets and be able to gain access to the forwarding group [4].

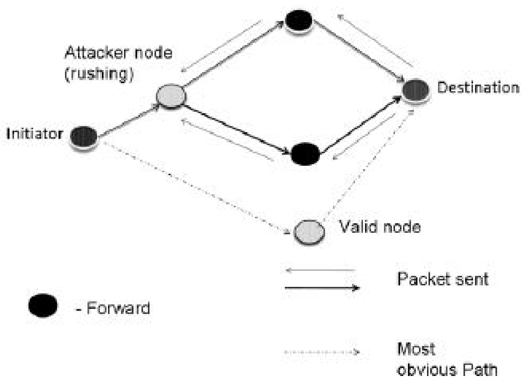


Figure 4. Rushing attack

4.3.8 Link spoofing attack

In a link spoofing attack, a malicious node advertises fake links with non-neighbors to disrupt routing operations. An attacker can advertise a fake link with a target's two-hop neighbors. This causes the target node to select the malicious node to be its multipoint relay node (MPR). As an MPR node, a malicious node can then manipulate data or routing traffic, i.e. modifying or dropping the routing traffic. They can also perform some other types of DOS attacks [13].

4.3.9 Byzantine attack

Byzantine attack can be launched by a single malicious node or a group of nodes that work in cooperation. A compromised intermediate node works alone or set of compromised intermediate nodes works in collusion to form attacks. The compromised nodes may create routing loops, forwarding packets in a long route instead of optimal one, even may drop packets. This attack degrades the routing performance and also disrupts the routing services [8].

4.3.10 Sybil attack

A Sybil attack is a computer hacker attack on a peer-to-peer (P2P) network. It is named after the novel Sybil, which recounts the medical treatment of a woman with extreme dissociative identity disorder. The attack targets the reputation system of the P2P program and allows the hacker to have an unfair advantage in influencing the reputation and score of files stored on the P2P network. Several factors determine how bad a Sybil attack can be, such as whether all entities can equally affect the reputation system, how easy it is to make an entity, and whether the program accepts non-trusted entities and their input. Validating accounts is the best way for administrators to prevent these attacks, but this sacrifices the anonymity of users [10].

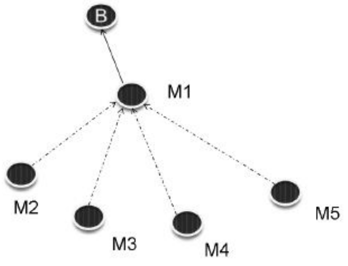


Figure 5. Sybil attack

4.4 Attacks at MAC layer

4.4.1 MAC Denial of service attack (DOS)

At the MAC layer DOS can be attempted as:

There is a single channel which is used frequently, keeping the channel busy around a particular node leads to a denial of service attack at that node.

An attacker node continuously sends spurious packets to a particular network node this leads to drain the battery power of the node, which further leads to a denial of service attack.

4.4.2 Traffic monitoring & Analysis

Traffic analysis is a passive type of attack in nature this kind of analysis is done by attacker to find out which type of communication is going on.

4.4.3 Bandwidth Stealth

In this kind of attack the attacker node illegally stealth the large fraction of bandwidth due to this congestion is happened in the network.

4.4.4 MAC targeted attack

MAC layer plays an important role in every piece of data that is exchanged through several nodes, ensuring that data is collected efficiently to its intended recipient. The MAC targeted attacks disrupt the whole MAC procedure [13].

4.4.5 WEP targeted attacks

The wired equivalent privacy (WEP) is designed to enhance the security in wireless communication that is privacy and authorization. However it is well known that WEP has number of weaknesses and is subject to attacks. Some of them are:-

1. WEP protocol does not specify key management.

2. The initialization vector (IV) is a 24 bit field which is the part of the RC4 encryption key. The reuse of IV and weakness of RC4 help to produce analytic attacks.

3. The combined cure of non cryptographic integrity algorithm, CRC32, with the stream cipher has a security risk [11].

4.5 Attacks at physical layer

4.5.1 Jamming attack (Denial of service attack)

DOS attack is also happened at physical layer. Due to DOS there is denial of services accessed by a legitimate network user. Example is jamming attack.

Due to jamming & interference of radio signals messages can be lost or corrupt. Signals generated by a powerful transmitter are strong enough to overwhelm the target signals and can disrupt communication. Pulse and random noise are most common type of signal jamming [3].

4.5.2 Stolen or compromised attack

These kinds of attacks are happened from a compromised entities or stolen device like physical capturing of a node in MANET.

4.5.3 Malicious message injecting

Attacker inject false streams into the real message streams which is routing through the intermediate nods, due to malicious message injecting the functionality of network is disrupted by the attacker.

4.5.4 Eavesdropping attack

Eavesdropping is the reading of messages and conversation by unintended receivers. The nodes in MANET share a wireless medium and the wireless communication use RF spectrum and broadcast by nature which can easily intercepted with receivers tuned to proper frequency. As a result transmitted messages can be overheard as well as fake messages can be injected into the network [3].

Table1. Attacks corresponding to different layers

MANET Layer	Type of Attack
Application Layer	Repudiation attack, Attacks by virus & worms
Transport Layer	TCP SYN attack (DOS in nature), TCP session hijacking, Jelly Fish attack
Network Layer	Flooding attack, Route tracking, Message Fabricate, modification, Blackhole attack, Wormhole attack, Link spoofing attack Grayhole attack, Rushing attack, Byzantine attack, Sybil attack
MAC Layer	Mac DOS (Denial of service) attack, Traffic monitoring & analysis, Bandwidth stealth, MAC targeted attack, WEP targeted attack
Physical Layer	Jamming attack (DOS in nature), Stolen or compromised attack, Malicious massage injecting, Eavesdropping attack

5. DETECTION TECHNIQUES

There are some schemes which are used to secure the MANET & in the detection of anomalies. Some of these are discussed below:-

5.1 Intrusion Detection Technique

IDS detect different threats in MANET communication There is proposed architecture [1] for IDS which is used by MANET given below:-

In the proposed architecture of IDS for MANET every node participates in the detection process and responds to activities. This detection process is done by detecting the intrusion behavior in the two ways:-

a). Locally

b). Independently

This act is performed by an agent who is known as IDS agent who is inbuilt in all devices (stations). Each node performs detection locally and independently but there is also a situation if a node detects an anomaly but it has not sufficient investigation results to figure out which type of anomaly it is, so it share its result to the other nodes in the communication range and ask them to search this anomaly in their respective security logs to trace out the possible characteristics of that intruder.

There are four functional modules in conceptual model of the IDS:-

5.1.1 Local data collection module

Local data collection module deals with data gathering issues. Data come from various resources through a real time data audit.

5.1.2 Local detection engine

It inspects any anomaly shown in the data which was collected by local data collection modules. This detection engine rely on the statistical anomaly detection technique which distinguish anomaly in the basis of the comparison which is done by taking a deviation between the current observation data and the normal profile (generated on the basis of normal behavior of the system) of system.

5.1.3 Cooperative detection engine

All time it is not possible the attacks which are happened on MANET known to the system (IDS). So there is some need to find more evidence for particular attack, so we have to initiate a cooperative detection process in these circumstances. In cooperated detection process participants will share the information regarding the intrusion detection to all their neighboring nodes. On the basis of information received a node can calculate new intrusion state. In this process they used certain algorithms such as a distributed consensus algorithm with weight. We may assume that the majority of node in MANET are actual (are not attacker nodes) so we can trust the results produced by any of the participants that the network is under attack.

5.1.4 Intrusion response module

When an intrusion is confirmed intrusion response module will response to that. It responses to reinitialize the communication channel. Re-initialization is done such as reassigning the key or reorganizing the network. In reorganization of the network we remove all the compromised nodes. This response varies corresponding to different kind of intrusion.

5.2 Cluster-Based Intrusion Detection Technique [13]

We have discussed cooperative intrusion detection architecture for the ad hoc network in the previous part which has some drawbacks. In cooperative intrusion detection technique there is mechanism of participation of all nodes in detection process which cause huge power consumption for all the participating nodes.

In MANET power supply is limited which may cause some node may behave in selfish way i.e. they are not cooperative with other nodes to save their battery power. So the actual aim is violet in cooperative intrusion detection mechanism. To solve this problem a cluster based intrusion detection technique is used. In this technique MANET can organized into number of clusters. The organization is done in such a way that every node is a member of at least one cluster and there will be only one node per cluster that will take the responsibility of monitoring. In a certain period of time this node is known as cluster head. A cluster contain several node that reside within the same radio range with each other, so when a node is selected as cluster head all the nodes in this cluster should be within 1-hop distance. When a cluster selection process is going on there is the necessity to ensure two things:-

- aFairness.
- Efficiency.

5.2.1 Fairness

Fairness contains two levels of meanings: the probability of every node in the cluster head should be equal and each node should act as the cluster node for the same amount of time.

5.2.2 Efficiency

Efficiency of cluster head selection process means that there should be some method that can select a node from the cluster periodically which has high efficiency. Cluster information is used in cluster based intrusion detection technique. Basically there are four states in the cluster information protocol:-

1. Initial

2. Clique.

3. Done

4. Lost.

At the beginning all nodes are at initial state. In initial state node will monitor their own traffic and detects intrusion behavior independently. There are two steps that we need to finish before we get the cluster head of the network:-

- Cluster computation.
- Cluster head computation.

A cluster is a group of nodes in which every pair of member can communicate via direct wireless link. Once the protocol is finished every node is aware of fellow clique member. Then a node will randomly select from the queue to act as the cluster head. There are two other protocols that assist the cluster to do some validation and recovery which are:-

- Cluster valid assertion protocol.
- Cluster recovery protocol.

5.2.2.1 Cluster valid assertion protocol:-

It is generally used in following two situations This protocol is used by a node to check if the connection between the cluster head and itself is maintained or not. The node does this task periodically. If connection is not maintained the node will check to see if it belong to another cluster, and if in this situation it also get a negative answer then the node draw a conclusion and will enter into the LOST state and initiate a routing recovering request.

To keeps the fairness and security in the whole cluster a mandatory reelection time out is also needed for the cluster head. If the time out expires, all the nodes switch from DONE state to INITIAL state, thus they begin a new round of cluster head election.

5.2.2.2 Cluster recovery protocol:-

It is mainly used in a case when a node losses its connection with previous cluster head, for a cluster head losses all its connected stations than they enter into LOST state and initiate cluster recovery protocol to elect a new cluster head.

5.3 Misbehavior detection through cross layer analysis [13]

In some cases attacker attacks on multiple layer of MANET simultaneously but they keep the attack stay below the detection threshold so as to escape from detection by the single-layer misbehavior detector. This kind of attack is also called as cross-layer attack. So cross-layer attacks are more threatening to a single-layer detector because they can be easily skipped by the single-layer misbehavior detector. So we have to used some different techniques in these circumstances, this attack scenario can be detected by cross layer misbehavior detector. In this technique the inputs from all layer of MANET stack are combined and analyzed by the cross layer detector. But a problem is arisen here, how to make the cross layer detection more effective and efficient, how to cooperate between single-layer detectors to make the detection process effective. Single-layer detectors deal with attacks to corresponding layers, so we have to take some different viewpoints in these circumstances when a single attack is observed in different layers of MANET. So it is necessary to clubbed out the different results produced by different layers to make a possible solution. There is second thing, we need to find out how much the system resources and network overhead will be increased due to the use of cross layer detector compared with the original single layer detector. Limited battery power of the nodes in MANET is also an issue here, the system and network overhead brought by the cross layer detection should be consider and compared with the performance gain caused by the use of cross layer detection technique.

6. CONCLUSION

In this paper, we try to inspect the security attacks at different layers of MANET, which produces lots of trouble in the MANET operations. Due to the dynamic nature of MANET it is more prone to such kind of attacks. In MANET the solutions are designed corresponding to specific attacks they work well in the presence of these attacks but they fail under different attack scenario.

Therefore, our aim is to develop a multi-functional security system for MANET, which will cover multiple attacks at a time and also some new attacks.

7. FUTURE WORK

This paper can be further extended to give the solutions corresponding to these attacks which we discussed at different layers of MANET, we can add more detection techniques if it is possible to invent them.

8. REFERENCES

[1] Boora, S. et. al (2011). A Survey on Security Issues in Mobile Ad-Hoc Networks, International Journal of Computer Science & Management Studies, Vol. 11, Issue 2.

[2] Biswas, K. et. al (2007). Security threats in Mobile Ad-hoc Network, Master theses, Department of Interaction & System Design, Blekinge Institute of Technology, Sweden.

[3] Gua, Y. (2008). a dissertation on Defending MANET against flooding attacks by detective measures, Institute of Telecommunication Research, The University of South Australia.

[4] Hu,Y-C. et. al (2003). Rushing Attacks and Defense in Wireless Ad Hoc Network Routing Protocols, Proceedings of ACM WiSe 2003, San Diego, CA.

[5] Hu,Y-C. et. al (2006). Wormhole attacks in Wireless Networks, IEEE JSAC, Vol. 24, No. 2.

[6] Ishrat, Z. (2011). Security issues, challenges & solution in MANET, IJCST, Vol. 2, Issue 4.

[7] Khokhar, R. et. al (2008). A review of current routing attacks in Mobile Ad-Hoc Networks, International Journal of Computer Science & Security, Vol. 2, Issue 3.

[8] Mamatha, G. S. et. al (2010). Network Layer Attacks and Defense Mechanisms in MANETS- A Survey, International Journal of Computer Applications, Vol. 9, No. 9.

[9] Nguyen, H. et. al (2006). Study of Different Types of Attacks on Multicast in Mobile Ad Hoc Networks, International Conference on Mobile Communications and Learning Technologies.

[10] Pandey, A. et. al (2010). A Survey on Wireless Sensor Networks Security, International Journal of Computer Applications, Vol. 3, No. 2.

[11] Rai, P. et. al (2010). A Review of MANETs Security Aspects and Challenges, IJCA Special Issue on “Mobile Ad-hoc Networks”.

[12] Sivakumar, K. et. al (2013). overview of various attacks in manet and countermeasures for attacks, International Journal of Computer Science and Management Research, Vol. 2.

[13] Wazid, M. et. al (2011). A Survey of Attacks Happened at Different Layers of Mobile Ad-Hoc Network & Some Available Detection Techniques, International Conference on Computer Communication and Networks CSI-COMNET.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 9, 547 - 549, 2014, ISSN: 2319–8656

A Review on a web based Punjabi to English Machine Transliteration System

Navpreet kaur
C.S.E Department
G.Z.S PTU Campus
Bathinda, India

Paramjeet Singh
C.S.E Department
G.Z.S PTU Campus
Bathinda, India

Shveta Rani
C.S.E Department
G.Z.S PTU Campus
Bathinda, India

Abstract: The paper presents the transliteration of noun phrases from Punjabi to English using statistical machine translation approach.Transliteration maps the letters of source scripts to letters of another language.Forward transliteration converts an original word or phrase in the source language into a word in the target language.Backward transliteration is the reverse process that converts the transliterated word or phrase back into its original word or phrase.Transliteration is an important part of research in NLP.Natural Language Processing (NLP) is the ability of a computer program to understand human speech as it is spoken.NLP is an important component of AI.Artificial Intelligence is a branch of science which deals with helping machines find solutions to complex programs in a human like fashion.The transliteration system is going to developed using SMT.Statistical Machine Translation (SMT) is a data oriented statistical framework for translating text from one natural language to another based on the knowledge.

Keyword: Transliteration,Mapping,Translation,Dictionary

1. INTRODUCTION

Transliteration is a process that maps the sounds of one language to scripts of another language.The system performs the process of transliteration of noun phrases of Punjabi to English using SMT approach.Punjabi Language is written from left to right using gurmukhi script and Punjabi language consist of consonents, vowels, halant, punctuation and numerals.The gurmukhi script was derived from sharda script.The Punjabi Language contains Thirty-five distinct letters.English language is written in roman scripts.There are 26 letters in English.Out of which 21 is consonants and 5 are vowels.Punjabi language is an official language of Punjab.It can be understand or read by the person who knows Punjabi.Opposite to it English is an international language.so the person who have no knowledge about Punjabi can convert the file Written in Punjabi into English using Punjabi to English transliteration system.SMT uses the concept of development of Machine learning system from the existing names stored in the database system.Development of database table for uni-gram,bi-gram,tri-gram,four-gram,five-gram,six-gram and upto ten-gram to store the results obtained from the learning phase of the system.Various algorithms for conversion of anmollipi into Unicode is used so that it can be used as input to the system This topic of machine transliteration has been used in different language to convert from one language to another language.Various techniques has been applied to this system Diect mapping like rule based approach etc.Transliteration is different from Translation system.Translation from Punjabi to English means to translate each word in Punjabi to its English equivalent whereas the transliteration means to write them sensing the characters in the word e.g. “nvdIp “in Punjabi is transliterated in English as “navdeep” where n for “n” v for “v” d for “d” p for “p” .This system can be developed using transliteration process using a database of transliterating characters.To develop this system first of all we have to collect names of proper nouns from various sources such as person names,cities rivers,countries,states etc.We have to store these names in Punjabi and its English equivalent in database.Then we have to develop an algorithm to convert the Punjabi font into Unicode so that it can be given as input to the system.Then to develop the algorithm for learning phase of the system.The system will learn from existing data entries.

Three Main Approaches are used for machine translation:

Direct Machine Translation (DMT) system is a simple form of machine translation system. In DMT, a word to word translation of the input text is performed and the result is obtained in the DMT, a language which is called a source language (Punjabi) is given as input and the output is received which is called a target form of output text.

Rule Based Machine Translation (RBMT) is also known as Knowledge Based Machine Translation system. It is a system which is based on linguistic infomation related to source and target languages and retrieves this information from dictionaries (bilingual) and grammars which includes semantic and syntactic information of each language. RBMT system generates output text from this information.

Statistical Machine Translation (SMT) is a new approach which is based on statistical models and in this approach; a word is translated to one of a number of possibilities based on the probability. The whole process is performed by dividing sentences into N-grams. N-gram is a contiguous sequence of n items from a given text. The items can be phonemes, letters, and words. An N-gram of size 1 is known as a unigram; size 2 is a bigram; size 3 is a trigram. Larger sizes are represented by the value of n i.e. four-gram, five-gram and so on. Statistical system will analyze the position of N-grams in relation to one another within sentences.

2. EXISTING WORK

Transliteration and translation has been studied in different languages.These systems has been developed in different languages pairs.We have studied different literature related to transliteration system.Gurpreet singh josan and gurpreet singh lehal has developed Punjabi to hindi machine transliteration system by combining character to character mapping using rule based approach.This paper shows that the system produced transliteration in hindi from Punjabi with an accuracy of 73% to 85%.Vishal goyal and Gurpreet singh has developed hindi to Punjabi machine translation system using the rule based techniques.The overall efficiency of this system hindi to Punjabi is 95%.Another system has been developed by Kamaldeep and Dr. Vishal goyal of using hybrid approach for Punjabi to English transliteration system.This paper presents the Punjabi to English machine transliteration using letter to letter mapping as baseline and try to find out the improvements by statistical methods.To improve the accuracy various rules has been developed.Author has developed hybrid (statistical + rules) approach based transliteration system.Independent vowel mapping,dependent vowel mapping,consonant mapping,mapping of special symbols table is defined.The Overall accuracy of the system comes out to be 95.23%.Kamaljeet kaur batra and G.S.Lehal has developed rule based machine translation of noun phrases from punjabi to English.The paper presents the automatic translation of noun phrases from Punjabi to English using transfer approach.The system has analysis,translation and synthesis components.The steps involved are preprocessing,tagging,ambiguity resolution,translation and synthesis of words in target language.The accuracy is calculated for each step and the overall accuracy of the system is calculated to be about 85% for a particular type of noun phrases

3. PROBLEM

The problem domain to which this project is concerned is machine transliteration.In foreign and in some areas of india other than Punjab,most of population is not so familiar with Punjabi.As we know that all the data of government sector of Punjab is in Punjabi language because Punjabi is an official language of Punjab,people who are unaware of Punjabi can't understand it.For e.g.punjab state government has to send the report of malnutrition children to UNO.As all the reports are generally created in Punjabi language but it is not useful in foreign so there is a need to present it in English language,here the transliteration system is useful.Existing systems has been developed with mostly rule based techniques and hybrid techniques.we can't make as many rules as possible.We can develop this system with the help of SMT technique which can increase the efficiency of the system.In existing system some errors are occur e.g. sometimes when a name is pronounced in Punjabi it correspond to many English words. e.g."rxjIq" is convert in english as ranjeet,ranjit.So that system fail to guess which one is the best.Sometime user does not enter correct data due to which output is also not correct.e.g."mRIq" it is wrongly enter data we cannot use "R" with "m" in Punjabi language.Another issue related to the difference in the number of characters in Punjabi and English languages.There is a difference in the number of vowels and consonents.Sometime single character to multiple mapping are occur e.g. "v" can be used as v,w.So there is a need to develop algorithm to select the appropriate character at different situations.Existing system is developed on the bases of direct and rule based approach.They are using direct approach due to which the accuracy of system is very low.

4. CONCLUSION

In this paper we have discussed about the transliteration system which has been developed in different languages.Different techniques has been used to develop this system.the accuracy of each system is studied.The paper has addressed the problem arising in transliteration of Punjabi to English.This system can be developed with additional efforts.There are many issues left for further improvement.the system could be improved by improving the techniques.The system can be effectively developed with the help of using SMT technique.SMT take the view that every sentence in the target language is a translation of the source language sentence with some probability.The best translation is the sentence that has highest probability.The system can be develop by using database table for uni-gram,bi-gram and upto ten-gram to store the results obtain from the learning phase of the system.In Punjab state most of the official work is done in Punjabi language,so this transliteration system will help them a lot to transliterate Punjabi to English.

5. REFERENCES

[1] Gurpeet Singh josan and Gurpreet Singh lehal,A Punjabi to Hindi machine transliteration system,Computational Linguistics and Chinese language processing vol.15.no.2.june 2010 ,pp.77-102

[2]Vishal Goyal and Gurpreet Singh Lehal,Evaluation of hindi to Punjabi machine translation system,IJCSI international Journal of computer science issues,vol.4.no.1,2009 ISSN(Online):1694-0784

[3]Kamaldeep ,Dr.vishal Goyal,hybrid approach for punjabi to English transliteration system International journal of computer applications (0975-8887) volume 28-no.1,August 2011.

[4]Sumita rani,Dr.Vijay Laxmi,A review on machine Transliteration of related languages:Punjabi to Hindi

international journal of science,Engineering and technology research (IJSETR) volume 2,issue 3,march 2013

[5]Gurpreet Singh Josan1& Jagroop Kaur, "Punjabi to Hindi statistical machine transliteration" International Journal of Information Technology and Knowledge Management July-December 2011, Volume 4, No. 2, pp. 459-463.

Mathematical Approach to Complexity-Reduced Antenna Selection Technique for Achieving High Channel Capacity

Priya Dhawan
Department of ECE
Amritsar College of Engineering and Technology
Amritsar-143001,Punjab,India

Narinder Sharma
Department of EEE
Amritsar College of Engineering and Technology
Amritsar-143001,Punjab,India

Abstract: In this paper channel state information is exploited for improving system performance. The performance parameters of the Multiple Input Multiple Output system is better and are even achieved using additional RF modules that are required as multiple antennas are employed. To reduce the cost associated with the multiple RF modules, antenna selection techniques can be used to employ a smaller number of RF modules than the number of transmit antennas. The exploiting of information for complexity reduced antenna selection is performed for achieving high channel capacity. Simulation results show that the channel capacity increases in proportion to the number of the selected antennas.

Keywords: MIMO systems, RF modules, Antenna Selection, Channel State Information, Signal to Noise Ratio.

1. INTRODUCTION

In typical digital communication system, Signal parameters on which multipath channel have effect that are independent path gain, independent path frequency offset, independent path phase shift, independent path time delay etc. To remove ISI from the signal, many kinds of equalizers can be used. Different techniques are used to handle the changes made by the channel,receiver requires knowledge over CIR to combat with the received signal for recovering the transmitted signal. CIR is provided by the separate channel estimator. Usually channel estimation is based on the known sequence of bits, which is unique for a certain transmitter and is repeated in every transmission burst. Which enables the channel estimator to estimate CIR for each burst separately by using the known transmitted signal and the corresponding received signal. Multiple Input Multiple Output (MIMO) systems takes advantage of multipath propagation signals by sending and receiving more than one data signal in the same frequency band at the same time by using multiple transmit and receive antennas. Orthogonal frequency division multiplexing (OFDM) is also has capability to handle the effect of ISI and Inter carrier interference (ICI). OFDM converts the frequency selective wide band signal into frequency flat multiple orthogonally spaced narrow band signals also resulting in high bandwidth efficiency [1].

2. ANTENNA SELECTION TECHNIQUE

The antenna selection technique is one of the major issue that is to be taken care in the communication system. MIMO systems have better performance which can be achieved without using additional transmit power or bandwidth extension.[2] However, it requires additional high-cost RF modules are required as multiple antennas are employed. In general, a transmitter does not have direct access to its own channel state information. Therefore, some indirect means are required for the transmitter. In time division duplexing system, we can exploit the channel reciprocity between opposite links (downlink and uplink). Based on the signal received from the opposite direction, it allows for indirect channel estimation. In frequency division duplexing (FDD) system, which usually does not have reciprocity between opposite directions, the transmitter relies on the channel feedback information from the receiver. In other words, CSI must be estimated at the receiver side and then, fed back to the transmitter side. To reduce the cost associated with the multiple RF modules, antenna selection techniques can be used to employ a smaller number of RF modules than the number of transmit antennas. Figure 1 illustrates the end-to-end configuration of the antenna selection in which only Q RF modules are used to support N_T transmit antennas since Q RF modules are selectively mapped to Q of N_T transmit antennas.[2]

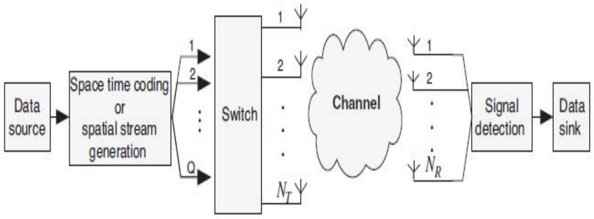


Figure 1: Antenna selections with Q RF modules and N_T transmit antennas $\left(Q < N_T\right)$ [10]

Since Q antennas are used among N_T transmit antennas, the effective channel can now be represented by Q columns of

$H \in \Re^{N_R \times N_T}$. Let p_i denote the index of the ith selected column, $i = 1, 2, \cdots, Q$. Then, the corresponding effective channel will be modeled by $N_R \times Q$ matrix, which is denoted by $H_{\{p_1, p_2, \cdots p_Q\}} \in \Re^{N_R \times Q}$ [3]. Let $X \in \Re^{Q \times 1}$ denote the space-time-coded or spatially-multiplexed stream that is mapped into Q selected antennas. Then, the received signal y is represented as

$$y = \sqrt{\frac{E_X}{Q}} H_{\{p_1, p_2, \cdots p_Q\}} X + Z$$

(1)

where $z \in \Re^{N_R \times 1}$ is the additive noise vector. The channel capacity of the system in Equation (1) will depend on the number of transmit antennas that are chosen.

3. COMPLEXITY-REDUCED ANTENNA SELECTION TECHNIQUE

The Complexity-Reduced Antenna Selection Technique is one of the type of antenna selection technique. As compared to the optimal antenna technique ,complexity reduced antenna selection technique is better. Optimal antenna selection requires too much complexity depending on the total number of available transmit antennas. In order to reduce its complexity, we proposed a sub-optimal method. We adopted an approach in which additional antenna is selected in ascending order of increasing the channel capacity i.e., one antenna with the highest capacity is first selected as

$$p_1^{subopt} = \arg\max_{p_1} C_{\{p_1\}} \qquad (2)$$

$$= \arg\max_{p_1} \log_2 \det\left(I_{N_R} + \frac{E_X}{QN_0} H_{\{p_1\}} H_{\{p_1\}}^H \right)$$

Given the first selected antenna, the second antenna is selected such that the channel capacity is maximized i.e.

$$p_2^{subopt} = \arg\max_{p_2 \neq p_1^{subopt}} C_{\{p_1^{subpot}, p_2\}}$$

$$= \arg\max_{p_2 \neq p_1^{subopt}} \log_2 \det\left(I_{N_R} + \frac{E_X}{QN_0} H_{\{p_1^{subpot}, p_2\}} \right)$$

After the nth iteration which provides $\left\{ p_1^{subopt}, p_2^{subopt}, \cdots p_n^{subopt} \right\}$, the capacity with an additional antenna, say antenna l, can be updated as

$$C_l = \log_2 \det\left\{ I_{N_R} + \frac{E_X}{QN_0} H_{\{p_1^{subopt}, p_2^{subopt}, \cdots p_n^{subopt}\}} H^H_{\{p_1^{subopt}, p_2^{subopt}, \cdots p_n^{subopt}\}} + H_{\{l\}} H^H_{\{l\}} \right\}$$

$$= \log_2 \det\left\{ I_{N_R} + \frac{E_X}{QN_0} H_{\{p_1^{subopt}, p_2^{subopt}, \cdots p_n^{subopt}\}} H^H_{\{p_1^{subopt}, p_2^{subopt}, \cdots p_n^{subopt}\}} \right\}$$

$$+ \log_2 \left\{ 1 + \frac{E_X}{QN_0} H_{\{l\}} \left(I_{N_R} + \frac{E_X}{QN_0} H_{\{p_1^{subopt}, p_2^{subopt}, \cdots p_n^{subopt}\}} H^H_{\{p_1^{subopt}, p_2^{subopt}, \cdots p_n^{subopt}\}} \right)^{-1} H^H_{\{l\}} \right\}$$

It can be derived using the following identities:

$$\det\left(A + uv^H \right) = \left(1 + V^H A^{-1} u \right) \det(A)$$

$$\log_2 \det\left(A + uv^H \right) = \log_2 (1 + V^H A^{-1} u) \det\left(A \right) + \log_2 \left(1 + V^H A^{-1} u \right)$$

Where

$$A = I_{N_R} + \frac{E_X}{QN_0} H_{\{p_1^{subopt}, p_2^{subopt}, \cdots p_n^{subopt}\}} H^H_{\{p_1^{subopt}, p_2^{subopt}, \cdots p_n^{subopt}\}}$$

$$u = v = \sqrt{\frac{E_X}{QN_0}} H_{\{l\}}$$

The additional (n+1) th antenna is the one that maximizes the channel capacity , that is,

$$P_{n+1}^{subopt} = \arg\max_{l \notin \{p_1^{subopt}, p_2^{subopt}, \cdots p_n^{subopt}\}} C_l$$

This process continues until all Q antennas are selected.

Also the same process can be implemented by deleting the antenna in descending order of decreasing channel capacity. Let S_n denote a set of antenna indices in the nth iteration. In the initial step, we consider all antennas, $S_l = \{1, 2, \cdots, N_T\}$, and select the antenna that contributes least to the capacity, that is,

$$p_1^{deleted} = \arg\max \log_2 \det\left(I_{N_R} + \frac{E_X}{QN_0} H_{S_1-\{p_1\}} H^H_{S_1-\{p_1\}} \right)$$

A good literature on exploitation of CSI for channel estimation and the types of antenna selection techniques can be found in [4-15].The antenna selected from above Equation will be deleted from the antenna index set, and there remaining antenna set is updated to $S_2 = S_1 - \left\{ p_1^{deleted} \right\}$. If $|s_2| = N_T - 1 > Q$ we choose another antenna to delete. This will be the one that contributes least to the capacity now for the current antenna index set S2, that is,

$$P_2^{deleted} = \arg\max \log_2 \det\left(I_{N_R} + \frac{E_X}{QN_0} H_{S_2-\{p_2\}} H^H_{S_2-\{p_2\}} \right)$$

Again, the remaining antenna index set is updated to $S_3 = S_2 - \left\{ p_2^{deleted} \right\}$. This process will continue until all Q antennas are selected, that is, $|S_n| = Q$ .The complexity of selection method in descending order is higher than that in ascending order.

From the performance perspective, however, the selection method in descending order outperforms that in ascending order when 1 < Q <N_T. This is due to the fact that the selection method in descending order considers all correlations between the column vectors of the original channel gain before choosing the first antenna to delete.

When Q = 1, the selection method in descending order produces the same antenna index set as the optimal antenna selection method produces Equation (2). When Q = 1, however, the selection method in ascending order produces the same antenna index as the optimal antenna selection method in Equation (2) and achieves better performance than any other selection methods. In general, however, all these methods are just suboptimal, except for the above two special cases. Figure above shows the channel capacity with the selection method in descending order for various numbers of selected antennas with $N_T = 4$ and $N_R = 4$. [6]

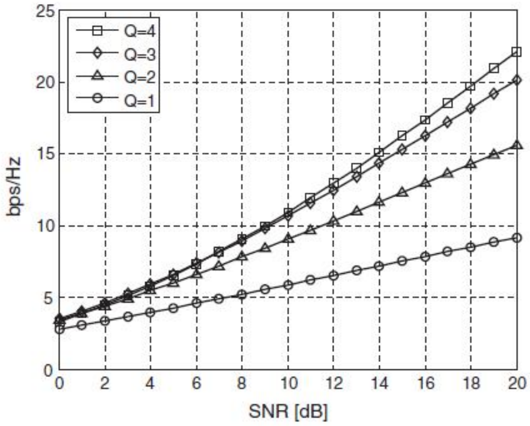


Figure 2: Channel capacities for antenna selection method in descending order.

4. CONCLUSION

In this paper, The Complexity-Reduced Antenna Selection Technique is discussed. As compared to the optimal antenna technique, complexity reduced antenna selection technique is better. Optimal antenna selection requires too much complexity depending on the total number of available transmit antennas. In order to reduce its complexity, a proposed a sub-optimal method is used. We adopted an approach in which additional antenna is selected in ascending order of increasing the channel capacity i.e., one antenna with the highest capacity is first selected we have used transmission techniques that can be used to exploit the CSI on the transmitter side. The CSI can be known completely or partially. Sometimes, only statistical information of the channel state is available. We have exploited such information for optimum antenna selection and hence for achieving the high channel capacity. Simulation results show that the channel capacity increases in proportion to the number of the selected antennas.

5. REFERENCES

[1] Menghui Yang, Tonghong Li, WeikangYang,Xin Su And Jing Wang (2009). A Channel Estimation Scheme for STBC-Based TDS-OFDM MIMO System. Eighth IEEE International Conference On Embedded Computing (2009). Page: 160-166.

[2] Dinesh B. Bhoyar, Dr. C. G. Dethe, Dr. M. M. Mushrif, Abhishek P. Narkhede (2013). Leaky Least Mean Square (Llms) Algorithm for Channel Estimation in BPSK-QPSK-PSK MIMO-OFDM.System. International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing, (2013). Page: 623.

[3] Jun Shikida, Satoshi Suyama, Hiroshi Suzuki, and Kazuhiko Fukawa (2010). Iterative Receiver Employing Multiuser Detection and Channel Estimation for MIMO-OFDMA. IEEE 71st Vehicular Technology Conference (2010). Page: 1-5.

[4] MeikD¨Orpinghaus, Adrian ISPAS and Heinrich Meyr (2011). Achievable Rate With Receivers Using Iterative Channel Estimation in Stationary Fading Channels. IEEE 8th International Symposium on Wireless Communication Systems, (2011). Page: 517-521.

[5] Bharath B. N. and Chandra R. Murthy. (2012) Channel Estimation At The Transmitter In a Reciprocal MIMO Spatial Multiplexing System. IEEE National Conference (2012), Page: 1-5.

[6] Reduced-Rank Estimation Of Non stationary Time-Variant Channels Using Subspace Selection. L. H. Xing, Zh. H. Yu, Zh. P. Gao, And L. Zha (2006)Channel Estimation For Transmitter Diversity OFDM Systems. 1st IEEE Conference on Digital Object Identifier (2006). Page: 1-4.

[7] JiaMeng, Wotao Yin, Yingying Li, Nam Tuan Nguyen, and Zhu Han *(2012)*. IEEE Journal Of Selected Topics In Signal Processing, 6(1) February 2012. Page: 15-25.

[8] Thomas Zemen, and Andreas F. Molisch, (2012) Adaptive IEEE Transactions, 61(9) (2012). Page: 4042-4056

[9] Osama Ullah Khan, Shao-Yuan Chen, David D. Wentzloff, And Wayne E. Stark (2012). Impact of Compressed Sensing With Quantization On UWB Receivers With Multipath Channel Estimation. IEEE Journal on Emerging and Selected Topics In Circuits And Systems, 2(3), September 2012. Page: 460-469.

[10] Mihai-AlinBadiu, Carles Navarro Manch´On, and Bernard Henri Fleury (2013). Message-Passing Receiver Architecture with Reduced-Complexity Channel Estimation. IEEE Communications Letters, 17(7) (2013). Page: 1404-1407.

[11] ErenEraslan, BabakDaneshrad, and Chung-Yu Lou (2013). Performance Indicator For MIMO MMSE Receivers In The Presence of Channel Estimation Error. IEEE Wireless Communications Letters, 2(2), April 2013. Page: 211-214.

[12] HarisGacanin (2013). Joint Iterative Channel Estimation and Guard Interval Selection of Adaptive Power line Communication Systems. IEEE 17th International Symposium On Power Line Communications And Its Applications (2013). Page: 197-202.

[13] Chao-Wei Huang, Tsung-Hui Chang, Xiangyun Zhou, and Y.-W. Peter Hong *(2013)*. Two-Way Training For Discriminatory Channel Estimation in Wireless MIMO Systems. IEEE Transactions On Signal Processing, 61(10), May 15, 2013. Page: 2724-2738.

[14] Mohamed Marey, Moataz Samir, And Mohamed Hossam Ahmed (2013). Joint Estimation Of Transmitter And Receiver IQ Imbalance With Ml Detection For Alamouti OFDM Systems. IEEE Transactions on Vehicular Technology, 62(6), July 2013. Page: 2847-2853.

[15] Joham, M., Utschick,W., and Nossek, J.A., "Linear transmit processing in MIMO communications systems" , IEEE Transactions on Signal Processing, vol 53(8), 2005.Page: 2700–2712.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 9, 554 - 558, 2014, ISSN: 2319–8656

Comparative Study of Diabetic Patient Data's Using Classification Algorithm in WEKA Tool

P.Yasodha
Pachiyappa's college for women,
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya
Kanchipuram, India

N.R. Ananthanarayanan
Pachiyappa's college for women,
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya
Kanchipuram, India

Abstract: Data mining refers to extracting knowledge from large amount of data. Real life data mining approaches are interesting because they often present a different set of problems for diabetic patient's data. The research area to solve various problems and classification is one of main problem in the field. The research describes algorithmic discussion of J48, J48 Graft, Random tree, REP, LAD. Here used to compare the performance of computing time, correctly classified instances, kappa statistics, MAE, RMSE, RAE, RRSE and to find the error rate measurement for different classifiers in weka .In this paper the data classification is diabetic patients data set is developed by collecting data from hospital repository consists of 1865 instances with different attributes. The instances in the dataset are two categories of blood tests, urine tests. Weka tool is used to classify the data is evaluated using 10 fold cross validation and the results are compared. When the performance of algorithms, we found J48 is better algorithm in most of the cases.

Keywords- Data Mining, Diabetics data, Classification algorithm, Weka tool

1. INTRODUCTION

The main focus of this paper is the classification of different types of datasets that can be performed to determine if a person is diabetic. The solution for this problem will also include the cost of the different types of datasets. For this reason, the goal of this paper is classifier in order to correctly classify the datasets, so that a doctor can safely and cost effectively select the best datasets for the diagnosis of the disease. The major motivation for this work is that diabetes affects a large number of the world population and it's a hard disease to diagnose. A diagnosis is a continuous process in which a doctor gathers information from a patient and other sources, like family and friends, and from physical datasets of the patient. The process of making a diagnosis begins with the identification of the patient's symptoms. The symptoms will be the basis of the hypothesis from which the doctor will start analyzing the patient. This is our main concern, to optimize the task of correctly selecting the set of medical tests that a patient must perform to have the best, the less expensive and time consuming diagnosis possible. A solution like this one, will not only assist doctors in making decisions, and make all this process more agile, it will also reduce health care costs and waiting times for the patients. This paper will focus on the analysis of data from a data set called Diabetes data set.

2. RELATED WORK

The few medical data mining applications as compared to other domains. [4] Reported their experience in trying to automatically acquire medical knowledge from clinical databases. They did some experiments on three medical databases and the rules induced are used to compare against a set of predefined clinical rules. Past research in dealing with this problem can be described with the following approaches: (a) Discover all rules first and then allow the user to query and retrieve those he/she is interested in. The representative approach is that of templates [3]. This approach lets the user to specify what rules he/she is interested as templates. The system then uses the templates to retrieve the rules that match the templates from the set of discovered rules. (b) Use constraints to constrain the mining process to generate only relevant rules. [12] Proposes an algorithm that can take item constraints specified by the user in the association rule mining processor that only those rules that satisfy the user specified item constraints are generated.

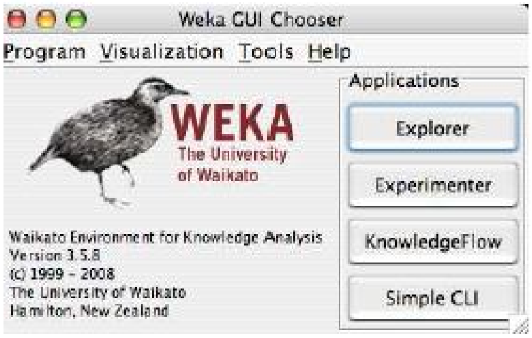
The study helps in predicting the state of diabetes i.e., whether it is in an initial stage or in an advanced stage based on the characteristic results and also helps in estimating the maximum number of women suffering from diabetes with specific characteristics. Thus patients can be given effective treatment by effectively diagnosing the characteristics.

Our research work based on the concept from Data Mining is the knowledge of finding out of data and producing it in a form that is easily understandable and comprehensible to humans in general. These further extended in this to make an easier use of the data's available with us in the field of Medicine.

The main use of this technique is the have a robust working model of this technology. The process of designing a model helps to identify the different blood groups with available Hospital Classification techniques for analysis of Blood group data sets. The ability to identify regular diabetic patients will enable to plan systematically for organizing in an effective manner. Development of data mining technologies to predict treatment errors in populations of patients represents a major advance in patient safety research.

3. MATERIALS AND METHODS

The **WEKA** (Waikato Environment for Knowledge Analysis) software was developed in the University of New Zealand. A number of data mining methods are implemented in the WEKA software. Some of them are based on decision trees like the J48 decision tree, some are rule-based like ZeroR and decision tables, and some of them are based on probability and regression, like the Naïve Bye's algorithm. The data that is used for WEKA should be made into the ARFF (Attribute Relation file format) format and the file should have the extension dot ARFF (.arff). WEKA is a collection of machine learning algorithms for solving real world data mining problems. It is written in Java; WEKA runs on almost any platform and is available on the web at www.cs.waikato.ac.nz/ml/weka.



Waikato Environment for Knowledge Analysis
Version 3.5.8
(c) 1999 - 2008
The University of Waikato
Hamilton, New Zealand

3.1. DATA PREPROCESSING

An important step in the data mining process is data preprocessing. One of the challenges that face the knowledge discovery process in medical database is poor data quality. For this reason we try to prepare our data carefully to obtain accurate and correct results. First we choose the most related attributes to our mining task.

3.2. DATA MINING STAGES

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the health datasets. The testing method adopted for this research was parentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Sixty six percent (66%) of the health dataset which were randomly selected was used to train the dataset using all the classifiers. The validation was carried out using ten folds of the training sets. The models were now applied to unseen or new dataset which was made up of thirty four percent (34%) of randomly selected records of the datasets. Thereafter interesting patterns representing knowledge were identified.

3.3 PATTERN EVALUATION

This is the stage where strictly interesting patterns representing knowledge are identified based on given metrics.

3.4 EVALUATION MATRICS

In selecting the appropriate algorithms and parameters that best model the diabetes forecasting variable, the following performance metrics were used:

3.4.1. **Time:** This is referred to as the time required to complete training or modeling of a dataset. It is represented in seconds

3.4.2. Kappa Statistic: A measure of the degree of nonrandom agreement between observers or measurements of the same categorical variable.

3.4.3. Mean Absolute Error: Mean absolute error is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error.

3.4.4. Mean Squared Error: Mean-squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value. The mean-squared error is simply the square root of the mean-squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values.

3.4.5. Root relative squared error: Relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value. As with the root mean-squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicted value.

3.4.6. Relative Absolute Error: Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values.

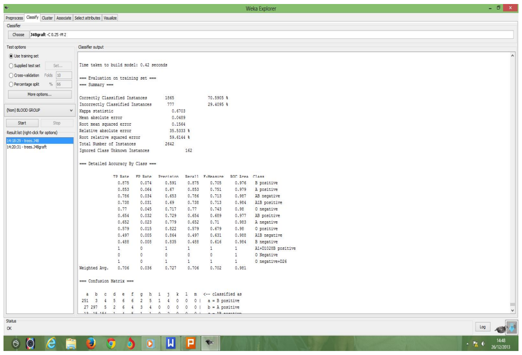
4. METHODOLOGY

4.1. CLASSIFICATION

Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy". Popular classification techniques include decision trees and neural networks.

4.2. J48 Pruned Tree

J48 is a module for generating a pruned or unpruned C4.5 decision tree. When we applied J48 onto refreshed data, we got the results shown as below on Figure .



J48graft -C 0.25 -M 2

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) BLOOD GROUP

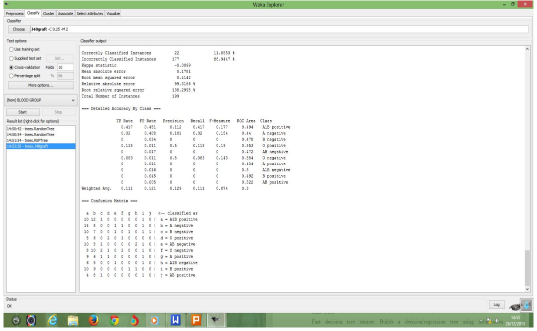
14:18:29 - trees.J48
14:20:31 - trees.J48graft

```
Time taken to build model: 0.42 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1865               70.5905 %
Incorrectly Classified Instances       777               29.4095 %
Kappa statistic                          0.6703
Mean absolute error                      0.0489
Root mean squared error                  0.1564
Relative absolute error                 35.5333 %
Root relative squared error             59.6144 %
Total Number of Instances             2642
Ignored Class Unknown Instances                162

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.875     0.074      0.591     0.875     0.705      0.976    B positive
                 0.853     0.064      0.67      0.853     0.751      0.979    A positive
                 0.786     0.034      0.653     0.786     0.713      0.987    AB negative
                 0.738     0.031      0.69      0.738     0.713      0.984    A1B positive
                 0.77      0.045      0.717     0.77      0.743      0.98     O negative
                 0.654     0.032      0.729     0.654     0.689      0.977    AB positive
                 0.652     0.023      0.779     0.652     0.71       0.983    A negative
                 0.579     0.015      0.822     0.579     0.679      0.98     O positive
                 0.497     0.005      0.864     0.497     0.631      0.988    A1B negative
                 0.488     0.008      0.835     0.488     0.616      0.984    B negative
                 1         0          1         1         1          1        A1+D1028B positive
                 0         0          0         0         0          1        O Negative
                 1         0          1         1         1          1        O negative+D26
Weighted Avg.    0.706     0.036      0.727     0.706     0.702      0.981

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   <-- classified as
 251   3   4   5   6   6   2   5   1   4   0   0   0 |   a = B positive
  27 297   5   2   6   4   3   4   0   0   0   0   0 |   b = A positive
```

Fig- 1: J48 Tree

4.3. J48 graft

Perhaps C4.5 algorithm which was developed by Quinlan [13] is the most popular tree classifier till today. Weka classifier package has its own version of C4.5 known as J48 or J48graft



Classifier: J48graft -C 0.25 -M 2

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) BLOOD GROUP

Result list (right-click for options)
14:50:42 - trees.RandomTree
14:50:54 - trees.RandomTree
14:51:54 - trees.REPTree
14:53:20 - trees.J48graft

Classifier output

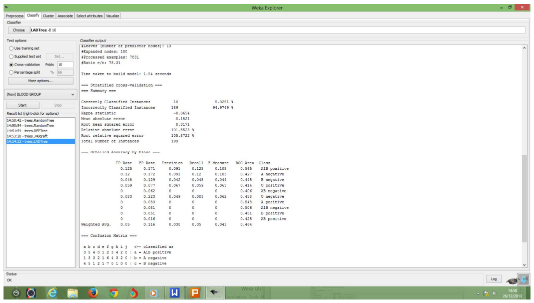
```
Correctly Classified Instances          22               11.0553 %
Incorrectly Classified Instances       177               88.9447 %
Kappa statistic                         -0.0099
Mean absolute error                      0.1781
Root mean squared error                  0.4142
Relative absolute error                 99.3166 %
Root relative squared error            138.2995 %
Total Number of Instances              199

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.417     0.451      0.112     0.417     0.177      0.494    A1B positive
                 0.32      0.408      0.101     0.32      0.154      0.44     A negative
                 0         0.034      0         0         0          0.478    B negative
                 0.118     0.011      0.5       0.118     0.19       0.553    O positive
                 0         0.017      0         0         0          0.472    AB negative
                 0.083     0.011      0.5       0.083     0.143      0.584    O negative
                 0         0.011      0         0         0          0.404    A positive
                 0         0.016      0         0         0          0.5      A1B negative
                 0         0.045      0         0         0          0.492    B positive
                 0         0.005      0         0         0          0.522    AB positive
Weighted Avg.    0.111     0.121      0.129     0.111     0.074      0.5

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j   <-- classified as
 10 12  1  0  0  0  0  0  1  0 |  a = A1B positive
 14  8  0  0  1  1  0  0  1  0 |  b = A negative
 10  7  0  0  1  0  1  0  1  1 |  c = B negative
  8  6  0  2  0  1  0  0  0  0 |  d = O positive
 10  8  1  0  0  0  0  2  1  0 |  e = AB negative
  8 10  2  1  0  2  0  0  1  0 |  f = O negative
  9  6  1  1  0  0  0  0  1  0 |  g = A positive
  6  5  0  0  1  0  0  0  1  0 |  h = A1B negative
 10  9  0  0  0  0  1  1  0  0 |  i = B positive
  4  8  1  0  0  0  0  0  1  0 |  j = AB positive
```

Status
OK

Fig-2: J48 Graft

4.4. LAD tree

LADTree is a class for generating a multiclass alternating decision tree using logistics strategy. LADTree produces a multi- class LADTree. It has the capability to have more than two class inputs. It performs additive logistic regression using the Logistics Strategy.



Classifier: LADTree -B 10

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) BLOOD GROUP

Result list (right-click for options)
14:50:42 - trees.RandomTree
14:50:54 - trees.RandomTree
14:51:54 - trees.REPTree
14:53:20 - trees.J48graft
14:54:22 - trees.LADTree

Classifier output

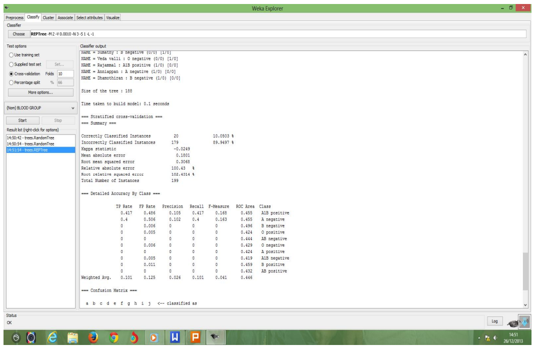
```
#Leaves (number of predictor nodes): 13
#Expanded nodes: 100
#Processed examples: 7831
#Ratio e/n: 78.31

Time taken to build model: 1.54 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          10                5.0251 %
Incorrectly Classified Instances       189               94.9749 %
Kappa statistic                         -0.0654
Mean absolute error                      0.1821
Root mean squared error                  0.3171
Relative absolute error                101.5523 %
Root relative squared error            105.8722 %
Total Number of Instances              199

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.125     0.171      0.091     0.125     0.105      0.565    A1B positive
                 0.12      0.172      0.091     0.12      0.103      0.427    A negative
                 0.048     0.129      0.042     0.048     0.044      0.445    B negative
                 0.059     0.077      0.067     0.059     0.063      0.414    O positive
                 0         0.062      0         0         0          0.406    AB negative
                 0.083     0.223      0.049     0.083     0.062      0.458    O negative
                 0         0.083      0         0         0          0.548    A positive
                 0         0.081      0         0         0          0.506    A1B negative
                 0         0.051      0         0         0          0.451    B positive
                 0         0.016      0         0         0          0.425    AB positive
Weighted Avg.    0.05      0.116      0.038     0.05      0.043      0.464

=== Confusion Matrix ===

 a b c d e f g h i j   <-- classified as
 3 5 4 0 1 2 3 4 2 0 | a = A1B positive
 1 3 3 2 1 6 4 3 2 0 | b = A negative
 4 5 1 2 1 7 0 1 0 0 | c = B negative
```

Status
OK

Fig-3: LAD Tree

4.5. REP Tree

Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).



Classifier: REPTree -M 2 -V 0.0010 -N 3 -S 1 -L -1

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) BLOOD GROUP

Result list (right-click for options)
14:50:42 - trees.RandomTree
14:50:54 - trees.RandomTree
14:51:54 - trees.REPTree

Classifier output

```
NAME = Sumathy : B negative (0/0) [1/0]
NAME = Veda valli : O negative (0/0) [1/0]
NAME = Rajammal : A1B positive (1/0) [0/0]
NAME = Anniappan : A negative (1/0) [0/0]
NAME = Dhamothiran : B negative (1/0) [0/0]

Size of the tree : 188

Time taken to build model: 0.1 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          20               10.0503 %
Incorrectly Classified Instances       179               89.9497 %
Kappa statistic                         -0.0249
Mean absolute error                      0.1801
Root mean squared error                  0.3068
Relative absolute error                100.43   %
Root relative squared error            102.4314 %
Total Number of Instances              199

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.417     0.486      0.105     0.417     0.168      0.455    A1B positive
                 0.4       0.506      0.102     0.4       0.163      0.455    A negative
                 0         0.006      0         0         0          0.496    B negative
                 0         0.005      0         0         0          0.424    O positive
                 0         0          0         0         0          0.444    AB negative
                 0         0.006      0         0         0          0.429    O negative
                 0         0          0         0         0          0.424    A positive
                 0         0.005      0         0         0          0.419    A1B negative
                 0         0.011      0         0         0          0.459    B positive
                 0         0          0         0         0          0.432    AB positive
Weighted Avg.    0.101     0.125      0.026     0.101     0.041      0.446

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j   <-- classified as
```

Status
OK

5. RESULT AND DISCUSSION

J48 algorithm was selected for the prediction because out of the five classifiers used to train the data, it had the best performance measures.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     py1
Instances:    2804
Attributes:   11
              NAME
              GENDER
              AGE
              HEIGHT
              BLOOD GROUP
              BLOOD SUGAR(F)
              BLOOD SUGAR (PP)
              BLOOD SUGAR (R)
              URINE SUGAR(F)
              URINE SUGAR(PP)
              URINE SUGAR (R)
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------
J48 pruned tree
------------------
AGE <= 46
|   AGE <= 35
|   |   GENDER = Male
|   |   |   AGE <= 26: B positive (2.0/1.0)
|   |   |   AGE > 26: A positive (3.0/1.0)
|   |   GENDER = Female
|   |   |   AGE <= 34: O negative (2.0)
|   |   |   AGE > 34: A positive (2.0/1.0)
|   AGE > 35: B positive (7.0/4.0)
AGE > 46
|   GENDER = Male
|   |   AGE <= 60: O positive (5.0/3.0)
|   |   AGE > 60: AB positive (4.0/2.0)
|   GENDER = Female
|   |   AGE <= 63
|   |   |   AGE <= 55: AB positive (2.0/1.0)
|   |   |   AGE > 55: A1B positive (4.0/2.0)
|   |   AGE > 63: A negative (2.0/1.0)
Number of Leaves  :         10
Size of the tree :      19

Time taken to build model: 0.29 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances                          1865
70.5905%
Incorrectly Classified Instances                         777
29.4095%

Kappa statistic                     0.6703
Mean absolute error                    0.0489
Root mean squared error                   0.1564
Relative absolute error              35.5333 %
```

```
Root relative squared error          59.6144%
Total Number of Instances            2642
Ignored Class Unknown Instances              162
```

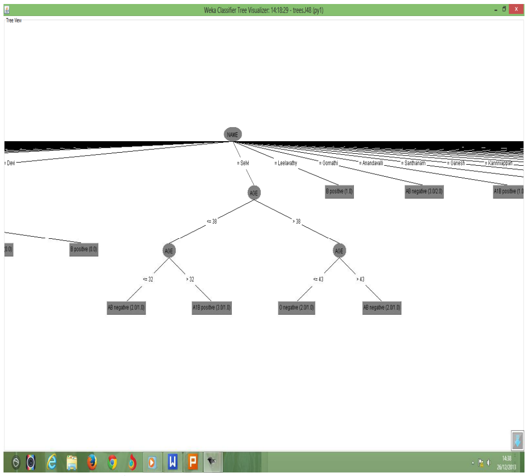


Fig -6: **VISUALISE THE TREE**

CLASSIFIER	CORRECTLY CLASSIFIED INSTANCES	TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE	ROC AREA
J48	1865 (70.5%)	0.706	0.036	0.727	0.706	0.702	0.981
J48 GRAFT	1524 (57.6%)	0.607	0.024	0.678	0.520	0.600	0.781
LAD TREE	553 (20.9%)	0.05	0.116	0.038	0.05	0.043	0.464
RANDOM TREE	350 (13.2%)	0.111	0.122	0.098	0.111	0.07	0.464
REP TREE	348 (0.13%)	0.132	0.132	0.017	0.132	0,031	0.5

Table-1: DIFFERENT PERFORMANCE METRICES RUNNING IN WEKA

In this study, we examine the performance of different classification methods that could generate accuracy and some error to diagnosis the data set. According to above Table 1 , we can clearly see the highest accuracy is 70.5% belongs to J48 and lowest accuracy is 0.13% that belongs to REP. The total time required to build the model is also a crucial parameter in comparing the classification algorithm.

	J48	**J48GRAFT**	**RANDOM TREE**	**REP**	**LAD**
TIME	0.29	0.42	0.02	0.05	1.85
CORRECTLY CLASSIFIED INSTANCES	1865 (70.5%)	1524 (57.6%)	350 (13.2%)	348 (0.13 %)	553 (20.9%)
KAPPA STATISTIC	0.011	0.6700	0.011	0.012	0.0654
MAE	0.0123	0.0480	0.1798	0.1377	0.1821
RMSE	0.1154	0.1560	0.3199	0.2624	0.3171
RAE%	12.53%	35.50%	100.24 %	99.98 %	101.55 %
RRSE%	22.61%	58.63%	106.82 %	100%	105.87 %

Table- 2: ERRORS MEASUREMENT FOR DIFFERENT CLASSIFIERS IN WEKA

Based on above table, we can compare errors among different classifiers in WEKA. We clearly find out that J48 is the best, second best is the j48 graft ,LAD, REP & random. An algorithm which has a lower error rate will be preferred as it has more powerful classification capability and ability in terms of medical and bio informatics fields.

6. CONCLUSION AND FUTURE WORK

The objective of this study is to evaluate and investigate FIVE selected classification algorithms based on WEKA. The best algorithm in WEKA is J48 classifier with an accuracy of 70.59% that takes 0.29 seconds for training. They are used in various healthcare units all over the world. In future to improve the performance of these classification.

I had been use the data mining classifiers to generate decision tree format. In this paper WEKA software for my experiment. Identify the diabetic patient's behavior using the classification algorithms of data mining. The analysis had been carried out using a standard blood group data set and using the J48 decision tree algorithm implemented in WEKA. The research work is used to classify the diabetic patient's based on the gender, age, height & weight, blood group, blood sugar(F), blood sugar(PP), urine sugar(F), urine sugar(PP). The J48 derived model along with the extended definition for identifying regular patients provided a good classification accuracy based model.

The distribution of blood groups in both positive and negative are shown in Table-1. Overall blood group A was the commonest (24.03 %), followed by B (18.77%), AB (19.11%), O (23.65) and A1B (17.14%).

Blood group spectrum	**Nos (%)**	**+ve (%)**	**–ve (%)**
A	635 (24.03)	348 13.17	287 10.85
B	496 (18.77)	289 (10.93)	207 (7.83)
AB	505 (19.11)	196 (7.41)	309 (11.69)
A1B	453 (17.14)	300 (11.35)	153 (5.79)
O	625 (23.65)	345 (10.59)	280 (13.05)

Table-3: Spectrum of Blood groups +ve and -ve in major population. (n-2642)

In the present blood group-A was the predominant (24.03%) while A1B was the least common (17.14%). Blood group "A" was the most predominant (24.03%) in both positive and negative subjects, followed by blood group A, B,O,A1B and AB.

The future work will be focused on using the other classification algorithms of data mining. It is a known fact that the performance of an algorithm is dependent on the domain and the type of the data set. Hence, the usage of other classification algorithms like machine learning will be explored in future.

The future work can be applied to blood groups to identify the relationship that exits between diabetic, diagnosing cancer patients based on blood cells or predicting the cancer types on the blood groups, blood pressure, personality traits and medical diseases.

7. REFERENCES

[1] Mats Jontell, Oral medicine, Sahlgrenska Academy, Göteborg University (1998) "A Computerised Teaching Aid in Oral Medicine and Oral Pathology. " Olof Torgersson, department of Computing Science, Chalmers University of Technology, Göteborg.

[2] T. Mitchell, "Decision Tree Learning", in T. Mitchell, Machine Learning (1997) the McGraw- Hill Companies, Inc., pp. 52-78.

[3] Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I (1994) "Finding interesting rules from large sets of discovered association rules," CIKM.

[4] Tsumoto S., (1997)"Automated Discovery of Plausible Rules Based on Rough Sets and Rough Inclusion," Proceedings of the Third Pacific-Asia Conference (PAKDD), Beijing, China, pp 210-219.

[5] Liu B., Hsu W., (1996) "Post-analysis of learned rules," AAAI, pp. 828-834.

[6] Liu B., Hsu W., and Chen S., (1997) "Using general impressions to analyze discovered classification rules," Proceedings of the Third ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[7] Stutz J., P. Cheeseman. (1996) Bayesian classification (autoclass): Theory and results. In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press

[8] Witten Ian H., E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Ch. 8, © 2000 Morgan Kaufmann Publishers

[9] http://www.cs.waikato.ac.nz/ml/weka/, accessed 06/05/21.

[10] http://grb.mnsu.edu/grbts/doc/manual/ J48_Decision_T rees.html, accessed

[11] Wikipedia, ID3-algorithm (accessed 2007/12/09) (URL: http://en.wikipedia.org/wiki/ID3_algorithm)

[12] Srikant,R.,Vu,Q.andAgrawal,R.,(1997), "Mining association rules with item constraints," Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, USA, pp 67-73.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 9, 559 - 563, 2014, ISSN: 2319–8656

A Novel Document Image Binarization For Optical Character Recognition

Varada V M Abhinay
S.V. College of Engineering
Tirupati, Andhra Pradesh, India

P.Suresh Babu
S.V. College of Engineering
Tirupati, Andhra Pradesh, India

Abstract: This paper presents a technique for document image binarization that segments the foreground text accurately from poorly degraded document images. The proposed technique is based on the Segmentation of text from poorly degraded document images and it is a very demanding job due to the high variation between the background and the foreground of the document. This paper proposes a novel document image binarization technique that segments the texts by using adaptive image contrast. It is a combination of the local image contrast and the local image gradient that is efficient to overcome variations in text and background caused by different types degradation effects. In the proposed technique, first an adaptive contrast map is constructed for a degraded input document image. The contrast map is then binarized by global thresholding and pooled with Canny's edge map detection to identify the text stroke edge pixels. By applying Segmentation the text is further segmented by a local thresholding method that. The proposed method is simple, strong, and requires minimum parameter tuning.

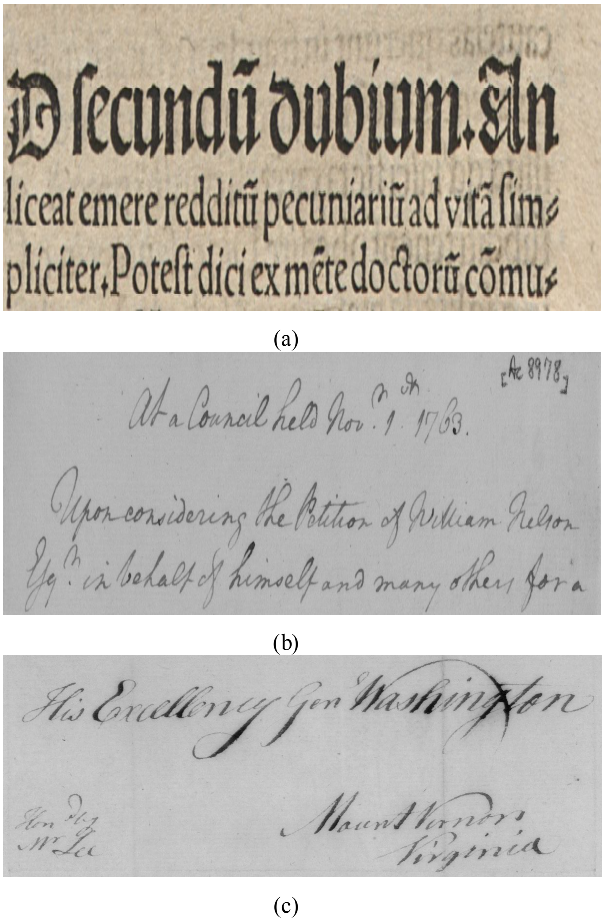
Keywords:Adaptive image contrast, document analysis, pixel intensity, pixel classification.

1. INTRODUCTION

Document image binarization is a preprocessing stage for various document analyses. As more and more number of text document images is scanned, speedy and truthful document image binarization is becoming increasingly important. As document image binarization [1] has been studied for last many years but the thresholding techniques of degraded document images is still an unsettled problem. This can be explained by the difficulty in modeling different types of document degradation such as change in image contrast, uneven illumination, smear and bleeding-through that exist in many document images as illustrated in Fig. 1.

The printed text within the degraded documents often shows a certain amount of variation in terms of the stroke brightness, stroke connection, stroke width and document image background. A large number of document image thresholding techniques have been reported in the literature. For document images of a good quality, global thresholding is efficiently capable to extract the document text.

But for document images suffering from different types of document degradation, adaptive thresholding, which estimates a local threshold for each document image pixel, is usually capable of producing much better binarization results. One of the typical adaptive thresholding approach [2] is window based, which estimates the local threshold based on image pixels within a neighborhood window. However, the performance of the window-based methods depends heavily on the window size that cannot be determined properly without prior knowledge of the text strokes.



D ſecundū dubium.ſ.an
liceat emere redditū pecuniariū ad vitā ſim-
pliciter. Poteſt dici ex mēte doctorū cōmu-

(a)

At a Council held Nov.r 1 1763. [Ac 8978]
Upon considering the Petition of William Nelson
Esq.r in behalf of himself and many others for a

(b)

His Excellency Gen Washington
Hon by Mr Lee
Mount Vernon, Virginia

(c)

Figure 1: Degraded Document Images from DIBCO Datasets.

Whereas, some window-based method Nib lack's often introduces a large amount of noise and some method such as Sauvola's[3] is very sensitive to the variation of the image contrast between the document text and the document background.

The proposed method is simple, straightforward and able to handle different types of degraded document images with minimum parameter tuning. It use of the adaptive image contrast that mixes the local image contrast and the local image gradient adaptively and therefore is liberal to the text and background variation caused by different types of degradations of document images. In particular, the proposed technique addresses the over-normalization problem of the local maximum minimum algorithm. At the same time, the parameters used in the algorithm can be adaptive estimated.

2. RELATED WORK

Many degraded documents do not have a clear bimodal pattern; global thresholding is usually not a suitable approach for the degraded document binarization. Adaptive thresholding [2], which estimates a local threshold for each document image pixel, is again a better approach to deal with different types variations in degraded document images. The early window-based adaptive thresholding [2] techniques estimate the local threshold by using the mean and the standard variation of image pixels within a local neighborhood window.

The weakness of these window-based thresholding techniques is that the thresholding performance depends deeply on the window size and hence the character stroke width. The other different approaches have also been reported, including background subtraction, texture analysis[4], recursive method [5], decomposition method, contour completion, Markov Random Field [3], cross section sequence graph analysis. These methods combine different types of image information and domain knowledge and are often complex. These methods are very useful features for segmentation of text from the document image background because the document text usually has certain image contrast to the neighboring document background. They are very effective and have been used in many document image binarization techniques.

3. PROPOSED METHOD

This section describes the proposed document image binarization techniques

A. Contrast Image Construction.

B. Canny Edge Detector.

C. Local Threshold Estimator.

D. Post Processing Procedure.

In the proposed technique, first an adaptive contrast map is constructed for an input image degraded badly. Then the binarized contrast map is combined with edge map obtained from canny edge detector to identify the pixels in edges of text stroke. By using local threshold the foreground text is further segmented which is based on the intensities of detected text stroke edge pixels within a local window. The block diagram of proposed method is as shown in figure 2.

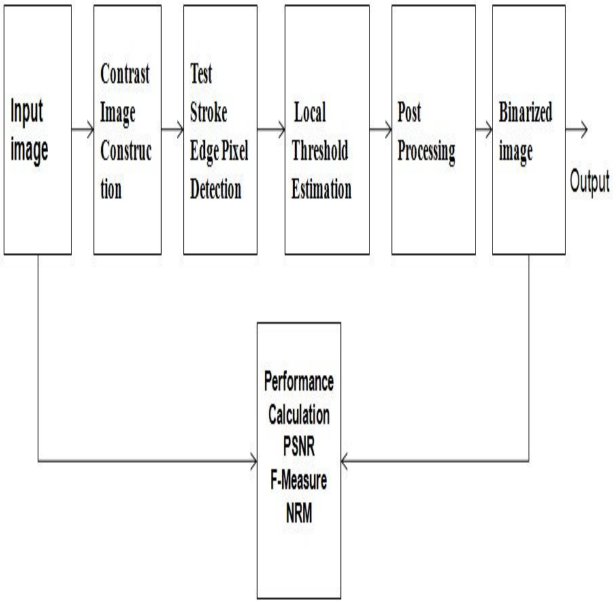

Input image
Contrast Image Construc tion
Test Stroke Edge Pixel Detection
Local Threshold Estimation
Post Processing
Binarized image
Output
Performance Calculation
PSNR
F-Measure
NRM

Figure 2: Block diagram of the proposed method

3.1 Contrast Image Construction

The image gradient has been extensively used for edge detection from uniform background image. Degraded document may have certain variation in input image because of patchy lighting, noise, or old age documents, bleed-through, etc. In Bernsen's paper, the local contrast is defined as follows:

$$C(i,j) = Imax(i,j) - Imin(i,j) \qquad (1)$$

where C(i, j) denotes the contrast of an image pixel (i, j), *I*max*(i, j)* and *I*min*(i,)* denote the maximum and minimum intensities within a local neighborhood windows of (i, j), respectively.

If the local contrast C(i, j) is smaller than a threshold, the pixel is set as background directly. Otherwise it will be classified into text or background by comparing with the mean of *I*max*(i, j)* and *I*min*(i, j)* in Bernsen's method. The earlier proposed a novel document image binarization method [1] by using the local image contrast that is evaluated as follows

$$C(i,j) = \frac{Imax(i,j) - Imin(i,j)}{Imax(i,j) + Imin(i,j) + €} \qquad (2)$$

Where € is a positive but infinitely small number that is added in case the local maximum is equal to 0. By comparing with Bernsen's contrast in Equation ,and the local image contrast in Equation 2 introduces a normalization factor by extracting the stroke edges properly; the image gradient can be normalized to recompense the image variation within the document background. To restrain the background variation the local image contrast is evaluated as described in Equation 2.

In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image difference that is similar to the traditional image gradient. The denominator is a normalization factor that suppresses the image variation within the document background. For pixels within bright regions of a image, it will produce a large normalization factor to neutralize the numerator and accordingly result in a relatively low image contrast. For the pixels within dark regions of an image, it will produce a small denominator and accordingly result in a relatively high image contrast.

3.2 Canny's Edge Detection

Through the contrast image construction the stroke edge pixels are detected of the document text. The edges can be detected through canny edge detection algorithm, firstly by smoothing the noise from the image and then algorithm finds for the higher magnitude of image accordingly the edges of image gradient will be marked. While marking only local edges of image should be marked.

As these methods are evaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both sides of the text stroke will be selected as the high contrast pixels. The binary map can be improved further through the combination with the edges by Canny's edge detector, through the canny edge detection the text will be identified from input image.

3.3 Local Threshold Estimation

Once the text stroke edges are detected, then the document text can be extracted based on the observation that the document text is surrounded by text stroke edges and also has a lower intensity level compared with the detected stroke edge pixels[2]. The document text is extracted based on the detected text Stroke edges as follows:

$$R(x, y) = \begin{cases} 1 & I(x, y) \leq E_{\text{mean}} + \frac{E_{\text{std}}}{2} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where *E*mean and *E*std are the mean and standard deviation of the intensity of the detected text stroke edge pixels within a neighborhood window *W*.

3.4 Post-Processing Procedure

Document image thresholding often introduces a certain amount of error that can be corrected through a series of post-processing operations. Document thresholding error can be corrected by three post-processing operations based on the estimated document background surface and some document domain knowledge. In particular, first remove text components (labeled through connected component analysis) of a very small size that often result from image noise such as salt and pepper noise. The real text components are usually composed based on the observation that of much more than 3 pixels, the text components that contain no more than 3 pixels in our system is simply removed.

Next, remove the falsely detected text components that have a relatively large size. The falsely detected text components of a relatively large size are identified based on the observation that they are usually much brighter than the surrounding real text strokes. Then observations are then captured by the image difference between the labeled text component and the corresponding patch within the estimated document background surface.

4. APPLICATION

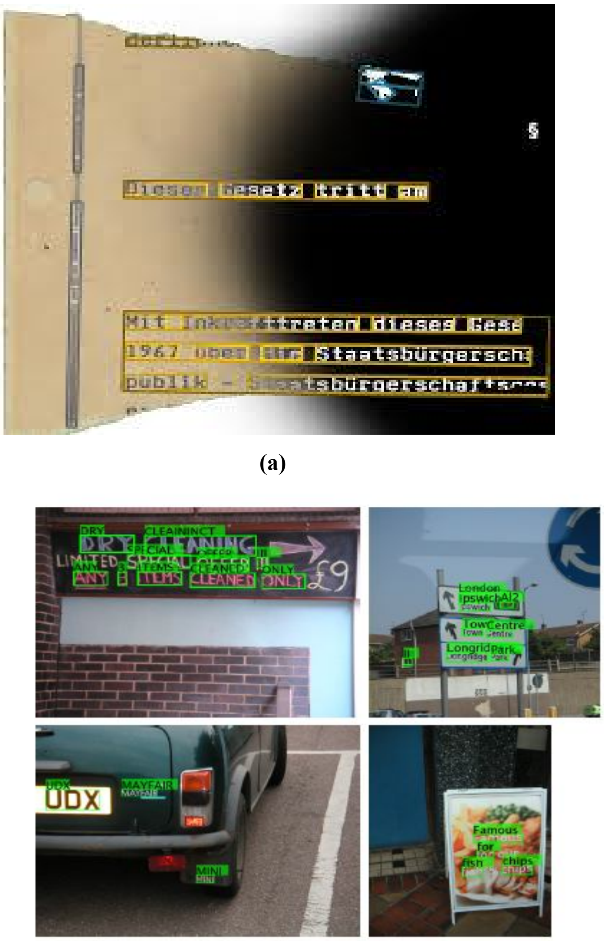
Foreign language data acquired via Arabic OCR is of vital interest to military and border control applications. Various hardcopy paper types and machine- and environment-based treatments introduce artifacts in scanned images. Artifacts such as speckles, lines, faded glyphs, dark areas, shading, etc. complicate OCR and can significantly reduce the accuracy of language acquisition. For example, Sakhr Automatic Reader, a leader in Arabic OCR, performed poorly in initial tests with noisy document images. We hypothesized that performing image enhancement of bi-tonal images prior to Arabic OCR would increase the accuracy of OCR output. We also believed that increased accuracy in the OCR would directly correlate to the success of downstream machine translation.

We applied a wide variety of paper types and manual treatments to hardcopy Arabic documents. The intent was to artificially model how documents degrade in the real world. Four hardcopies of each document were created by systematically applying four levels of treatments. Subsequent scanning resulted in images that reflect the progressive damage in the life-cycle of each document – the Manually Degraded Arabic Document (MDAD) corpus. Applying the assigned image enhancement settings, three types of images were captured for each document:

- Without image enhancement,
- With Fujitsu TWAIN32 image enhancement, and
- With both Fujitsu TWAIN32 and ScanFix image enhancement.

The MDAD corpus default scans already established the images without image enhancement. The dynamic threshold capability (i.e., SDTC) was disabled in order to gain full control of the scan brightness. Discovering the ideal brightness setting involved re-scanning and reducing the brightness setting repeatedly until white pixels appeared inside glyphs. The last scan with solid black glyphs was selected as the optimal scan. The three types of images for each document were then processed through the OCR tool. CP1256 files were output and compared against the ground truth using the UMD accuracy tool.

We discovered that the evaluation metrics may not be reflecting the OCR output well. We have already mentioned that the OCR tool expects clean documents and on noisy documents it attempts to recognize speckles as characters. For noisy documents, the OCR tool produced several failure characters in the output file or caused Automatic Reader to abnormally end. Since accuracy was calculated as the number of correct characters minus error characters, divided by the number of correct characters, the tool produced negative and zero values.



(a)

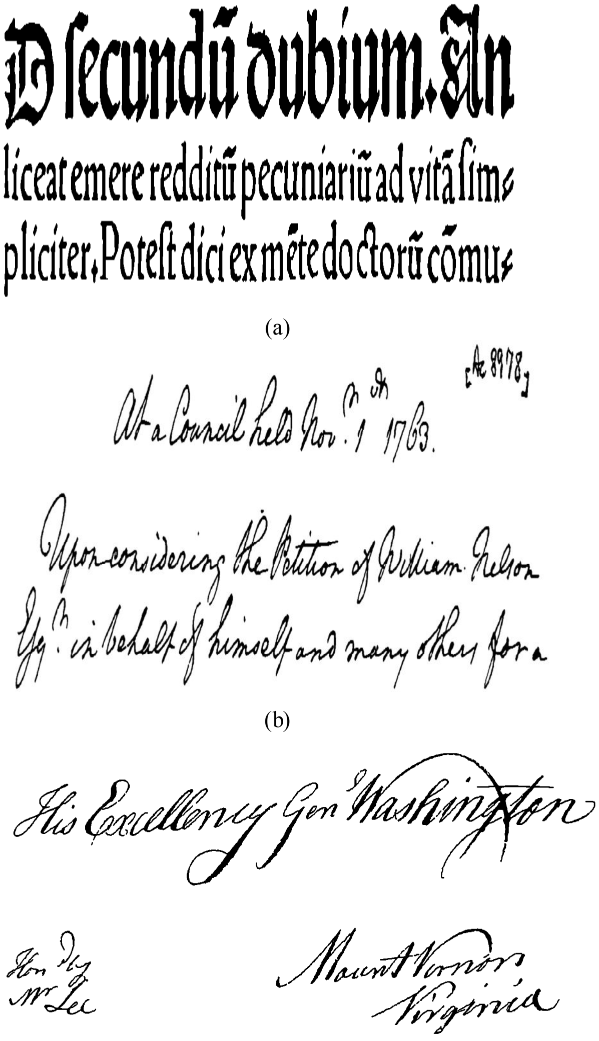
(b)

Figure 3. Text localization and recognition results of proposed binarization method.

5. DISCUSSION

As described in previous sections, the proposed method involves several parameters, most of which can be automatically estimated based on the statistics of the input document image. This makes our proposed technique more stable and easy-to-use for document images with different kinds of degradation. Binarization results of the sample document images are as shown in figure 4.

The superior performance of our proposed method can be explained by several factors. First, the proposed method combines the local image contrast and the local image gradient that help to suppress the background variation and avoid the over-normalization of document images with less variation. Second, the combination with edge map helps to produce a precise text stroke edge map. Third, the proposed method makes use of the edges of the text stroke that help to extract the foreground text from the document background accurately.



D ſecundū dubium. An liceat emere redditū pecuniariū ad vitā ſim-
pliciter. Poteſt dici ex mēte doctorū cōmu-

(a)

[Ac 8978]

At a Council held Nov. 1st 1763.

Upon considering the Petition of William Nelson Esq. in behalf of himself and many others for a

(b)

His Excellency Gen Washington

Hon. by Mr Lee

Mount Vernon Virginia

(c)

Figure 4: Binarization results of the sample document images as shown in figure 1.

6. CONCLUSION

The proposed method follows numerous different steps, Firstly pre-processing procedure collect the document image information, then proposed technique makes use of the local image contrast that is valuated based on the local maximum and minimum. Through canny edge detection the stroke edges are detected based on the local image variation, then local threshold is estimated based on the detected stroke edge pixels within a local neighborhood window and then through post processing procedure the quality of binarized result is improved.

7. REFERENCES

[1] Bolan Su, Shijian Lu, and Chew Lim Tan, ―Robust Document Image Binarization Technique for Degraded Document Images‖ IEEE TRANS ON IMAGE PROCESSING, VOL. 22, NO. 4, APRIL 2013.

[2] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," Pattern Recognit., vol. 39, no. 3, pp. 317–327, 2006.

[3] T. Lelore and F. Bouchara, "Document image binarisation using Markov field model," in Proc. Int. Conf. Doc. Anal. Recognit., pp. 551–555, Jul. 2009.

[4] Y. Liu and S. Srihari, "Document image binarization based on texture features," IEEE Trans. Pattern Anal. Mach. In tell., vol. 19, no. May 1997.

[5] M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding technique for image segmentation," in Proc. IEEE Trans. Image Process., June 1998.

Agent based Personalized e-Catalog Service System

M.Thangaraj,
Department of Computer Science
Madurai Kamaraj University ,Madurai ,
Tamilnadu

M Chamundeeswari,
Department of Computer science
V.V.V College for Women
(Affiliated to Madurai Kamaraj university)
Virudhunagar, Tamil Nadu

Abstract: With the emergence of the e-Catalog, there has been an increasingly wide application of commodities query in distributed environment in the field of e-commerce. But e-Catalog is often autonomous and heterogeneous, effectively integrating and querying them is a delicate and time-consuming task. Electronic catalog contains rich semantics associated with products, and serves as a challenging domain for ontology application. Ontology is concerned with the nature and relations of being. It can play a crucial role in e-commerce as a formalization of e-Catalog. User personalized catalog ontology aims at capturing the users' interests in a working domain, which forms the basis of providing personalized e-Catalog services. This paper describes a prototype of an ontology-based Information retrieval agent. User personalized catalog ontology aims at capturing the users' interests in a working domain, which forms the basis of providing personalized e-Catalog services. In this paper, we present an ontological model of e-Catalogs, and design an Agent based personalized e-Catalog service system (ABPECSS), which achieves match user personalized catalog ontology and domain e-Catalog ontology based on ontology integrated

Keywords: personalization, semantic web, information retrieval, ontology, re-ranking algorithms, knowledge base ,user profile,e-catalog

1. INTRODUCTION

As Internet technologies develop rapidly, companies are shifting their business activities to e-Business on the Internet. Worldwide competition among corporations accelerates the reorganization of corporate sections and partner groups, resulting in a break of the conventional steady business relationships. For instance, a marketplace would lower the barriers of industries and business categories, and then connect their enterprise systems. Electronic catalogs contain the data of parts and products information used in the heavy electric machinery industry. They contain not only the commercial specifications for parts (manufacturer name, price, etc.), but also the technical specifications (physical size, performance, quality, etc.). Clearly defined product information is a necessary foundation for collaborative business processes. Furthermore, semantically enriched product information may enhance the quality and effectiveness of business transactions. As a multifunctional applied system, it serves for advertisement, marketing, selling and client support, and at the same time it is a retail channel.

As the number of Internet users and the number of accessible Web pages grow, it is becoming more and more difficult for users to find documents among e-Catalogs that are relevant to their particular needs. Users can search with a search engine which allows users to enter keywords to retrieve e-Catalogs that contain these keywords. The navigation policy and search have their own problems. Indeed, approximately one half of all retrieved documents have been reported to be irrelevant. The main reasons for obtaining poor search results are that (1) many words have multiple meanings (2) key words are not enough to express the rich concepts and the natural semantics of customers' queries. (3) The property query lacks of semantic support, and is difficult to search for knowledge, and has other problems of mechanisms. (4) Related merchandises cannot be returned. What is needed is a solution that will *personalize* the e-Catalog selection and be presented to each user. A semantically rich user model and an efficient way of processing semantics are the keys to provide personalized e-Catalog services. In view of the existing limitations, we develop a personalized ontology based on user model, called user personalized catalog ontology, which has the same level of semantics as domain ontology.

The rest of this paper is structured as follows: Section 2 describes related work. Section 3 , explains the theory of propose system. Section 4 we put forward our modeling methodology for generating user personalized catalog and product domain ontology. Then in Section 5, we present the implementation of the system and its evaluation. Conclusion and future work are drawn in Section 6

2. RELATED WORK

E-catalogues play a critical role in e-procurement marketplaces. They can be used in both the tendering (pre-award) and the purchasing (post-award) processes. Companies use e-catalogues to exchange product information with business partner’s .Suppliers use e-catalogues to describe goods or services that they offer for sale. Mean while buyers may use e-catalogues to specify the items that they want to buy [1, 2] Matching a product request from a buyer with products e-catalogs that have been provided by the suppliers, helps companies to reduce the efforts needed to find partners in e-marketplaces [5, 7]

.

2.1 E-Catalog Ontology Design

Researches in recent years show that applying ontology to e-commerce scenarios would bring benefits such as solving the interoperability problems between different e-commerce systems [3, 4]. Especially, e-Catalog, which is a key component of e-commerce systems, seems to be the most adequate domain within e-commerce scenarios where ontology can realize the expression of e-Catalog on a semantic level. It is possible for e-business systems to offer diverse interoperable services by sharing well-defined e-Catalog model containing rich semantics. Fensel [5] described in principle how ontology's can support the integration of heterogeneous and distributed information in ecommerce scenarios which is mainly based on product catalogs, and what tasks are needed to be mastered. E-Catalog ontology model is defined as ECO (concepts, relationship, properties, axioms and individuals).

The traditional key-based retrieval method cannot satisfy massive heterogeneous personalized catalog service, then [8] introduce meta search engines, but this method is passive service. [9] Provided an intelligent catalog recommend method using customer requirements mapping with product categories. [10] Brought forward personalized e-Catalog model based on customer interests and [11] is a personalized catalog service community, WebCatalog [12] designed enterprise e-Catalog based on customer behavior. The knowledge representation and acquisition of client catalog turns into the key problems. In order to reach an effective method, K-clustering algorithm and e-Catalog segmentation approach are described in [13] and [14] described the customer segmentation method based on brand and product, price. In [15] the author researched personalized catalog service with one-to-one market by association rules and CART. In recent years, personalized ontology's (also known as private ontology, such as [9] are introduced into e-Catalog service, Peter Haase put forward personalized ontology learning theory based on user access and interest coordination [16]. In distributed system, there are sharing concepts of domain ontology's and personalized knowledge ontology's [17]. Therefore, it has important theoretical and practical significance to apply personalized ontology's to personalized e-Catalog service.

3. PROPOSED ARCHITECTURE

The personalized information retrieval system based on multi-agent adopts the working fashion of multi-agent cooperation, multi-agent collaborate mutually and communicate to one another for accomplishing task.

The system consists of User Agent, Query Generation Agent, Reasoning and Expanding Agent, Searching Agent and Filtering Agent, Personalized Ranking Agent and Knowledge Base. It is shown in Figure 1.[23] All agents are monitored entirely to fulfill proprietary system functions, including information retrieval and Knowledge Base update.

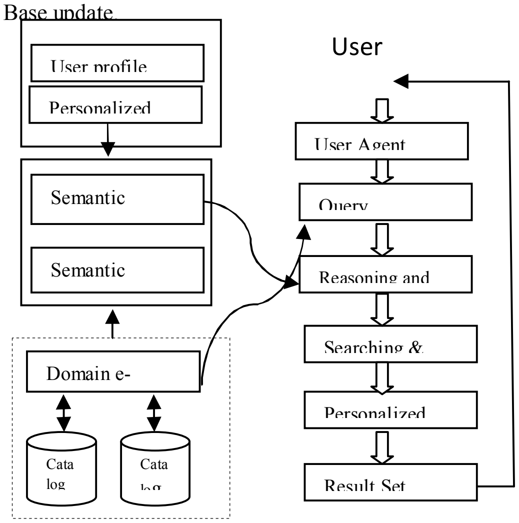


Figure 1 Architecture of Agent based personalized e-Catalog service system (ABPESS)

(1) User Agent: User Agent is the mutual interface between user and system, and provides a friendly platform to users. User Agent also takes over result from Personalized ranking agent and presents personally these results to user. User's browsing or evaluating behavior can be stored and learned by User Agent, so user interest model may be updated and improved in time.

(2). Query Generating Agent: QGA incepts user's retrieval request, which is transformed to prescriptive format, and transmits the formatted user request to Reasoning and expanding agent.

(3) Reasoning and expanding agent: In the personalized information retrieval system, Reasoning and expanding agent takes charge of receiving formatted user request from QGA, and the user request is expanded according to user interest model. Afterwards, the perfected user request is transmitted to Searching & Filtering Agent.

(4) Searching Agent and Filtering Agent: Searching Agent collects all data from initiative Searching Agent or meta-Searching Agent, takes out invalid links, deleting excrescent information, and finally processed data are transmitted to Personalized Re-ranking agent . Filtering Agent analyses the returned data from Searching Agent, filtrating useless information, and processed results are send to Personalized Re-ranking agent.It also completes search result statistic, user browse statistic, and retrieval keywords statistic, etc. Various statistic outcomes are stored in Knowledge Base.

Algorithm of e-Catalog- searching and filtering:

Constructing semantic results SR, where DO is domain ontology, expanding ontologies, SRD is r.
Keyset KS= { $k_1,k_2..k_n$ }
Input: keyword, basic ontology DO;
Output: semantic results SR;
Search(KS,DO)

```
Begin

for(each KS) {finding  DO mapping $K_i$ , according to the semantic mapping table; }

for(sub-ontology s in DO){
if( $R^{w}_{d}$ (Oi,s)≥m && s isn't in DO)
find the result s for semantic query
copy the components of s to SR;}
return SR;
End
```

(4) Personalized Re-ranking agent : it is the decision-making center of personalized information retrieval system based on multi-agent, and assorts with data communication and task assignment. Personalized Re-ranking agent use re-ranking alg. To find the new score based on user interest.

```
PR (uid)
Begin
If uid exits{        Re-ranking(CP,uid,interest)}
else
{
```

```
For each user entered
{
  userProfiledb()->uid,uinterest ,keyword weight
 For each search
  {
    Usersearchdb()->  uid,keyword,interest
    Apply Assoicationarlg(uid,keyword,interest)
   Cp()<-keyword,interest }}}
```

(5) Knowledge Base**:** This is an auxiliary component, used by the integration mechanism. It contains semantically-enhanced inter-domain and intra-domain knowledge bases representing dependencies and relationships between various user, item and context features. The data stored in the knowledge bases facilitate resolving the heterogeneities in the obtained user modeling data. For example, it allows reconciliation of the ontology's exploited by various recommender systems, converting the terms used by certain systems to a standard representation, and even provides machine translation tools resolving cross-lingual dependencies.

(6) Semantic ontology: It contains some product knowledge used to generate the queries. It was designed as a hierarchical tree, with a frame based representation approach. This ontology must be at some degree context free, but it has to point elements of the search engines used by the Query Generation module.

4. METHODOLOGY:

4.1 Method of Designing User Catalog ontology:

In order to satisfy customer's personalized requirement, we should master more information of the customers. Sometimes customers also cannot describe their own thought, to understand their potential mind, we need user e-Catalog ontology. Based on consumer behavior, we propose a personalized approach to build personalized catalog ontology (PCO).

PCO supposed to be formed by

- First, build user personal ontology (PCO) based on users' personal information and preferences
- Second, extract user catalog information from user purchase history, user searching keywords, user browsing catalog, user feedback information
- Third, web resource according to user catalog ontology information

Agent based e-catalog organizes a group of keywords expressing users' interest through PCO, when users puts semantic query, it is no longer a simple keywords match, but considering users' personal preference and information, and tightly integrates the users and products, so that the system can improve the semantic query precision rate and recall rate, as well as be conducive to sort query results.

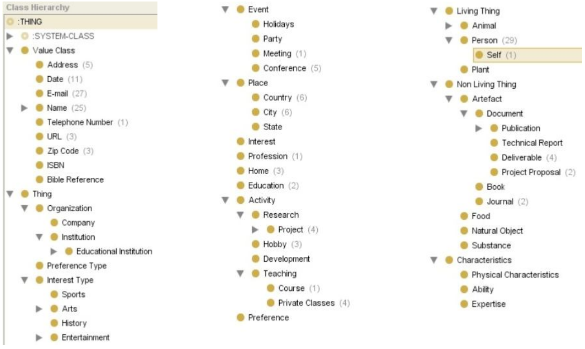


Figure 2 framework of user Personalized Catalog Ontology

Figure 2 shows a user catalog ontology framework, in which we describe user interest information, user preference and product concepts, properties and individuals that users are interested in, including product area, brand and quality authentication. Users associate with the product by property hasPreference, and we set aside a weight interface in property "has Preference", indicating the fact users' different observation extent about different properties of a product which is shown in Figure 3.

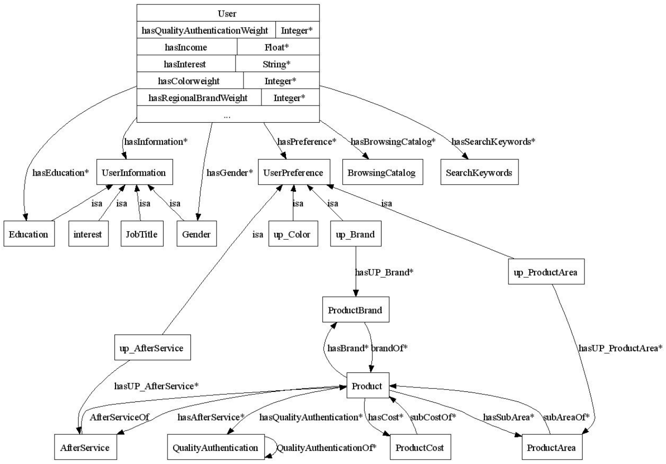


Figure 3 The Relationship of user Personalized Catalog Ontology

Generating Semantic Catalog ontology (SCO):

Generation domain e-Catalog ontology is divided into three steps:

- Extraction of the core concepts and properties for domain e-Catalog ontology's, according to the UNSPSC standards, wordNet standards and semantic catalog dictionary.
- Construction of a SCO model..
- Acquisition standardized DECO by e-Catalog ontology pruning subsystem, combining WorldNet and semantic catalog dictionary.

4.2 Semantic Match Based on Ontology

One critical step of semantic match is that calculation semantic match degree between the terms of ontology concepts. There have been many methods to calculate conceptual semantic match in e-commerce scenarios [18]. Common calculation methods and models are: (1) Identifier-based method [19], which uses word-building to find the semantic match degree between the concepts, and primarily reflects the linguistic similarity of the two concepts; (2) Synonym dictionary-based method [20], which organizes all concepts to a tree hierarchy structure according to synonym dictionary where there is only one path between any two nodes and this path length is taken as a measure of semantic distance of the two concepts; (3) Feature Match-based model [21], which calculates semantic match of concepts by the collection of properties; and (4) Semantic relationship-based model [22], also known as the semantic distance-based model, which calculates semantic match of concepts based on hierarchy information and is mainly used in the same ontology. In this paper, we need to calculate the semantic match of UPCO and DECO using Individual-based Semantic Match methods.

4.3. Individual-based Semantic Match

To query user preferences product, we should get the product similar with user preferences, namely calculating the instance similarity between SCO individual and PCO individual. We calculate the semantic match of the individuals by the property value-based method.

calculate the semantic match method based on linguistics, when we calculate semantic match degree of the property values

$$smi(C_1,C_2)=\frac{ed(C_1,C_2)}{(|C_1|+|C_2|)/2}$$

Explanation:

| C_1 | is the length of the string $C_{1,}$| C_2 | the length of the string C_2, $ed(A,C_2)$ is the same number of characters in C_1 and C_2.

String C_1 and C_2 are input parameters, in the process, which are the properties values of two products calculate the individual semantic match of the two products through comparing several groups property semantic match degree.

4.4 Basic function of ABPECSS:

To implement agent based E-service first of all, personalized user catalog ontology's are customized according to consumers(PCO) ; secondly, we need to build domain e-Catalog ontology's(SCO) ; thirdly, we match the two kinds of ontology's by match algorithm through semantic reasoning and expanding agent which generates match result sets.

The basic the theory of distributed semantic query based on e-Catalog ontology is: users input key words, phrases, sentences or paragraphs (users' queries, U_q) in user querying interface; query generator module translates U_q to ontology descript; query reasoning and expanding module is responsible for reasoning and expanding the descript using the semantic match result set is, then outputs semantic queries (S_q) in forms of Sparql and finally extract data from distributed e-Catalog database. Searching and Filtering module combines the distributed results and filters repetitive and invalid results .personalized ranking agent rearrange the result sets and recommended to the user.

4.5 Results Personalization

The personalization helps in getting relevant results for the user's query. As shown in the query-processing steps, the personalization starts with the query enrichment step, where we utilize the user profile to expand the query and to fill in the incomplete query templates. Here, we go into more detail with the results personalization steps and show how we capture the user's feedback.

Results personalization steps

Personalizing the results involves presenting the results in the most effective way possible through several steps. The first step is answering the user's query in the same language he asks it in, regardless of the language of the ontology and the knowledge base, which has the annotated data. The second step is answering the user's query in appropriate syntax based on the question type; a confirmation question is different than a subjective question, as the user expects a "yes" or "no" answer in the first type, while s/he expects a list of items in the second type. So, an answer is personalized to express the understanding of the query and to be familiar to the user. The third step is ranking the results based on the user's preferences and interests. Finally, it filters the non-relevant food or health information based on the user profile.

4.6 User's feedback

Continuous feedback collection is required to sharpen the user's experiences. Feedback is not only explicit, but also implicit, as it can be collected through different measures. Many measures could help in reflecting the implicit feedback, such as time spent in browsing the results, clicks on the data sources, clicks on the result facets related to the search results, etc. All interactions and feedback are recorded and logged in the usage log which is analyzed after each query to know how effective the results are and how we can improve the future recommendations. This is reflected in the user profile ontology

5. IMPLEMENTATION AND EXPERIMENTATION

In this section experiments carried out to evaluate the performance of proposed system will be discussed from a quantitative point of view by running some experiments to evaluate the precision of the results. The basic idea of the experiment is to compare the search result from keyword based search engine with proposed one on the same category and the same keywords.

The proposed system ABPECSS is implemented in C#.Net as Web-based system using Visual Studio 2008, .NET Framework 3.5, and SQL Server 2005. The system was evaluated by having 20 users implement the system to create personal ontology's. The user was given a query interface to input his/her query parameters and view each one of their concepts and every concept from the SCO that had been matched to the personalize catalog concept. Also the user was able to decide which concept or property was not needed when reasoned and expanded the query. In the experiment, we take different electronic items as an example. The user was asked to compare the semantic query result and that from the keyword-based search engines and decide if ABPECSS was the better. Therefore, we manually create the domain e-Catalog ontology (SCO) and user personalized catalog ontology (PCO) and calculate semantic match degree in the system.

Table 1 Experimental results statistics for query manipulation

concepts	**Total found concept s**	**concept Found correct**	**Correct concepts manually**	**Precision**	**Recall**
Dell Inspiron 15R i3531-1200BK	**89**	**71**	**74**	**91.36%**	**80.43%**
Dell Alienware 18 Gaming Laptop.	**89**	**71**	**93**	**71 %**	**5.54%**
Canon EOS 6D Black SLR Digital.	**50**	**16**	**56**	**78.00%**	**84.21%**
Nikon D810 DSLR Camera (Body Only)	**90**	**53**	**78**	**90.00%**	**81.54%**
Nikon 1 AW1 14.2MP Waterproof.	**50**	**10**	**13`**	**89.00%**	**86.92%**
Bargains Depot USB Cable Lead Cord	**45**	**19**	**39**	**93.00%**	**8.95%**

We evaluated the system with two measures, precision and relevance, shown in Figure 4 Precision measures the number of relevant pages that were seen vs. the total number of pages that were seen. Relevance measures the number of relevant pages seen plus the number irrelevant pages not seen vs. the total number queried

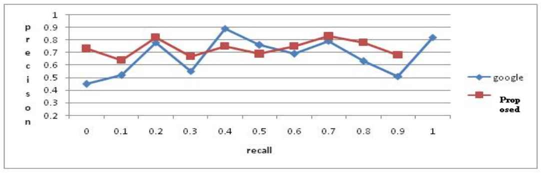


Figure 4 Precision Vs Recall graph for proposed system Vs GOOGLE

The next experiment aims at determining the importance of personalization by using generated dynamic user model during using the system. The user model is used to re-rank the retrieved documents to match the user interest

Personalization time:

Time to retrieve any information depends on the type of search engine, size of data set, relevancy between query and doc. User history & re-ranking algorithm used.

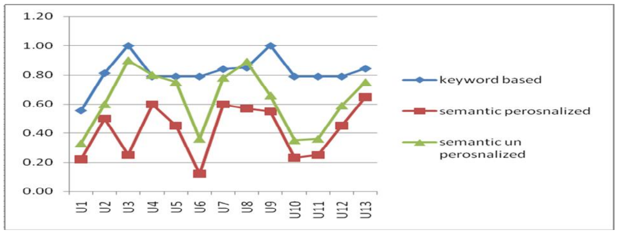


Figure 5 Performance efficiency of the new system

Figure 5 discuss the performance efficiency of the system when the system uses to retrieve the result.

It is observed that 80% users, out of 30 users in our data set, have found improved precision with the proposed approach in comparison to the standard search engine(Google) results, while 34% users have achieved equal precision with both approaches. It has been observed that users who posed Queries in unpopular context than well liked context got better performance. In addition, when the system can extract the exact context of user's need, the Precision and recall is found better than other search engine results.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose a framework for semantic query manipulation and personalization of Electronic catalog service systems. We present the user profile ontology and its relation to other domain ontology's. Then, we explain the semantic query processing steps and present the result personalization steps. A complete scenario is illustrated to visualize the framework followed by experimental results. The empirical evaluation shows promising improvements in the relevancy of the retrieved results and of the user's satisfaction. It can be used in other domain by editing the domain ontology using export option of new system and building the domain concepts weight table .In future work, we will focus on: (1) automatically learn e-Catalog ontological concepts, properties and relationship from web to build PCO; (2) add business properties besides general properties to SCO; (3) construct the Reasoning and Expending Module of ABPECSS, to set rules onto SCO.

7. REFERENCES

[1]. J. de Bruijn, D. Fensel, and M. Kerrigan, Modeling Semantic Web Services, Heidelberg: Springer-Verlag,2008, pp. 30-52.

[2] . E. Casasola, ProFusion personal assistant: An agent for personalized Information filtering on the WWW, M.S. thesis, The University of Kansas, Kansas, KCK, U.S.A., 1998.

[3]. I. Chen, J. Ho, and C. Yang, On hierarchical web catalog integration with conceptual relationships in
thesaurus, in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and
Development in Information Retrieval, Washington, 2006, pp. 635-636.

[4]. O. Corcho, A. Gómez-Pérez, Solving integration problems of e-Commerce standards and initiatives through ontological mappings, in Proceedings of the 17th International Joint Conference on Artificial Intelligence,
Seattle, 2001.

[5]. R. Cyganiak, A relational algebra for SPARQL. Digital Media Systems Laboratory HP Laboratories Bristol.

HPL-2005-170, September 28, 2005.

[6]. Z. Cui, D. Jones, and P. O'Brien, Semantic B2B Integration: Issues in Ontology-based Approaches, SIGMOD Record, vol. 31, no. 11, 2002.

[7]. S. Gauch, J. Chaffee, and A. Pretschner, Ontology-based personalized search and browsing, Web Intelligence and Agent Systems, vol. 1, no. 3-4, pp. 219-234, 2003.

[8]. L. Kwon and C. O. Kim, Recommendation of e-commerce sites by matching category-based buyer query and product e-Catalogs, Computers in Industry, vol. 59, no. 4, pp. 380-394, 2008.

[9] J. Lee and T. Lee, Massive catalog index based search for e-Catalog matching, in Proceedings of the 9thIEEE International Conference on e-Commerce Technology. Tokyo. IEEE Computer Society, 2007, pp. 341-348

[10]. H. Lee, J. Shim, S. Lee, and S. Lee, Modeling considerations for product ontology, in Lecture Notes in Computer Science, Advances in Conceptual Modeling: Theory and Practice, vol. 4231, Tucson, AZ: Springer, 2006, pp. 291-300.

[11]. J. Leukel, V. Schmitz, and F. Dorloff, A modeling approach for product classification systems, in Proceedings of 13th International Conference on the Database and Expert Systems Applications. Aix-en-Provence, 2002, pp. 868-874.

[12]. H. Li, XML and industrial standards for electronic commerce, Knowledge and Information Systems, vol. 2,no. 4, pp. 487-497, 2000.

[13]. S. Liao, C. Chen, C. Hsieh, and S. Hsiao, Mining information users' knowledge for one-to-one marketing on information appliance, Expert Systems with Applications, vol. 36, no. 3, pp. 4967-4979, 2009.

[14]. L. Lim and M. Wang, Managing e-Commerce catalogs in a DBMS with native XML support, in Proceedings of the IEEE International Conference on e-Business Engineering, Beijing, 2005, pp. 564-571.

[15]. C. Lin and C. Hong, Using customer knowledge in designing electronic catalog, Expert Systems with Applications, vol. 34, no. 1, pp. 119-127, 2008.

[16]. D. Liu, Y. Lin, and C. Chen, Deployment of personalized e-Catalogues: An agent-based framework integrated with XML metadata and user models, Journal of Network and Computer Applications, vol. 24, no. 3, pp. 201-228, 2001.

[17]. K. Masanobu, D. Kobayashi, D. Xiaoyong, and I. Naohiro, Evaluating word similarity in a semantic network,Informatics, 2000, vol. 24, no. 1, pp. 192-202.

[18]. H. Paik and B. Benatallah, Personalised organisation of dynamic e–Catalogs, in Web Services, e-Business,and the Semantic Web (C. Bussler, R. Hull, S. McIlraith, M. E. Orlowska, B. Pernici and J. Yang, Eds.).Heidelberg, Berlin: Springer Verlag, 2002, pp. 139-152.

[19] E. Prud'hommeaux and A. Seaborne. (2005, July) SPARQL Query Language for RDF. W3C Working Draft. [Online]. Available: http://www.w3.org/TR/2005/WD-rdf-sparql-query-20050721/.

[20]. R. Rada, H. Mili, E. Bicknell, and M. Blettner, Development and application of a metric on semantic nets, IEEE Transaction on System, Man and Cybernetics, vol. 19, no. 1, pp. 17-30, 1989.

[21]. H. Sun-Young and K. Eun-Gyung, A study on the improvement of query processing performance of OWL data based on Jena, in Proceedings of the International Conference on Convergence and Hybrid Information Technology, Daejeon, 2008, pp. 678-681.

[22]. A. Tversky, Feature of similarity, Psychological Review, vol. 84, no. 4, pp. 327-352, 1977.

[23] Dr.M.Thangaraj and Mrs. M.Chamundeeswari Agent Based personalized Semantic Web Information Retrieval System in (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 8, 2014

A Posteriori Perusal of Mobile Computing

Yusuf Perwej
Department of Computer Science & Engg., Al Baha University, Al Baha, Kingdom of Saudi Arabia (KSA)

Shaikh Abdul Hannan
Department of Computer Science & Engg.,
Al Baha University,
Al Baha,
Kingdom of Saudi Arabia (KSA)

Firoj Parwej
Department of Computer Science & Engg.,
Singhania University,
Pacheri Bari,
Distt. Jhunjhunu,
Rajasthan, India

Nikhat Akhtar
Department of Computer Science & Engg.,
Integral University,
Lucknow, India

Abstract: The breakthrough in wireless networking has prompted a new concept of computing, called mobile computing in which users tote portable devices have access to a shared infrastructure, independent of their physical location. Mobile computing is becoming increasingly vital due to the increase in the number of portable computers and the aspiration to have continuous network connectivity to the Internet irrespective of the physical location of the node. Mobile computing systems are computing systems that may be readily moved physically and whose computing ability may be used while they are being moved. Mobile computing has rapidly become a vital new example in today's real world of networked computing systems. It includes software, hardware and mobile communication. Ranging from wireless laptops to cellular phones and WiFi/Bluetooth- enabled PDA's to wireless sensor networks; mobile computing has become ubiquitous in its influence on our quotidian lives. In this paper various types of mobile devices are talking and they are inquiring into in details and existing operation systems that are most famed for mentioned devices are talking. Another aim of this paper is to point out some of the characteristics, applications, limitations, and issues of mobile computing.

Keywords: Mobile Computing, Mobile Devices, Mobile Computing Security, Cache Management, Mobile Operating Systems, Mobile Limitations.

1. INTRODUCTION

Mobile computing refers to technologies that employ small portable devices and wireless communication networks that allow user mobility by providing access to data anytime, anywhere. Mobile computing systems are computing systems that may be easily moved physically and whose computing capabilities may be used while they are being moved. Examples are laptops, [1] personal digital assistants (PDAs), and mobile phones. Mobile computing technology improves healthcare in a number of ways, such as by providing healthcare professionals access to reference information and electronic medical records and improving communication among them. Mobile computing is associated with the mobility of hardware, data and software in computer applications. Respectively, mobile software deals with the requirements of mobile applications. Also, hardware includes the components and devices which are needed for mobility. Communication issues include ad-hoc and infrastructure networks, protocols, communication properties, data encryption and concrete technologies. Mobile computing means being able to use a computing device while changes location properties. The study of this new area of computing has prompted the need to rethink carefully about the way in [2] which mobile network and systems are conceived. Mobile phones are one of the most ubiquitously used devices around. With different brands like the Android, Windows Mobile, and the iPhone, mobile phones have revolutionized the way we look at computing. There are thousands of applications such as social networking and games that have cropped up on mobile phones. With the help of cloud services, even sophisticated applications such as multi-player games, image processing, and speech processing has become feasible.

2. A HISTORY OF MOBILE COMPUTING

Mobile computing is the discipline for creating an information management platform, which is free from spatial and temporal constraints. The freedom from these constraints allows its users to access and process desired information from anywhere in the space. In the figure 1shows a timeline of mobile computing development. One of the very first computing machines, [3] the abacus, which was used as far back as 500 B.C., was, in effect, a mobile computing system because of its small size and portability. As technology progressed, the abacus evolved into the modern calculator. A mobile computing system, as with any other type of computing system, can be connected to a network. Connectivity to the network, however, is not a prerequisite for being a mobile computing system. The late 1960s, networking allows computers to talk to each other. Networking two or more computers together requires some medium that allows the signals to be exchanged among them. This was typically achieved through wired networks. By the 1970s, communication satellites began to be commercialized. With the new communication satellites, the quality of service and reliability improved enormously. Still, satellites are expensive to build, launch, and maintain. So the available bandwidth provided by a series of satellites was limited. In the 1980s, cellular telephony technologies became commercially viable and the exciting world of mobile computing is only in existence since the 1990s. Since then, the devices have been developed for mobile computing has taken over the wireless industry. This new type of communication is a very powerful tool for business and private purposes. Mobile computing is defined as the ability to use technology that is not physically connected to the static network [4]. He really used for a radio transmitter on a stable, most often with the help of a large antenna. Mobile computing has evolved from a two-way radio that use large antennas to communicate a simple message, to three inches of personal computers that can do almost everything a normal computer does. Today, most laptops and personal digital assistants all have wireless cards or Bluetooth interface built them for convenient mobile Internet access. Mobile solutions are right under your nose all day, and connectivity has never been easier.

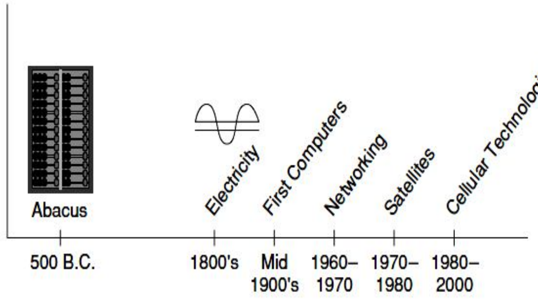


Figure 1. A Timeline of Mobile Computing

3. THE CHARACTERISTICS OF MOBILE COMPUTING

Mobile computing is accomplished using a combination of computer hardware, system and applications software and some form of communications medium. Mobile hardware includes mobile devices or device components that receive or access the service of mobility. They would range from Portable laptops, Smart phones, Tablet Pc's, Personal Digital Assistants. These devices will have receptor medium that is capable of sending and receiving signals. These devices are configured to operate in full-duplex, whereby they are capable of sending and receiving signals at the same time. They don't have to wait until one device has finished communicating for the [2] other device to initiate communications. The characteristics of mobile computing hardware are defined by the size and form factor, weight, microprocessor, primary storage, secondary storage, screen size and type, means of input, means of output, battery life, communications capabilities, expandability and durability of the device. Mobile computers make use of a wide variety of system and application software. The most common system software and operating environments used in mobile computers includes MSDOS, Symbian, Windows 3.1/3.11/95/98/NT, UNIX, android, a specialized OS like Blackberry shows in figure 2.

Figure 2. The Symbol of Most Common Operating Environments

Mobile software is the actual program that runs on the mobile hardware. It deals with the characteristics and requirements of mobile applications. This is the engine of that mobile device. In other terms, it is the operating system of that appliance. It's the [5] essential component that makes the mobile device operate. Since portability is the main factor, this type of computing ensures that users are not tied or pinned to a single physical location, but are able to operate from anywhere. It will incorporate all aspects of wireless communications. Finally, the most useful software - end user application like messaging, sales force automation, public query, data collection, etc.

The last few years have witnessed a phenomenal growth in the wireless industry, both in terms of mobile technology and its subscribers. A mobile radio communication system by definition consists of telecommunication infrastructure serving users that are on the move (i.e., mobile). The communication between the users and the infrastructure is done over a wireless medium known as a radio channel. Telecommunication systems have [6] several physical components such as: user terminal/equipment, transmission and switching/routing equipment, etc. There has been a clear shift from fixed to mobile cellular telephony, especially since the turn of the century. By the end of 2010, there were over four times more mobile cellular subscriptions than fixed telephone lines. Both the mobile network operators and vendors have felt the importance of efficient networks with equally efficient design.

Many more designing scenarios have developed with not only 2G networks, but also with the evolution of 2G to 2.5G or even to 3G networks. Along with this, interoperability of the networks has to be considered. 1G refers to analog cellular technologies; it became available [7] in the 1980s. 2G denotes initial digital systems, introducing services such as short messaging and lower speed data. CDMA2000 1xRTT and GSM are the primary 2G technologies, although CDMA2000 1xRTT is sometimes called a 3G technology because it meets the 144 kbps mobile throughput requirement. EDGE, however, also meets this requirement. 2G technologies became [8] available in the 1990s. 3G requirements were specified by the ITU as part of the International Mobile Telephone 2000 (IMT-2000) project, in which digital networks had to provide 144 kbps of throughput at mobile speeds, 384 kbps at pedestrian speeds, and 2 Mbps in indoor environments. UMTS-HSPA and CDMA2000 EV-DO are the primary 3G technologies, although recently WiMAX was also designated as an official 3G technology. 3G technologies began to be deployed last decade. The ITU [9] has recently issued requirements for IMT-Advanced, which constitutes the official definition of 4G. Requirements include operation in up-to-40 MHz radio channels and extremely high spectral efficiency. The ITU recommends operation in upto-100 MHz radio channels and peak spectral efficiency of 15 bps/Hz, resulting in a theoretical throughput rate of 1.5Gbps. The Fourth generation (4G) will provide access [10] to a wide range of telecommunication services, including advanced mobile services, supported by mobile and fixed networks, which are increasingly packet based, along with a support for low to high mobility applications and a wide range of data rates, in accordance with service demands in multi-user environment. There are many communications technologies available today that enable mobile computers to communicate.

4. MOBILE COMPUTING DEVICES

Mobile computing is not limited to, Mobile Phones only, but also there are various gadgets available in the market helping mobile computing. Example for personal digital assistant/enterprise digital assistant, smart phone, tablet computer, ultra-mobile PC, and wearable computer. They are usually classified in the following categories.

4.1 Personal Digital Assistant (PDA)

The main purpose of this device was to act as an electronic organizer or day planner that is portable, easy to use and capable of sharing information with you with a computer system. The PDA was an extension of the PC, not a replacement. These systems were capable of sharing information with a computer system through a process or service known as synchronization. Where both devices will access each other to check for changes or updates in the individual devices. The use of infrared and Bluetooth [11] connections enabled these devices to always be synchronized. With

PDA devices, a user could; browsers the internet, listen to audio clips, watch video clips, edit and modify office documents, and many more services. They had a stylus and a touch sensitive screen for input and output purposes.

Figure 3. Personal Digital Assistant (PDA)

4.2 Smart Phones

This kind of phone combines the features of a PDA with that of a mobile phone or camera phone. It has a superior edge over other kinds of mobile phones. The smart phone has the capability to run multiple programs concurrently. These phones include high-resolution touch enabled screens, web browsers that can access and properly display standard web pages rather than just mobile-optimized sites, and high-speed data access via Wi-Fi and high speed cellular broadband. The most common [12] mobile operating systems (OS) used by modern Smart phones include Google's Android, Apple's iOS, Nokia's Symbian, RIM's Blackberry OS, Samsung's Bada, Microsoft's Windows Phone, and embedded Linux distributions such as Maemo and MeeGo. Such operating systems can be installed on many different phone models, and typically each device can receive multiple OS software updates over its lifetime.

Figure 4. Smart Phone

4.3 Tablet PC and I-Pads

This mobile device is larger than a mobile phone or a personal Digital Assistant and integrates into a touch screen and operated using touch sensitive motions on the screen. They are often controlled by a pen or touch of a finger. They are usually in slate form and are light in weight. Examples would include; Ipads, Galaxy Tabs, Blackberry Playbooks etc.

Figure 5. Tablet PC and I-Pads

They offer the same functionality as portable computers. They support mobile computing in a far superior way and have enormous processing horse power [13]. User can edit and modify documents, files, access high speed intern1et, stream video and audio data, receive and send e-mails, perform lectures and presentations among very many other functions. They have an excellent screen resolution and clarity.

4.4 Ultra-Mobile PC

An ultra-mobile PC (ultra-mobile personal computer or UMPC) is a small form factor version of a pen computer, a class of laptop whose specifications were launched by Microsoft and Intel in spring 2006. Sony with its Vaio U series had manufactured the first attempt in this direction in 2004, which was however only sold in Asia. UMPCs are smaller than sub notebooks operated like tablet PCs, with a TFT display measuring (diagonally) about 12.7 to 17.8 cm, and a touch screen or a stylus. There is no distinct boundary between sub notebooks and ultra-mobile PCs. The first-generation UMPCs were just simple PCs with Linux or an adapted version of Microsoft's tablet PC operating system. With the announcement of the UMPC, Microsoft dropped the licensing requirement that tablet PCs must support proximity sensing of the stylus, which Microsoft termed "hovering". Second-generation UMPCs use less electricity and can therefore be used longer (up to five hours) and also support Windows Vista. Originally codenamed Project Origami, the project was launched in 2006 as a collaboration between Microsoft, Intel, Samsung, and a few others. Despite predictions of the demise of UMPC device category, according to CNET the UMPC category appears to continue to be in existence, however, it has largely been supplanted by tablet computers as evidenced by the introduction of Apple iPad, Google Android, Blackberry Tablet OS, and Nokia's MeeGo.

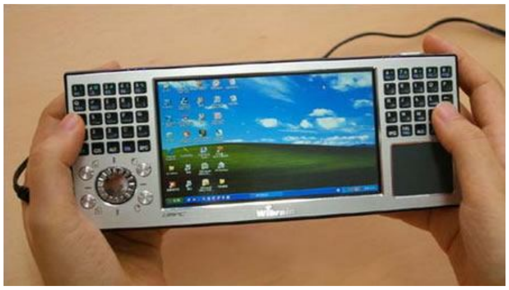


Figure 6. Ultra-Mobile PC

4.5 Wearable Computers

Wearable computers, also known as body-borne computers are miniature electronic devices that are worn by the bearer under, with or on top of clothing. This class of wearable technology has been developed for general or special purpose information technologies and media development. Wearable computers are especially useful for applications That require more complex computational support than just hardware coded logics. Figure 5 shows a wearable computer sample. One of the main features of a wearable computer is consistency [14]. There is a constant interaction between the computer and user, i.e. There is no need to turn the device on or off. Another feature is the ability to multi-task. It is not necessary to stop what you are doing to use the device; it is augmented into all other actions. These devices can be incorporated the user to act like a prosthetic. It can therefore be an extension of the user's mind and/or body. Many issues are common to the wearable as with mobile computing, ambient intelligence and ubiquitous computing research communities, including power management and heat dissipation, software architectures, wireless and personal area networks. The International Symposium on Wearable Computers is the longest-running academic conference on the subject of wearable computers.

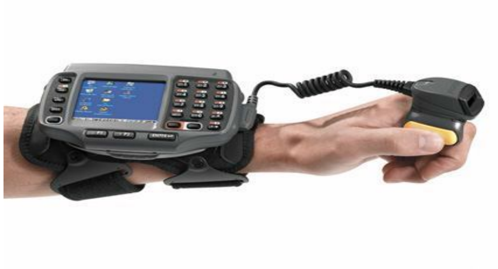


Figure 7. Wearable Computer Sample

4.6 E-Reader

An e-reader, also called an e-book reader, is designed primarily for the purpose of reading digital electronic books, magazines, and newspapers. Books from certain book sellers such as Amazon and others are available to be downloaded to the e-reader. E-readers usually have a seven inch screen, are designed with a longer battery life, and show text that can be read in the sunlight. Most recently, however, they have been designed to also connect to the Internet and have email capabilities. The older models do not use touch screens, but the newer ones do use them. They all have special operating systems designed just for them.

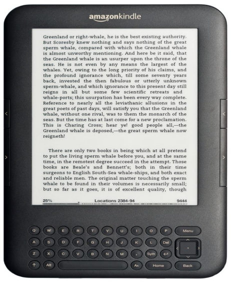


Figure 8. E-Reader

5. MOBILE OPERATING SYSTEM

A mobile operating system, also called a mobile OS, is an operating system that is specifically designed to run on mobile devices such as mobile phones, smart phones, PDAs, tablet computers and other handheld devices. The mobile operating system is the software platform on top of which other programs, called application programs, can run on mobile devices.

5.1 Symbian

Symbian OS is officially the property of Nokia. It means that any other company will have to take permission from Nokia before using this operating system. Nokia has remained a giant in low-end mobile market, so after Java, Symbian was the most used in the mobile phones till a couple of years ago. Still Symbian is widely used in low-end phones, but the demand rate has [15] continuously decreasing. By upgrading the Symbian mobile OS, Nokia has made it capable to run smartphones efficiently. Symbian ANNA and BELLE are the two latest updates which are currently used in Nokia's smartphones. Overall, the Symbian OS is excellently designed and is very user-friendly. Unfortunately, the Symbian OS graph is going downwards nowadays due to the immense popularity of Android and iOS. Some of the phones currently running on Symbian OS are Nokia C6-0, Nokia 700, Nokia 808 Pure View, Nokia E6 (ANNA) and Nokia 701 (BELLE). Symbian is a popular choice among nokia dual sim mobile phones as well. In February 2011, Nokia announced that it would replace Symbian with Windows Phone [16] as the operating system on all of its future smartphones. This transition was completed in October 2011, when Nokia announced its first line of Windows Phone 7.5 smartphones, Nokia Lumia 710 and Nokia Lumia 800. Nokia committed to support its Symbian based smartphones until 2016, by releasing further OS improvements, like Nokia Belle and Nokia Belle FP1, and new devices, like the Nokia 808 pure views.

5.2 Android

In September 20th 2008 was the date when Google released the first Android OS by the name of 'Astro'. After some time next upgrade versions 'Bender' and 'Cupcake' were also released. Google then adopted the trend of naming android versions after any dessert or a sweet in alphabetical order. The other releases are [17] Donut, Éclair, Froyo, Gingerbread, Honeycomb, Ice Cream Sandwich and Jelly Bean. Jelly Bean is so far the latest android version of google. Since the platform is not closed like IOS, there are too many great Android apps built by developers. Just after stepping into the smart phone and the tablet market, Android gained immense popularity due to its beautiful appearance and efficient working. Many new features were introduced which played a significant role in Android's success. Google Play is an official app market, which contains millions of different apps for android [18] devices. Samsung, HTC, Motorola and other top manufacturers are using Android in their devices. Currently, Android is one of the top operating systems and is considered a serious threat to the iPhone.

The system architecture consists of

- A modified Linux Kernel.

- Open source Libraries coded in C and C++.

- The Android Runtime, which considers core libraries that disposals the most core functions of Java. As virtual machines it uses Dalvin, which enables to execute Java applications.

- An Application Framework, which disposals services and libraries coded in Java for the application development.

- The Applications, which operate on it.

In an execution environment, local code is executed with full permission and has access to important system resources. On the other hand, application code is executed inside restricted areas called a sandbox. This restriction affects some specified operations such as: local file system access or invoking applications on the local system. Sandboxing enforces fixed security policies for the execution of an application. Some of the smartphones operating on the Android are HTC Desire, Samsung Galaxy Gio, Motorola Droid Razr, Samsung Galaxy S3, S4, S5 and HTC Wilfire.

5.3 Windows OS

All of you will be familiar with Windows OS because it is used in computers all over the world. Windows OS has been also been used in mobile phones, but normal mobile phone users find it a bit difficult to operate it, but at the same time it was very popular among people who were used to it. This was the case until Nokia and Microsoft joined hands to work together. The latest Windows release by Microsoft is known as Windows 7 which has gained immense popularity among all kinds of users. With its colorful and user friendly interface it has given Windows OS a new life and is currently in demand all over the world [19]. Another reason behind its success is that this latest OS is used in very powerful devices made by Nokia. The computer like look has totally vanished from the windows phones with the release of Windows 7. Samsung and HTC also released some Windows based phones, but they could not much place in the market.

Nokia Lumia series is completely windows based. Some of the latest Windows Phones are Nokia Lumia 800, Nokia Lumia 900, Samsung Focus and HTC Titan 2. If you are not on windows mobile OS and using windows for your pc, this is how you can run .jar games on your pc. Windows Phone uses technologies and tools, which are also, used in the station based application development, like the development environment Visual Studio and the Frameworks Silverlight, XNA and .NET Compact. Furthermore, Windows Phone considers a complete integration with the Microsoft Services Windows Live, Zune, Xbox Live and Bing. For sandboxing Windows Phone uses the same model like Android and iOS.

5.4 Apple iOS

The iOS was introduced in 29th June 2007 when the first iPhone was developed. Since then iOS has been under gone many upgrades and currently the latest one is the iOS 6. Apple has still not allowed any other manufacturer to lay hands on its operating system. Unlike Android, Apple has more concentrated on the performance rather than appearance. This is the reason that the basic appearance of iOS is almost the same as it was in 2007 [20]. Overall, it is very user-friendly and is one of the best operating systems in the world. So far iOS has been used in iPhone, iPhone 2G, iPhone 3G, iPhone 4 and iPhone 4S, not to mention their tablet pc's branded as iPad 3, iPad 2 and iPad [21].

The system architecture is identical to the MacOSX architecture and consists of the following components

- Core OS: The kernel of the operating system.

- Core Services: Fundamental system-services, which are subdivided in different frameworks and based on C and Objective C. For example, offers the CF Network Framework the functionality to work with known network protocols.

- Media: Considers the high-level frameworks, which are responsible for using graphic, audio and video technologies.

- Coca Touch: Includes the UIKIT, which is an Objective C based framework and provides a number of functionalities, which are necessary for the development of an iOS Application like the User Interface Management Like in the Android section mentioned, iOS uses a similar sandboxing model.

5.5 Blackberry OS

Blackberry OS is the property of RIM (Research In Motion) and was first released in 1999. RIM has developed this operating system for its Blackberry line of smartphones. Blackberry is much different from other operating systems. The interface style as well as the smart phone design is also different having a trackball for moving on the menu and a qwerty keyboard. Like Apple, Blackberry [22] OS is a close source OS and is not available from any other manufacturer. Currently the latest release of this operating system is Blackberry OS 7.1 which was introduced in May 2011 and is used in Blackberry Bold 9930. It is a very reliable OS and is immune to almost all the viruses. Some of the smartphones operating on Blackberry OS are Blackberry Bold, Blackberry Curve, Blackberry Torch and Blackberry 8520. The Blackberry OS uses an older model for application sandboxing. It uses different trust roles for assignments and applications have full [23] access to the complete device and data. It is also required to sign an application via Certificate Authorities (CA) or generated (self signed) certificate to run code on the device. Furthermore the signature provides information about the privileges for an application, which is necessary because applications have full access to Blackberry devices, because of its sandboxing model.

5.6 BADA

Like others Samsung also owns an operating system which is known as BADA. It is designed for mid range and high end smartphones. Bada is a quiet user friendly and efficient operating system, much like Android but unfortunately Samsung did not use Bada on a large scale for unknown reasons. The latest version Bada 2.0.5 was released on March 15th 2012. There are only 3 phones which are operating on Bada. These three smartphones are Samsung Wave, Samsung Wave 2 and Samsung Wave 3. I believe that Bada would have achieved much greater success if Samsung had promoted it properly. Read out how you can use Picasa on Bada mobiles [24].

Bada provides various UI controls to developers: It provides assorted basic UI controls such as List box, Color Picker and Tab, has a web browser control based on the open-source WebKit, and features Adobe Flash, supporting Flash 9, 10 or 11 (Flash Lite 4 with ActionScript 3.0 support) in Bada 2.0. Both the WebKit and Flash can be embedded inside native Bada applications. Bada supports OpenGL ES 2.0 3D graphics API and offers interactive mapping with point of interest (POI) features, which can also be embedded inside native applications. It supports pinch-to-zoom, tabbed browsing and cut, copy, and paste features. Bada supports many mechanisms to enhance interaction, which can be incorporated into applications. These include various sensors such as motion sensing, vibration control, face detection, accelerometer, magnetometer, tilt, Global Positioning System (GPS), and multi-touch. Native applications are developed in C++ with the Bada SDK, and the Eclipse based integrated development environment (IDE). GNU-based tool chains are used for building and debugging applications. The IDE also contains UI Builder, with which developers can easily design the interface of their applications by dragging and dropping UI controls into forms. For testing and debugging, the IDE contains an emulator which can run apps.

5.7 Palm OS (Garnet OS)

Palm OS was developed by Palm Inc in 1996 especially for PDAs (Personal Digital Assistance). Palm OS was basically designed to work on touch screen GUI. Some Years later it was upgraded and was able to support smartphones. Unfortunately, it could not make a mark on the market and currently is not being used in any of the latest top devices. It has been 5 and half years since we saw the latest update of Palm OS in 2007. Palm OS was used by many companies including Lenovo, Legend Group, Janam, Kyocera and IBM [25].

The key features of the current Palm OS Garnet are

• Simple, single-tasking environment to allow launching of full screen applications with a basic, common GUI set.

• Monochrome or color screens with resolutions up to 480x320 pixels.

• Handwriting recognition input system called Graffiti 2.

• HotSync technology for data synchronization with desktop computers.

• Sound playback and record capabilities.

• Simple security model: Device can be locked by password, arbitrary application records can be made private.

• TCP/IP network access.

• Serial port/USB, infrared, Bluetooth and Wi-Fi connections.

• Expansion memory card support.
• Defined standard data format for personal information management applications to store calendar, address, and task and note entries, accessible by third-party application.
.
• Included with the OS is also a set of standard applications, with the most relevant ones for the four mentioned PIM operations.

5.8 MeeGo

MeeGo was basically called a mobile platform, but it was actually designed to run multiple electronic devices including handhelds, in car devices, television sets and net books. All the devices on which MeeGo can run have the same core but the user interface is entirely different according to the device. In 2010 Moorestown Tablet PC was introduces at COMPUTEX Taipei which was also a MeeGo powered device. Most of you will have heard the name Nokia N9, but you will not be aware of the fact that this large selling device is operating in MeeGo [26] .

5.9 Maemo

Nokia and Maemo Community joined hands to produce an operating system for smartphones and internet tablets, known as Maemo. Like other devices the user interface of Maemo also comprised of a menu from which the user can go to any location. Like today's Android the home screen is divided into multiple sections which show Internet Search bar, different shortcut icons, RSS Feed and other such things. Later in 2010 at the MWC (Mobile World Congress) it was revealed that now Maemo project will be merged with Mobil in to create a fresh operating system known as MeeGo [27].

5.10 Open WebOS

Open WebOS also known as Hp WebOS or just WebOS, which was basically developed by Palm Inc but after some years it became the property of Hewlett Packard. WebOS was launched in 2009 and was used in number of smartphones and tablets. Hp promoted WebOS at a very high level by using it in high end smartphones and tablets. The latest device working on WebOS was the Hp Touch Pad. With the introduction of Android in the market sales of Hp WebOS based tablets got very less. At last Hp announced to discontinue WebOS based devices, but the existing users were assured that they will get regular updates of the operating system [28].

6. THE LIMITATIONS OF MOBILE COMPUTING

There are some general limitations for mobile computing devices. They are nominated and described in brief in follow:

6.1 Power Consumption

Power consumption plays a major part in the limitations of mobile computing, as it deals with the wireless networks battery back up are very poor in certain networks .When a power outlet is not available, mobile computers must rely entirely on battery power and most of the batteries have a back up of a few hours and need to but plugged in for future usage.

6.2 Insufficient Bandwidth

Wireless access is generally slower than the wired connection. This is mainly due to the band with allocation, mostly in developing countries. The most recent discovery in a wireless network is the 3G network where you can actually do a video conferencing. These networks are actually available within the range of near by cell phone towers; once you are out of your network access area you can't be using the latest discovery even though you have it with you. Users will be limited by the service providers .Transmission interferences also play a major role in bandwidth allocation. Connectivity in tunnels, certain buildings and in rural areas are often poor. The other major drawback chooses the network, for instance, certain phones are designed to work with CDMA and the same can't be used to using a GSM network. You need to have two different phones using both these networks. Then comes the Pay as You Go on which you can sign on a contract for one network and you get the handset to that particular network and the phone cannot be put aside to another network.

6.3 Health Hazards

Most occurrences of accidents are due to drivers who are using some form of mobile computers, most of them having a chat in their mobile phones. This occurred worldwide and many safety measures and instructions were given to the drivers regarding it and many awareness programs were conducted on it. There are allegations that the radiations from the phones cause serious health problems. World Health Organization's [29] study in 13 countries confirms radiations from the phone increases the risk of brain tumor. This is mainly due to the people who are exposed to microwaves that are emitted out from a cordless phone. Scientists have discovered that the chances of developing a glioma tumor are for people who use mobile phones for ten years. Even a normal user who uses a mobile phone for a short call will have adverse effects. Hungarian scientists have found out that 30% sperm decrease in intensive mobile phone users.

6.4 Human Interface with Device

The Screens and keyboards tend to be small, which may make them hard to use. Alternate input methods such as speech or handwriting recognition require training.

6.5 Transmission Interferences

Weather, terrain, and the range from the nearest signal point can all interfere with signal reception. Reception in tunnels, some buildings, and rural areas is often poor.

6.6 External Defects

There are various external defects, screen resolutions in some phones are poor and they don't suit to be used well on a bright sunny day, certain batteries are sensitive to high temperatures and need to be developed for charging at any condition. Touch screen plays a [30] great role with the upcoming mobile phones and it has its own drawbacks, care should be taken not to be dropped down, certain cases users need to wipe their hands dry before using their phones.

6.7 Security Standards

When working mobile, one is dependent on public networks, requiring careful use of VPN. Security is a major concern while concerning the mobile computing standards on the fleet. One can easily attack the VPN through a huge number of networks interconnected through the line.

7. APPLICATIONS OF MOBILE COMPUTING

Some of the applications of mobile computing are education and research, healthcare sector, pollution monitoring, tourism industries, airlines and railway industries, transportation industry, manufacturing and mining industries, banking and financial institutions, insurance and financial planning, hospitality industry etc. Mobile working infrastructure can deliver real time business benefits, companies of all sizes are walking up to the fact that they can improve productivity and increase profits by giving employees remote access to mission critical corporate IT system. The internet can be accessible from business, homes, and hot spots cyber cafes, available on cell phones. It is a critical business requirement, such as the oceanic fiber cuts that may result in loss of revenue and severe disruptions in networks. The required speeds have moved from supporting simple text terminals to email, the web, audio and video, requiring orders of magnitude increases in performance. It is no longer to a salesman come door to door for selling shelves full of dictionaries and encyclopedias. Rather, one can use the search engines such as Google, online dictionaries, Wikipedia etc. The written word is increasingly enhanced and replaced with graphical images, sound clips and videos. New software technology allows cell phone and PDA users to download their medical records, making them quickly accessible in case of emergency, creating room for accessing the information about the status of an airline or railway tickets. The new software to be available in years to come which can even display animated 3D scans. The computer scientists predict that the technology will also enable students to do research using their portable devices. Social networking has also taken off with applications such as Facebook, Twitter and so on. The freedom of information via Google, blogs, photos, video (You Tube), Twitter, and Wikileaks are some good examples, or police brutality is often reported first by individuals. Intellectual property, e.g. The music industry's protective stand, or how much does say Facebook or Google know about you, who your friends are, where you live, where you work, for searches made, or mining all the emails etc. The smart phones bring mobility to the internet user.

8. ISSUES IN MOBILE COMPUTING

Mobile computing is a broad area that describes a computing environment where the devices are not restricted to a single place. It is the ability of computing and communicating while on the move. Wireless networks help in the transfer of information between a computing device and a data source without a physical connection between them. In this paper I will discuss the two new issues first security issues and second issues cache management issues introduced by mobile computing.

8.1 Mobile Computing Security Issues

So some of the new security issues introduced in mobile computing are originated from the security issues of wireless networks and distributed computing systems. In addition, poorly managed mobile devices introduce new security issues involving information exposure and compromise, especially when these devices like laptops, PDAs, iPhones, Blackberries, and others are loaded with sensitive information and are stolen or fallen into the hands of an unauthorized person. Hence the new types of threats and security challenges introduced by mobile computing. Wireless networks have their own [31] security issues and challenges. This is mainly due to the fact that they use radio signals that travel through the air where they can be intercepted by location-less hacker that is difficult to track down. In addition, most wireless networks are dependent on other private networks, owned and managed by others, and in a public-shared infrastructure where you have much less control of, and knowledge about, the implemented security measures. I will discuss the main mobile computing security issues introduced by the use of wireless networks.

• Denial of Service

This attack is characterized by an explicit attempt by attackers to prevent legitimate users of a service from using that service. DOS attacks are common in all kinds of networks, but they are particularly threatening in the wireless context. This is because, the attacker does not require any physical infrastructure and he gets the necessary anonymity in the wireless environment [32]. The attacker floods the communication server or access point with a large number of connection requests so that the server keeps responding to the attacker alone hindering legitimate users from connecting and receiving the normal service.

• Pull Attacks

The attacker controls the device as a source of propriety data and control information. Data can be obtained from the device itself through the data export interfaces, a synchronized desktop, mobile applications running on the device, or the intranet servers.

• Push Attacks

The attacker uses the mobile device to plant a malicious code and spread it to infect other elements of the network. Once the mobile device inside a secure network is compromised, it could be used for attacks against other devices in the network.

• Mobility and Roaming

The mobility of users and data that they carry introduces security issues related to the presence and location of a user, the secrecy and authenticity of the data exchanged, and the privacy of user profile. To allow roaming, certain parameters and user profiles should be replicated at different locations so that when a user roams across different zones, she or he should not experience any degradation in the access and latency times. However, by replicating sensitive data across several sites, the number of points of attack is increased and hence the security risks are also increased.

• Disconnections

The frequent disconnections caused by hand-offs that occur when mobile devices across different introduce new security and integrity issues. The transition from one level of disconnection to another may present an opportunity for an attacker to masquerade either the mobile unit or the mobile support station.

• Traffic Analysis

The attacker can monitor the transmission of data, measure the load on the wireless communication channel, capture packets, and reads the source and destination fields. In order to do this, the attacker only needs to have a device with a wireless card and listen to the traffic flowing through the channel. By doing such things, the attacker can locate and trace communicating users and gain access to private information that can be subject to malicious use.

• Eavesdropping

This is a well known security issue in wireless networks. If the network is not secure enough and the transmitted information is not encrypted then an attacker can log on to the network and get access to sensitive data, as long as he or she is within range of the access point.

• Session Interception and Messages Modification

The attacker can intercept a session and alter the transmitted messages of the session. Another possible scenario by an attacker is to intercept the session by inserting a malicious host between the access point and the end host to form what is called man-in-the-middle. In this case all communications and data transmissions will go via the attacker's host.

• Captured and Retransmitted Messages

The attacker can capture a full message that has the full credential of a legitimate user and replay it with some minor but crucial modification to the same destination or to another one to gain unauthorized access and privileged to the certain computing facilities and network services.

• Information Leakage

This potential security issue lies in the possibility of information leakage, through the inference made by an attacker masquerading as a mobile support station. The attacker may issue a number of queries to the database at the user's home node or to database at other nodes, with the aim of deducing parts of the user's profile containing the patterns and history of the user's movements.

• Forced De-authentication

The attacker transmits packets intended to convince a mobile endpoint to drop its network connection and reacquire a new signal, and then inserts a crook device between a mobile device and the genuine network.

• Multi-protocol Communication

This security issue is the result of the ability of many mobile devices to operate using multiple protocols, e.g. One of the 802.11 family protocols, a cellular provider's network protocol, and other protocols which may have well-known security loopholes. Although these types of protocols aren't in active usage, many mobile devices have these interfaces set "active" by default. Attackers can take advantage of this vulnerability and connect to the device, allowing them access to extract information from it or use its services.

• Delegation

The attacker can hijack mobile session during the delegation process. A delegation is a powerful mechanism to provide flexible and dynamic access control decisions. It is a temporary permit issued by the delegator and given to the delegate who becomes limited authorized to act on the delegator's behalf. Mobile [33] devices have to switch connections between different types of networks as they move and some kind of delegation has to be issues with different network access points. Delegations may be issued and revoked frequently as mobile device detach and reattach to different parts of the network system.

• Spoofing

The attacker may hijack a session and impersonate as an authorized legitimate user to gain access to unauthorized information and services.

8.2 Cache Management Issues in Mobile Computing

Mobile Computing environments are normally known as slow wireless links and relatively underprivileged hosts with limited battery powers, are prone to frequent disconnections. Caching data [34] at the hosts in a mobile computing environment can solve the problems which are associated with slow, limited bandwidth wireless links, by reducing latency and conserving bandwidth [35]. Cache replacement, Cache Consistency, Cache Invalid action is the most frequent technique used for data management in wireless networks.

• Cache Replacement

Caching the frequently data items is considered as an effective mechanism for improving the system performance. Cache replacement algorithms are providing the solution for finding a suitable group of items from the cache [36]. Most of the cache replacement existing algorithm is based on the time since last access ,entry time of the item in the cache, hit ratio, the expiration time of the item in the cache, location etc. Most of the time cache replacement algorithm has designed in the context of [37] operating system virtual memory management and database buffer management.

• Cache Invalidation

Frequently needed data items in the database server are cached to improve transaction throughput. It is necessary to maintain the data in the cache. It must be properly invalidated, for ensuring the consistency of data. Cache Invalidation strategies permit the mobile user to re-establish the cache state from invalid stage to valid stage. The even Cache validation algorithm should consider the scarce bandwidth and limited the resources [38]. For this technique most of the time the database server involved is cache

invalidation, by sending Invalidation report (IR) to all the mobile clients. It is necessary to develop the effective cache invalidation strategies that ensure the consistency between the cached data in the mobile clients and the original data stored in the database server [39].

• Cache Consistency

Caching frequently accessed data objects in the local buffer of a mobile user (MU) can significantly improve the performance of mobile wireless networks. Marinating the cache consistency in a mobile environment [40] is a challenging task due to frequent disconnections and mobility of MUs. Several cache consistency maintenance schemes have been [41] proposed for the for mobile wireless environments. The goals of these schemes and algorithms are to ensure valid data objects in the cache to enhance their availability and minimize overhead due to consistency maintenance [42].

9. CONCLUSION

Mobile computing is dramatically changing our day-to-day lives, especially with the popularity of small devices such as personal digital assistants (PDAs) and with the embedding of substantial processing capabilities in devices such as telephones and cameras. Mobile computing offers significant benefits for organizations that choose to integrate the technology into their fixed organizational information system. Mobile computing is made possible by portable computer hardware, software, and communications systems that interact with a non-mobile organizational information system while away from the normal, fixed workplace. Mobile computing may be implemented using many combinations of hardware, software, and communications technologies. It offers a lot of benefits for everyone, especially the end users; however, it requires high security measures. In this paper, we have discussed about some of the challenging issues, applications of mobile computing along with a few of the characteristics of Mobile computing. Here in this paper we have introduced new security issues and challenges. Data management issues exhibit new challenges for both global and local. The caching techniques reduce bandwidth consumption and data access delay. Finally the computational power will be available everywhere through mobile and stationary devices that will dynamically connect and coordinate to smoothly help users in accomplishing their tasks.

10. REFERENCES

[1] C. Mascolo, L. Capra, W. Emmerich. Mobile Computing Middleware. In Tutorial Notes of Int. Conf. Networking 2002. LNCS 2497. Springer.

[2] Muller, N. J. Mobile Telecommunications factbook. New York: McGraw-Hill.

[3] M. Satyanarayanan, "Mobile computing: where's the tofu ?" ACM SIGMOBILE Mobile Computing and Communications Review,vol. 1, no. 1, pp. 17–21, 1997

[4] Nadire Cavus, Mohammad Musa Al-Momani, "Mobile system for flexible education", Procedia Computer Science, Vol. 3, pp. 1475-1479, 2011.

[5] F. Bennett, T. Richardson, and A. Harter. Teleporting - making applications mobile. In Proc. of the IEEE Workshop on Mobile Computing Systems and Applications , pages 82–84, Santa Cruz, California, Dec. 1994. IEEE Computer Society Press.

[6] D. Chalmers and M Sloman. A Survey of Quality of Service in Mobile Computing Environments. IEEE Communications Surveys , Second Quarter:2–10, 1999.

[7] Pereira, Vasco & Sousa, Tiago. "Evolution of Mobile Communications: from 1G to 4G", Department of Informatics Engineering of the University of Coimbra, Portugal 2004.

[8] Kamarularifin Abd Jalil, Mohd Hanafi Abd. Latif, Mohamad Noorman Masrek, "Looking Into The 4G Features", MASAUM Journal of Basic and Applied Sciences Vol.1, No. 2 September 2009

[9] ITU (2010). "ITU Paves the Way for Next-Generation 4G Mobile Broadband Technologies". [Online] Available: http:// www.itu.

[10] Suk Yu Hui Kai Hau Yeung , City Univ. of Hong Kong, China; Challenges in the migration to 4G mobile systems Communications Magazine, IEEE ,Vol: 41, Issue: 12 ISSN: -6804 Dec. 2003

[11] Viken, Alexander (2009). "The History of Personal Digital Assistants 1980 – 2000". Agile Mobility. http://agilemobility.net/2009/04/thehistory-of-personal digital-assistants1

[12] Andrew Nusca (20 August 2009). "Smartphone vs. feature phone arms race heats up; which did you buy?". ZDNet.

[13] 31 Percent of U.S. Internet Users Own Tablets By Angela Moscaritolo, PC Magazine, 2012.

[14] Quincy, The invention of the first wearable computer, in The Second International Symposium on Wearable Computers: Digest of Papers, IEEE Computer Society, 1998.

[15] "Symbian Device – The OS Evolution" (PDF). Independent Symbian Blog.

[16] "Ericsson Introduces The New R380e". Mobile Magazine. 2011. Available at: http://www.mobilemag. com/2001/09/25/ericsson introducesthe- new-r380e.

[17] Android, "What is android." [Online]. Available: http://developer. android.com/guide/basics/what-is android. html

[18] "Android Overview". Open Handset Alliance. 2012. Available at:vhttp://www.openhandsetalliance.com/android_ overview.html.

[19] "Windows Embedded Homepage". Microsoft.com. 2010. Available at: http://www.microsoft.com/ windows/ embedded/default.mspx.

[20] Apple, "ios overview." [Online]. Available: http: //developer.apple.com/library/ios/#referencelibrary/GettingStarted/ URL iPhone OS Overview/

[21] Apple, "iphone in business security overview." [Online]Available:http://images.apple.com/iphone/business/docs/iPhone Security.pdf

[22] "BlackBerry – Manuals and Guides". blackberry.com. 2011. Available at: http://docs.blackberry.com/en/.

[23] BlackBerry, "Blackberry enterprise server - policy reference guide." [Online]. Available: http://na.blackberry.

com/eng/deliverables/ 1417/BlackBerry Enterprise Server Policy Reference Guide.pdf

[24] "bada: un système d'exploitation pour les cellulaires Samsung". Maximejohnson.com. 2010.

[25] Piloting Palm, Andrea Butter & David Pogue, Wiley 2002.

[26] Grabham, Dan (2010-02-15). "Intel and Nokia merge Moblin and Maemo to form MeeGo". techradar.com. Retrieved 15 February 2010.

[27] Fremantle closed packages" (wiki). Mæmo. Retrieved 10 June 2013.

[28] Open webOS :: Open webOS Architecture". Openwebosproject.org. Retrieved 14 June 2013.

[29] Mobile Computing - Social Implications and Challenges, see http://wiki.mediaculture. org.au/index.php/ Mobile_Computing_-_Social_Implications_and_Challenges

[30] McKimmy, P.B. 2003. Wireless mobile instructional labs: Issues and opportunities. International Journal of Instructional Media, 30 (1) :111

[31] Adrian Leunga,*,1, Yingli Shengb, Haitham Cruickshankb, "The security challenges for mobile ubiquitous services" Information Security Group, Royal Holloway, University of London, Egham, UKbCentre for Communication Systems Research, University of Surrey, Guildford, Surrey, UK , 2007

[32] John Edwards, DOD tackles security challenges of mobile computing, Defense Systems, June 13, 2011.

[33] Q. Pham , J. Reid , A. McCullagh , Ed Dawson, "Commitment issues in delegation process", Proceedings of the sixth Australasian conference on Information security, Jan. 2008, Wollongong, Australia.

[34] Zhijun Wang ,Sajal Das ,Hao Che ,Mohan Kumar Dynamic Cache Consistency Schemes for Wireless Cellular Networks, Ieee Transactions On Wireless Communications, Vol. 5, No. 2, February 2006.

[35]Song; G. Cao, " Cache-miss-initiated prefetch in mobile environments ", IEEE International Conference on Mobile Data Management, 2004 Page(s):370 – 381

[36] Hazem Hiary, Qadri Mishael, Saleh Al-Sharaeh," Investigating Cache Technique for Location of Dependent Information Services in Mobile Environments", European Journal of Scientific Research,ISSN 1450-216X Vol.38 No.2, pp.172-179,2009.

[37] N. Megiddo and D.S. Modha, "ARC: A Self-Tuning, Low Overhead Replacement Cache," Proc. Usenix Conf. File and Storage Technologies (FAST 2003), Usenix, 2003, pp. 115-130.

[38] Miraclin Joyce Pamila J.C. and Thanushkodi K ," Performance Analysis of Improved Cache Invalidation Scheme in Mobile Computing Environment", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.9, September 2009.

[39] G. Cao, "A Scalable Low-Latency Cache Invalidation Strategy for Mobile Environments," Proc. Sixth Ann. ACM/IEEE Int'l Conf. Mobile Computing and Networking (MobiCom 2000), pp. 200-209, Aug. 2000.

[40] Kahol,Sandeep K. S et . al , "A Stretegy to manage cache consistency in a disconnected Distributed Environment " ,IEEE Transactions on Parallel and Distributed System,Volume 12 Issue 7,July 2001.

[41] A. Kahol et al., "A Strategy to Manage Cache Consistency in a Distributed Mobile Wireless Environment," IEEE Trans.Parallel and Distributed Systems, vol. 12, no. 7, 2001, pp. 686-700.

[42] J.Carter , J. BeNnett, and W. Z waenepoel ,"Techniques for reducing consistency-related communication in distributed shared memory systems," ACM Transactions on Computer Systems, 13(3):205–243,Aug.1995.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 9, 579 - 584, 2014, ISSN: 2319–8656

The Use of Intelligent Algorithms to Detect Attacks In Intrusion Detection System

Faezeh Mozneb khodaie
Department of computer,
Shabestar branch, Islamic Azad
University, Shabestar, Iran

Mohammad Ali Jabraeil Jamali
Department of computer,
Shabestar branch, Islamic Azad
University, Shabestar, Iran

Ali Farzan
Department of computer,
Shabestar branch, Islamic Azad
University, Shabestar, Iran

Abstract: More networks are connected to the Internet every day, which increases the amount of valuable data and the number of resources that can be attacked. Some systems have been designed and developed to secure these data and prevent attacks on resources. Unfortunately, new attacks are being created everyday, which makes the design of system that could catch these attacks harder. The need is not only for preventing the attack, but also to detect such an attack, if it happens. Intrusion Detection Systems is built to accomplish this task and complement other security systems. In this paper we build an Intrusion Detection System using Artificial neural networks (ANN) and Self-Organizing Map (SOM).

Keywords: intrusion detection systems; attacks; system security; artificial neural network; self-organizing map

1. INTRODUCTION

Heavy reliance on the internet and worldwide connectivity has greatly increased the potential damage that can be inflicted by remote attacks launched over the internet. It is difficult to prevent such attacks by security policies, firewalls, or other mechanisms because system and application software always contains unknown weaknesses or bugs, and because complex, often unforeseen, interactions between software components and/or network protocols are continually exploited by attackers. Intrusion detection systems are designed to detect attacks which inevitably occur despite security precautions [1]. Intrusion Detection Systems (IDS) is a piece of software or hardware that captures the inbound and outbound traffic, and analyzes it, in order to detect unusual flows. After detecting the abnormal flows, it notifies the system or the network administrator to take the appropriate action. IDS detects that a security breach happened, while firewall protects the system from security breaches. Hence they complement each other and should be used together [2].

The first concept of the IDS was introduced in 1980 by Anderson James P. [3]. In 1984 Fred Cohen mentioned that the percentage of detecting an attack will increase as the traffic increases [4]. Dorothy E. Denning introduced a model of IDS in 1986, which becomes the basic model of the current IDS models [5].

2. SECURITY OF COMPUTER SYSTEMS

Nowadays computer and Internet systems are used in almost all aspects of our lives. With the advent of personal computers and the growth of its use, Today all companies, universities and even small stores customer information, purchasing and sales and store in a computer database. One of the facilities, computer systems, computer networking systems is to establish a resource sharing among users. With the ability to connect multiple computers together, and create a computer network, protecting it from invaders came all this information and the machines. This information is critical for people trying to win others to use, alter, or destroy it.

Various measures to protect companies and home users computer resources available, But if you follow all the recommendations of the experts, the system will never be safe from attack. For this reason, users or the security of an organization, you should know the value of their and Risk analysis on it do to protect it [6].

A good security policy with a proper risk analysis by experts, the system is more resistant to the influence of many. Security is defined by three basic principles:

- Confidentiality: the attacker does not have access to confidential information.
- Integrity*:* Information may be altered or destroyed by the invaders.
- Access control: The system may be blocked so that it can not be normal.

One of the three major attempt to disturb the security of computer systems is called diffusion.

3. INTRUSION DETECTION SYSTEM

In order to combat computer systems and networks against hackers, Several methods have been established as a method for intrusion detection that The practice of monitoring the events occurring in a computer system or network play.

In general, an intrusion detection system to monitor the activities of the environment in which it operates and Eliminates unnecessary data from the data obtained, Usually a series of features to be extracted from the data collected, Then after assessment activities, the probability of an attack is considered that this procedure is done by recognizer. After identifying a suitable response system against invasions usually diagnosed offers. Most intrusion detection systems only detect attacks, and to warn the Nmayndv usually no preventive action is not the issue. The most important part of an intrusion detection system, detection is the main task is to check the data collected. Figure 1. An overview of Intrusion Detection System based on the definitions provided by the show.

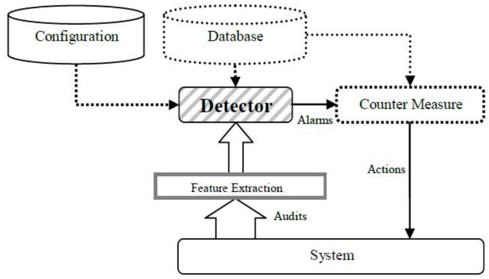


Figure. 1 Intrusion detection system

3.1 Types of IDS

There are two main types of IDS:

1) Host-based Intrusion Detection System (HIDS*):* HIDS is one of the first IDS types that were developed. Its main job is to monitor the information that flows to a computer by collecting the information that goes through and analyze it. Because of the nature of the HIDS, it has the ability to detect which process in the host computer is being under attack. This is its main advantage over other types.
2) Network-based Intrusion Detection System (NIDS):
Using the NIDS is more economical, which make it useful than any other types. The NIDS collects the packet that flows through the network to the different hosts of the network, then analyzes all the collected information and sends the results to a central system, in order to detect a possible attack. This is done by using different single purpose sensors that are placed in various points of the network [7].

3.2 Intrusion Detection Techniques

There are two basic techniques to detect an intruder, namely anomaly detection and misuse detection [2].

3.2.1 Anomaly Detection:

This technique has been developed to detect abnormal operations. It works by registering every activity in the system in a profile for hosts or network connection. If there is a sudden change in the profile, it will be treated as an abnormal activity. For example, if a normal user usually logs on to his account 2 times a day then, if in any one day he logs 20 times, the system will treat this as an abnormal and considers it as an attack.

3.2.2 Misuse Detection:

Misuse detection is also known as signature detection. It discovers any attempt to breach the Not every misuse is an attack, because some of them are just mistakes that were done by authorized ends, but every unauthorized attempt has to be taken seriously. Depending on the robustness and seriousness of a signature, some alarm, response, or notification should be sent to the proper authorities.

3.3 Types of network attacks

There are three main kinds of attacks that could be detected by the IDS: system scanning, denial of service (DoS) and system penetration. These attacks could be executed on the local machine or could be executed from a different remote machine. Every kind of these attacks should be treated differently.
1) Scanning Attacks: Before performing an attack, the attacker may search for a week point to use for attacking the system. This is performed by releasing a number of packets to some specific hosts, until vulnerable ports are discovered.
2) Denial of Service Attacks : Denial of Service (DoS) attacks is used to shut down a service that is being provided by a specific server, or to slow down the host network connection. This is done by sending infinite number of requests to the target host, until it will reach its limit and shut down.
3) Penetration Attacks: Penetration attacks target the system privileges, data and resources to alter them by an unauthorized party. This attacker could gain access to huge amount of information on the host machine and this makes it more dangerous than other attacks.

4. DATASETS KDD'99

Complex relationships exist between features, which are difficult for humans to discover. The IDS must therefore reduce the amount of data to be processed. This is very important if real-time detection is desired. The easiest way to do this is by doing an intelligent input feature selection. Certain features may contain false correlations, which hinder the process of detecting intrusions. Further, some features maybe redundant since the information they add is contained in other features. Extra features can increase computation time, and can impact the accuracy of IDS. Feature selection improves classification by searching for the subset of features, which best classifies the training data.

Feature selection is done based on the contribution the input variables made to the construction of the decision tree. Feature importance is determined by the role of each input variable either as a main splitter or as a surrogate. Surrogate splitters are defined as back-up rules that closely mimic the action of primary splitting rules. Suppose that, in a given model, the algorithm splits data according to variable 'protocol_type' and if a value for 'protocol_type' is not available, the algorithm might substitute 'flag' as a good surrogate. Variable importance, for a particular variable is the sum across all nodes in the tree of the improvement scores that the predictor has when it acts as a primary or surrogate (but not competitor) splitter.

The data for our experiments was prepared by the 1998 DARPA intrusion detection evaluation program by MIT Lincoln Labs MIT. The LAN was operated in a real environment, but was subjected to multiple attacks. For each TCP/IP connection, 41 various quantitative and qualitative features were extracted. The data set has 41 attributes for each connection record plus one class label. The data set contains 24 attack types that could be classified into four main categories.
1) DoS: Denial of service
Denial of service (DoS) is a class of attack where an attacker makes a computing or memory resource too busy or too full to handle legitimate requests, thus denying legitimate users access to a machine.

2) R2L: unauthorized access from a remote machine
A remote to user (R2L) attack is a class of attack where an attacker sends packets to a machine over a network, then exploits the machine's vulnerability to illegally gain local access as a user.

3) U2Su: unauthorized access to local super user (root)
User to root (U2Su) exploits are a class of attacks where an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system.

Probing: surveillance and other probing
4) Probing is a class of attack where an attacker scans a network to gather information or find known vulnerabilities. An attacker with a map of machines and services that are available on a network can use the information to look for exploits. Table 1. attacks in the KDD dataset based on each of the four classes above shows.

Table 1. Attacks in the KDD dataset

DOS	back, land, neptune, pod, smurf, teardrop
U2R	buffer_overflow, loadmodule, multihope, perl, rootkit
R2L	fip_write, guess_password, imap, phf, spy, warezclient, warezmaster
PROBE	ipsweep, nmap, portsweep, satan

5. NEURAL NETWIRKS

Human brain, composed of many elements, is capable of processing very elaborate and complex tasks. The brain contains billions of neurons, which are basically regarded as the most essential brain processing units. The information process is achieved by exchange of electrical pulses between these units. Neurons process information in parallel and they are connected through synaptic weights to each input in order to generate an output.

Synaptic weight refers to the significance of the established connection between an input value and a neuron. Since neurons process information in a distributed way it is possible to achieve high processing rates [8].

Neural networks terminology refers to the cluster of neurons that function or act together to solve a particular task and process information. These networks are also capable of learning through supervision or independently. Artificial neural networks (ANN) as processing models are inspired by the way nervous system work and they attempt to implement in computer systems neuron like capabilities. Three layers are present in a typical ANN: input layer, hidden layer and output layer. Each layer is composed of one or more nodes (neurons) and communication paths between them [9]. All layers connected together form a network of nodes (or neurons). Typically information flows from the input to the output layer, although in some ANN architectures a feedback flow is present. The input layer represents the stimulus or information forwarded to the network, while the output layer is the final product of the neural processing. Input layer nodes often carry out hidden relationships amongst them producing “hidden” nodes. The hidden nodes and the interaction weight between input nodes compose the hidden layer. Figure 2. shows the neural netwok layers.

The performance of neural networks depends on the architecture, algorithms and learning model chosen to collect and process data.

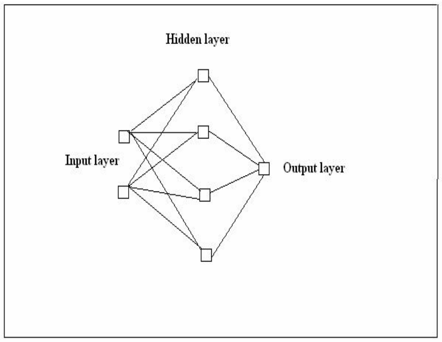


Figure. 2 ANN layers

Neural networks main features:

a) Architecture: Layer feed forward, multiple layer feed-forward, recurrent etc Single layer networks have only one layer of neurons connected individually to input points while multiple layers usually have several layers of neurons to process the data. In a single feed forward network the information move forward from input layer to output layer without backward feedback. Multiple layer models use algorithms such as back propagation to learn; output values are compared with the result values in order to correct errors. The acquired information is then forwarded back to the network for self correction. Recurrent networks use multiple layers and back propagation for learning [10].

b) Learning algorithms: There is a variety of algorithms used for learning including: error correction learning, Hebbian learning, competitive learning, self organizing maps, back propagation, snap-drift algorithm neocognition, feature map, competitive learning, adaptive resonance theory, principal component, perceptron, decision-based, multilayer perceptron, temporal dynamic model, hidden Markov model, Hamming net, Hopfield net, combinatorial optimization etc. Snapdrift in particular, performs well in frequently changing environments because of its ability to alter between minimalist learning when network performance is down and cautious learning when performance is up [11].

c) Learning model: Supervised or unsupervised. Supervised models have been the mainstream of neural development for some time. The training data consist of many pairs of input/output training patterns and the learning process relies on assistance (Kung,1993).While in the learning phase the neural network learn the desired output for a given input .Multiple layer perceptron (MLP) algorithm is used often with supervised models. In the case of unsupervised models, the network gain knowledge without specifying the required output during the learning phase. The self-organizing map (SOM) algorithm is associated frequently with unsupervised models [12].

6. SELF ORGANIZING MAP

The Self-Organizing Map is a neural network model for analyzing and visualizing high dimensional data. It belongs to the category of competitive learning network.The SOM figure 2. defines a mapping from high dimensional input data space onto a regular twodimensional array of neurons. It is a competitive network where the goal is to transform an input data set of arbitrary dimension to a one- or two-dimensional topological map. The model was first described by the Finnish professor Teuvo Kohonen and is thus sometimes referred to as a Kohonen Map. The SOM aims to discover underlying structure, e.g. feature map, of the input data set by building a topology preserving map which describes neighborhood relations of the points in the data set [13].

The SOM is often used in the fields of data compression and pattern recognition. There are also some commercial intrusion detection products that use SOM to discover anomaly traffic in networks by classifying traffic into categories. The structure of the SOM is a single feed forward network, where each source node of the input layer is connected to all output neurons. The number of the input dimensions is usually higher than the output dimension.

The neurons of the Kohonen layer in the SOM are organized into a grid, see figure 3. and are in a space separate from the input space. The algorithm tries to find clusters such that two neighboring clusters in the grid have codebook vectors close to each other in the input space. Another way to look at this is that related data in the input data set are grouped in clusters in the grid. The training utilizes competitive learning, meaning that neuron with weight vector that is most similar to the input vector is adjusted towards the input vector.The neuron is said to be the 'winning neuron' or the Best Matching Unit (BMU). The weights of the neurons close to the winning neuron are also adjusted but the magnitude of the change depends on the physical distance from the winning neuron and it is also decreased with the time.

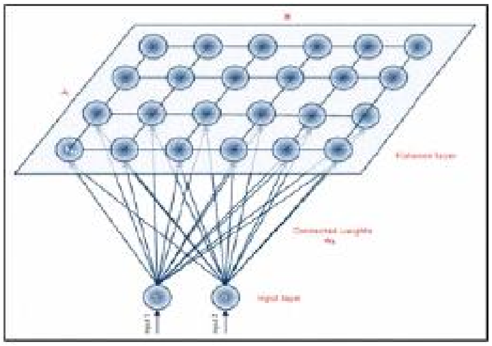


Figure. 3 Self-Organizing (Kohonen) Map

The learning process of the SaM goes as follows:

1) One sample vector x is randomly drawn from the input data set and its similarity (distance) to the codebook vectors is computed by using Euclidean distance measure [14]:

$$\| x - m_c \| = \min_i \{\| x - m_i \|\} \qquad (1)$$

2) After the BMU has been found, the codebook vectors are updated. The BMU itself as well as its topological neighbors are moved closer to the input vector in the input space I.e. the input vector attracts them. The magnitude of the attraction is governed by the learning rate. As the learning proceeds and new input vectors are given to the map, the learning rate gradually decreases to zero according to the specified learning rate function type. Along with the learning rate, the neighborhood radius decreases as well. The update rule for the reference vector of unit i is the following:

$$m_i(t+1) = m_i + a(t)h_{ci}(r(t))[x(t) - m_i(t)] \qquad (2)$$

3) The steps 1 and 2 together constitute a single training step and they are repeated until the training ends. The number of training steps must be fixed prior to trainingthe SaM because the rate of convergence in the neighborhood function and the learning rate are calculated accordingly.

After the training is over, the map should be topologically ordered. This means that *n* topologically close input data vectors map to *n* adjacent map neurons or even to the same single neuron.

5.1 Mapping Precision

The mapping precision measure describes how

accurately the neurons respond to the given data set. If the reference vector of the BMU calculated for a given testing vector xi is exactly the same xi, the error in precision is then 0. Normally, the number of data vectors exceeds the number of neurons and the precision error is thus always different from 0. A common measure that calculates the precision of the mapping is the average quantization error over the entire data set:

$$E_q = \frac{1}{N}\sum_{i=1}^{N} \| x_{i+} m_c \| \qquad (3)$$

5.2 Topology Preservation

The topology preservation measure describes how well the SOM preserves the topology of the studied data set. Unlike the mapping precision measure, it considers the structure of the map. For a strangely twisted map, the topographic error is big even if the mapping precision error is small. A simple method for calculating the topographic error:

$$E_q = \frac{1}{N}\sum_{i=1}^{N} u_x(x) \qquad (4)$$

Where $u(x_k)$ is 1 if the first and second BMUs of x_k are not next to each other. Otherwise $u(x_k)$ is 0.

7. SYSTEM ARCHITECTURE

Architecture for intrusion detection system based on self-organizing map and artificial Neural Networks. Figure 4. shows the general view of the system. This system uses two detection layers used to detect and isolate attacks. The first layer of a self-organizing map And the next layer of the neural network 1 and 2 were separately taken. The task of separating the first layer attacks from normal traffic. self-organizing map layer, first taught by normal traffic data. In fact, in this episode we have a sample of intrusion detection system to detect anomalies. This means that if the vector has been recognized by the SOM is determined as part of the normal traffic Otherwise be regarded as an attack. The main task of this layer is actually separating normal traffic from attack traffic. The next layer is the output layer, in general, two routes that one of the normal traffic that has been detected in the neural network together and other vectors in the direction of the attack has been detected in the neural network together.

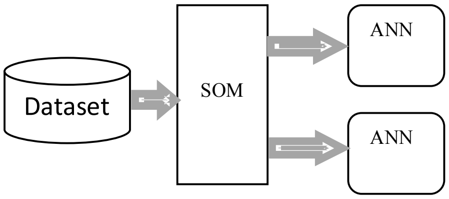


Figure. 4 System architecture

Firstly, using a dataset consisting of 13,472 normal vector is given, self-organizing map of the train. Then in the second step using a set of normalized data and attack that include Vector Data is 72,102, we label the neurons of the self-organizing map this means that each neuron is responsible for one type of attack. The third stage is the main test system, Test data including vector data is 18,794, First, we give to self-organizing map and after determining the direction (attack or normal) to one of the neural networks and this type of attack is detected by neural networks.

7.1 The proposed system simulation parameters

The first layer is a flat topology self-organizing map the dimensions of 50*40 interlocking hexagonal. Gaussian neighborhood function and the learning algorithm used is batch. The second layer of the neural network with separate entrance View of 41 neurons, 10 neurons in the middle and 1 output neuron is used. The size of the training data set includes 72,102 self-organizing map and neural network vector data. Table 2. Number of data vectors in the series to show each type of attack and the size of the testing data set included 18,794 data vectors.

Table 2. The number of data vectors in the training and testing data set self-organizing map and neural network the type of attack

Attack Type	Count (Train)	Count (Test)
Dos	45927	5741
U2R	52	37
R2L	995	2199
Probe	11656	1106
Normal	13472	9711

8. RESULTS AND EVALUATION CRITERIA

Evaluation criteria in the system is calculated as shown in Table 4. All these criteria are based on the accuracy is measured.

Table 4. Evaluation results

Criteria	Error Count	Percent %
Total Error	2735	85.45
False Positive	460	95.27
True Negative	1099	87.91
Attack Type DOS Error	272	95.27
Attack Type U2R Error	33	10.82
Attack Type R2L Error	1953	11.19
Attack Type PROBE Error	17	98.47

9. CONCLUSIONS

The new system has a very high accuracy and speed of detection compared to other methods of attack. The system is also able to detect and classify them by type of attacks.

10. REFERENCES

[1] Lippmann R., Haines J.W., Fried D. J.,Korba J., Das K., "Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation", . Recent Advances in Intrusion Detection 2000: 162-182, 2000.

[2] http://www.securityfocus.com/infocus/1520 - An introduction to IDS, (last checked 15/July/2009).

[3] Anderson, James P., "Computer Security Threat Monitoring and Surveillance," Washing, PA, James P. Anderson Co., 1980.

[4] Cohen, Fred, "Computer Viruses: Theory and Experiments," 7thDOD/NBS Computer Security Conference, Gaithersburg, MD, September 24-26, 1984.

[5] Denning, Dorothy E., "An Intrusion Detection Model," Proceedings of the Seventh IEEE Symposium on Security and Privacy, May 1986.

[6] Gollmann, D. (2002), "Computer Security", New Jersey, Wiley.

[7] Rebecca B., Peter M., "NIST Special Publication on Intrusion Detection System" http://danielowen.com /NIDS, (last checked 15/July/2009).

[8] Silva, L., Santos, A., Silva, J., Montes, A.: A neural network application for attack detection in computer networks (2004).

[9] Smith, S.: The Scientist & Engineer's Guide to Digital Signal Processing. California Technical Publishing, USA (1998).

[10] Hagan, T., Demuth, H., Beale, M.: Neural network design. PWS Publishing, USA (1996).

[11] Palmer-Brown, D., Lee, S.: Continuous reinforced snap-drift learning in a neural architecture for proxylet selection in active computer networks (2004).

[12] Planquet, J.: Application of neural networks to Intrusion Detection systems (2001).

[13] Kohonen, T, "Self-Organizing Maps", Springer Series in Information Sciences. Berlin, Heidelberg: Springer. 2006.

[14] P. Lichodzijewski, A. Zincir-Heywood, and M. Heywood. "Dynamic intrusion detection using self organizing maps", 2002.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 9, 585 - 588, 2014, ISSN: 2319–8656

Intrusion Detection System Using Self Organizing Map Algorithms

Faezeh Mozneb khodaie
Department of computer,
Shabestar branch, Islamic Azad
University, Shabestar, Iran

Mohammad Ali Jabraeil Jamali
Department of computer,
Shabestar branch, Islamic Azad
University, Shabestar, Iran

Ali Farzan
Department of computer,
Shabestar branch, Islamic Azad
University, Shabestar, Iran

Abstract: With the rapid expansion of computer usage and computer network the security of the computer system has became very important. Every day new kind of attacks are being faced by industries. Many methods have been proposed for the development of intrusion detection system using artificial intelligence technique. In this paper we will have a look at an algorithm based on neural networks that are suitable for Intrusion Detection Systems (IDS). The name of this algorithm is "Self Organizing Maps" (SOM). So far, many different methods have been used to build a detector that Wide variety of different ways in the covers. Among the methods used to detect attacks in intrusion detection is done, In this paper we investigate the Self-Organizing Map method.

Keywords: Intrusion Detection System; Self Organizing Maps; Attacks; Security; neural network

1. INTRODUCTION

The goal of intrusion detection is to discover unauthorized use of computer systems. Existing intrusion detection approaches can be divided into two classes - anomaly detection and misuse detection. Anomaly detection approaches the problem by attempting to find deviations from the established patterns of usage. Misuse detection, on the other hand, compares the usage patterns to known techniques of compromising computer security. Architecturally, an intrusion detection system can be categorized into three types – hostbased IDS, network-based IDS and hybrid IDS. Host-based IDS, deployed in individual hostmachines, can monitor audit data of a single host. Network-based IDS monitors the traffic data sent and received by hosts. Hybrid IDS uses both methods.

Self-Organizing Map has been successfully applied in complex application areas where traditional method has failed. Due to their inherently non-linear nature, they can handle much more complex situations than the traditional methods. One of those problems represents intrusion detection by intrusion detection systems. These systems deal with high dimension data on the input, which is needed to map to 2-dimension space. Designed architecture of the intrusion detection system is application of neural network SOM in IDS systems. Over the last few decades information is the most precious part of any organization. Most of the things what an organization does revolve around this important asset. Organizations are taking measures to safeguard this information from intruders. The rapid development and expansion of World Wide Web and local networks and their usage in any industry has changed the computing world by leaps and bounds [1][2].

2. INTRUSION DETECTION SYSTEMS

Intrusion Detection System is a system that identifies , in real time, attacks on a network and takes corrective action to prevent them. They are the set of techniques that are used to detect suspicious activity both at network and host level. There are two main approaches to design an IDS.

1) Misuse Based Ids (Signature Based)
2) Anomaly Based Ids.

In a misuse based intrusion detection system , intrusions are detected by looking for activities that correspond to know signatures of intrusions or vulnerabilities [3]. While an anomaly based intrusion detection system detect intrusions by searching for abnormal network traffic . The abnormal traffic pattern can be defmed either as the violation of accepted thresholds for frequency of events in a connection or as a user's violation of the legitimate profile developed for normal behavior.

One of the most commonly used approaches in expert system based intrusion detection systems is rule-based analysis using Denning's profile model [3]. Rule-based analysis depends on sets of predefined rules that are provided by an administrator. Expert systems require frequent updates to remain current. This design approach usually results in an inflexible detection system that is unable to detect an attack if the sequence of events is slightly different from the predefined profile [4]. Considered that the intruder is an intelligent and flexible agent while the rule based IDSs obey fixed rules . This problem can be tackled by the application of soft computing techniques in IDSs. Soft computing is a general term for describing a set of optimization and processing techniques. The principal constituents of soft computing techniques are Fuzzy Logic (FL), Artificial Neural Networks (ANNs), Probabilistic Reasoning (PR), and Genetic Algorithms (GAs) [4].

3. TYPES OF NETWORKING ATTACKS

There are four major categories of networking attacks. Every attack on a network can be placed into one of these groupings [4].

3.1 Denial of Service (DoS):

A DoS attacks is a type of attack in which the hacker makes a memory resources too busy to serve legitimate networking requests and hence denying users access to a machine e.g. apache, smurf, Neptune, ping of death, back, mail bomb, UDP storm, etc.

3.2 Remote to User attacks (R2L):

A remote to user attack is an attack in which a user sends packets to a machine over the internet, and the user does not have access to in order to expose the machines vulnerabilities and exploit privileges which a local user would have on the computer, e.g. xlock, guest, xnsnoop, phf, sendmail dictionary etc.

3.3 User to Root Attacks (U2R): These attacks are exploitations in which the hacker starts off on the system with a normal user account and attempts to abuse vulnerabilities in the system in order to gain super user privileges, e.g. perl, xterm.

3.4 Probing: Probing is an attack in which the hacker scans a machine or a networking device in order to determine weaknesses or vulnerabilities that may later be exploited so as to compromise the system. This technique is commonly used in data mining, e.g. satan, saint, portsweep, mscan, nmap etc.

4. SELF ORGANIZING MAP

The Self-Organizing Map [5] is a neural network model for analyzing and visualizing high dimensional data. It belongs to the category of competitive learning network. The SOM Figure 1. defines a mapping from high dimensional input data space onto a regular two dimensional array of neurons.

In designed architecture is input vector with six input values and output is realized to 2 dimension space. Every neuron i of the map is associated with an n dimensional reference vector $m_i\left[m_1,........,m_n\right]^T$, where n denotes the dimension of the input vectors. The reference vectors together form a codebook. The neurons of the map are connected to adjacent neurons by a neighborhood relation, which dictates the topology, or the structure, of the map. Adjacent neurons belong to the neighborhood Ni of the neuron i. In the SOM algorithm, the topology and the number of neurons remain fixed from the beginning. The number of neurons determines the granularity of the mapping, which has an effect on the accuracy and generalization of the SOM. During the training phase, the SOM forms an elastic net that is formed by input data. The algorithm controls the net so that it strives to approximate the density of the data. The reference vectors in the codebook drift to the areas where the density of the input data is high. Eventually, only few codebook vectors lie in areas where the input data is sparse.

The learning process of the SOM goes as follows:

1. One sample vector x is randomly drawn from the input data set and its similarity (distance) to the codebook vectors is computed by using Euclidean distance measure:

$$\| x - m_c \| = \min_i \{ \| x - m_i \| \}$$

2. After the BMU has been found, the codebook vectors are updated. The BMU itself as well as its topological neighbors are moved closer to the input vector in the input space i.e. the input vector attracts them. The magnitude of the attraction is governed by the learning rate. As the learning proceeds and new input vectors are given to the map, the learning rate gradually decreases to zero according to the specified learning rate function type. Along with the learning rate, the neighborhood radius decreases as well. The update rule for the reference vector of unit i is the following:

$$m_i(t+1) = m_i + a(t)h_{ci}(r(t))[x(t) - m_i(t)]$$

3. The steps 1 and 2 together constitute a single training step and they are repeated until the training ends. The number of training steps must be fixed prior to training the SOM because the rate of convergence in the neighborhood function and the learning rate are calculated accordingly.

After the training is over, the map should be topologically ordered. This means that n topologically close input data vectors map to n adjacent map neurons or even to the same single neuron.

4.1 Mapping precision

The mapping precision measure describes how accurately the neurons respond to the given data set. If the reference vector of the BMU calculated for a given testing vector xi is exactly the same xi, the error in precision is then 0. Normally, the number of data vectors exceeds the number of neurons and the precision error is thus always different from 0. A common measure that calculates the precision of the mapping is the average quantization error over the entire data set:

$$E_q = \frac{1}{N}\sum_{i=1}^{N}\left\| x_{i+} m_c \right\|$$

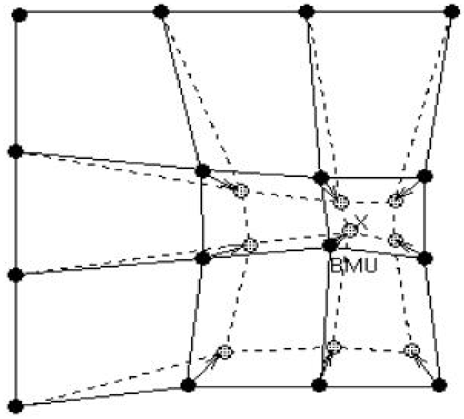


Figure. 1 General SOM topology

2.2 Topology preservation

The topology preservation measure describes how well the SOM preserves the topology of the studied data set. Unlike the mapping precision measure, it considers the structure of the map. For a strangely twisted map, the topographic error is big even if the mapping precision error is small.

A simple method for calculating the topographic error:

$$E_q = \frac{1}{N}\sum_{i=1}^{N} u_x\left(x\right)$$

where $u\left(x_k\right)$ is 1 if the first and second BMUs of x_k are not next to each other. Otherwise $u\left(x_k\right)$ is 0.

5. THE ARCHITECTURE SELF-ORGANIZING MAP METHOD

The Self-Organizing Map method is mapped to the data Normal initially trained Then the mixture of normal and attack data to be tagged. After this step, the experimental data to give mapped To determine whether the input vector The normal vector or a vector of attack. If that BMU selected is a normal neuron with labels In this case the normal vector of the detected Otherwise traffic in general, the attack is detected [6]. Figure 2. shows the architecture.

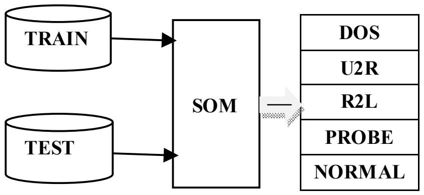


Figure. 2 Architecture Self-Organizing Map method

Training data set in this data set is a mixture of normal and attack. Table 1. the number and type of data in the training data set shows. The size of the test data sets are shown in Table 2.

Table 1. The number of data vectors in the training data set

Attack Type	Count
Dos	45927
U2R	52
R2L	995
Probe	11656
Normal	26944

Table 2. The number of data vectors in the data set to test the type of attack

Attack Type	Count
Dos	5741
U2R	37
R2L	2199
Probe	1106
Normal	9711

5. RESULTS AND EVALUATION CRITERIA

To simulate the Self-Organizing Map of the simulation tool box in MATLAB is used for Self-Organizing Map (SOM TOOLBOX, 2012). Evaluation criteria used are as follows:

5.1 Total Error

Percentage of The total number of errors made The data have been trained And test data.

5.2 False Positive

Event normal system as an attack is detected. This event is not an attack, but the attack was seen. When we tested the data And compare them with data from the trained If you have been attacked, and the attack has been detected This is an error.

5.3 True Negative

Activities or events without risk of That have been labeled as normal activity. The event of an attack, but the attack has not been seen.

Table 3. shows the results of the evaluation on the basis of these criteria indicates the number of error found.

Table 3. Evaluation results show that the Self-Organizing Map based on these criteria.

Method Criteria	Self-Organizing Map Method	
	Error Count	Accuracy
Total Error	9384	50.07
False Positive	9384	0.06
True Negative	0	0

6. CONCLUSIONS

The Self Organizing Map is an extremely powerful mechanism for automatic mathematical characterization of acceptable system activity. In the above paper we have described how we can use Self Organizing Maps for building an Intrusion Detection System. We have explained the system architecture and the flow diagram for the SOMe We have also presented the pros and cons of the algorithm.

The results show that Algorithms used in the Self-Organizing Map method gives the optimal solutions to large amounts of data. The Self-Organizing Map method is trained only with normal data Thus, errors can not be calculated for each type of attack.

7. REFERENCES

[1] Damiano Bolzoni, Sandro Etalle, Pieter H. Hartel, andEmmanuele Zambon. Poseidon: a 2-tier anomaly-based networkintrusion detection system. In Proceedings of the 4th IEEE International Workshop on Information

Assurance, 13-14 April 2006, Egham, Surrey, UK, pages 144-156, 2006.

[2] D. A. Frincke, D. Tobin, 1. C. McConnell, 1. Marconi, and D. Polla. A framework for cooperative intrusion detection. In Proc. 21st NIST-NCSC National Information Systems SecurityConference, pages 361-373, 1998.

[3] Denning D, "An Intrusion-Detection Model", IEEE Transactionson Software Engineering, Vol. SE-13, No 2, Feb 1987.

[4] Simon Haykin, "Neural Networks: A ComprehensiveFoundation", Prentice Hall, 2nd edition, 1999.

[5] Kohonen, T. 1995. Self-Organizing Maps, volume 30 of Springer Series in InformationSciences. Berlin, Heidelberg: Springer. (SecondExtended Edition 1997).

[6] Kohonen T., Oja E., Simula O., Visa A., Kangas J., , "Engineering applications of the self-selforganizing map.", Proceedings of the IEEE, Vol. 84, Issue: 10, Pages: 1358 – 1384, 1996.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 9, 589 - 591, 2014, ISSN: 2319–8656

Steganography using Interpolation and LSB with Cryptography on Video Images-A Review

Jagdeep Kaur
Computer Science Department
UIET, Kurukshetra University
Kurukshetra, India

Abstract: Stegnography is the most common term used in the IT industry, which specifically means, "covered writing" and is derived from the Greek language. Stegnography is defined as the art and science of invisible communication i.e. it hides the existence of the communication between the sender and the receiver. In distinction to Cryptography, where the opponent is permitted to detect, interrupt and alter messages without being able to breach definite security grounds guaranteed by the cryptosystem, the prime objective of Stegnography is to conceal messages inside other risk-free messages in a manner that does not agree to any enemy to even sense that there is any second message present. Nowadays, it is an emerging area which is used for secured data transmission over any public medium such as internet. In this research a novel approach of image stegnography based on LSB (Least Significant Bit) insertion and cryptography method for the lossless jpeg images has been projected. This paper is comprising an application which ranks images in a users library on the basis of their appropriateness as cover objects for some facts. Here, the data is matched to an image, so there is a less possibility of an invader being able to employ steganalysis to recuperate the data. Furthermore, the application first encrypts the data by means of cryptography and message bits that are to be hidden are embedded into the image using Least Significant Bits insertion technique. Moreover, interpolation is used to increase the density

Keywords: Cryptography, Stegnography, LSB

1. INTRODUCTION

As living in the society, human beings have repeatedly sought innovative and well-organized ways to communicate. The most primitive methods included smoke signals, cave drawings and drums. With the advancements of civilization introduced written language, telegraph, radio/television, and most newly electronic mail. Nowadays, almost each and every communication is carried out electronically; new requirements, issues and opportunities are born. At times when we communicate, we prefer that only the intended recipient have the ability to decipher the contents of the communication in order to keep the message covert. One of the common solution to resolve this problem is the use of encryption. Whilst encryption masks the significance of a communication, instances do exist where it would be preferred that the entire communication process is not obvious to any observer, even the fact that communication is taking place is kept secret. In this case, the communication taking place is hidden. Steganography can be used to conceal or cover the existence of communication. A major negative aspect to encryption is that the existence of data is not hidden. Data that has been encrypted, although unreadable, still exists as data. If given an adequate amount of time, someone could eventually decrypt that data. A solution to this dilemma is steganography.

2. DIFFERENT KINDS OF STEGNOGRAPHY

Approximately all digital file formats can be used for stegnography; but the formats that are more appropriate are those with a high level of redundancy. The term redundancy can be defined as the bits of an object that provide accurateness far greater than needed for the object's use and display. Also, the redundant bits of an object are those bits that can be changed without the alteration being detected easily. Image and audio files particularly meet the terms with this prerequisite, while research has also uncovered other file formats that can be used for information hiding.

Figure 1 shows the four main categories of file formats that can be used for steganography.

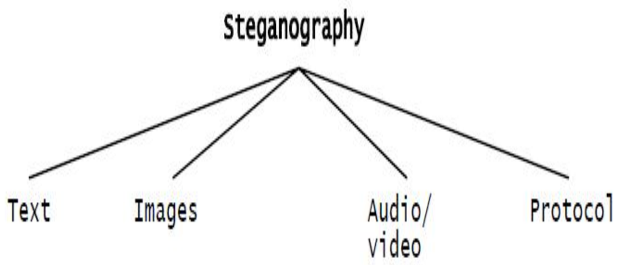


Figure 1 Types of Steganography.

Image steganography is about exploiting the inadequate powers of the human visual system (HVS). Within reason, any cipher text, plain text, images, or anything else that can be embedded in a bit stream can be concealed in an image.

Moreover, image steganography has come quite far in current years with the expansion of fast, influential graphical computers.

Digital image is the most important and common type of carrier used for steganography. A digital image is composed of finite number of elements each of which has a particular location and value (gray scale). The processing of these digital images by means of a digital Computer is referred as digital image processing. The images are used for steganography in the following ways.

The message or the data either in encrypted form or in the unique form is embedded as the covert message to be sent into a graphic file. This method results in the production of what is called a stego-image. An additional secret data may be required in the hiding process e.g. a stegokey. Furthermore, the stego-image is then transmitted to the receiver. After that, the recipient extracts the message from the carrier image. The message can only be extracted if both the sender and the recipient has a shared secret between them.

This could be the algorithm for extraction or a special parameter such as a key. A stego-analyst or attacker may try to intercept the stego-image. The computer based stenography allows changes to be made to what are known as digital carriers such as sounds or images. The changes represent the hidden message, but result is successful if their is no discernible change to the carrier. The information has nothing to do with the carrier sound or image. Information might be about the carrier such as the author or a digital watermark or fingerprint.

Stegnography applications that hide data in images generally use a variation of least significant bit (LSB) embedding . In LSB embedding, the data is hidden in the least significant bit of each byte in the image. The size of each pixel depends on the format of the image and normally ranges from 1 byte to 3 bytes. Each unique numerical pixel value corresponds to a color; thus, an 8-bit pixel is capable of displaying 256 different colors .Given two identical images, if the least significant bits of the pixels in one image are changed, then the two images still look identical to the human eye. This is because the human eye is not sensitive enough to notice the difference in color between pixels that are different by 1 unit. Thus, stegnography applications use LSB embedding because attackers do not notice anything odd or suspicious about an image if any of the pixel's least significant bits are customized.

3. CRYPTOGRAPHY

Cryptography[8] is the study of mathematical techniques related to aspects of information security such as confidentiality, data integrity, entity authentication, and data origin authentication. In this paper we will focus only on confidentiality, i.e., the service used to keep the content of information from all but those authorized to have it.

Cryptography protects the information by transforming it into an incomprehensible format. It is useful to achieve private transmission over a public network. Also, the original text, or *plaintext*, is transformed into a coded alike called *ciphertext* via any encryption algorithm. Only those who hold a secret key can decipher (*decrypt*) the ciphertext into plaintext. Cryptography systems can be broadly classified into symmetric-key systems that use a single key (i.e., a *password*) that both the sender and the receiver have for their piece of work and a public-key systems that use two keys, a public key known to everyone and a private key that is unique and only the recipient of messages uses it. In the rest of this paper, we will discuss only symmetric-key systems.

Cryptography and stegnography are close cousins in the spy craft family: the former scrambles a message so it cannot be understood and the latter hides the message so it cannot be seen. A cipher message, for illustration, might arouse suspicion on the part of the recipient whilst an invisible message created with stegnographic methods will not.

In fact, stegnography can be useful when the use of cryptography is forbidden; where cryptography and strong encryption are barred, steganography can get around such policies to pass message covertly. However, stegnography and cryptography differ in the way in which they are evaluated; stegnography fails when the "enemy" is able to access the content of the cipher message, while cryptography fails when the "enemy" detects that there is a secret message present in the stegnographic medium .

The disciplines that study techniques for deciphering cipher messages and detecting hide messages are called *cryptanalysis* and *steganalysis*. The former denotes the set of methods for obtaining the meaning of encrypted information, while the latter is the art of discovering covert messages

4. DIFFERENCE BETWEEN CRYPTOGRAPHY AND STEGNOGRAPHY

In cryptography, the system is broken when the attacker can read the secret message. Breaking a stegnographic system has two stages:

1. The attacker can detect that stegnography has been used.

2. Additionally, he is able to read the embedded message.

In our definition a stegnographic system is insecure already if the detection of stegnography is possible (first stage).

5. CONCLUSIONS

The Steganography has its place in the security. On its own, it won't serve much but when used as a layer of cryptography, it would lead to a greater security.

Although only some of the main image steganographic techniques were discussed in this paper, one can see that there exists a large selection of approaches to hiding information in images. All the major image file formats have different methods of hiding messages, with different strong and weak points respectively. Where one technique lacks in payload capacity, the other lacks in robustness.

Steganography, particularly pooled with cryptography is a commanding tool which enables people to converse without possible eavesdroppers even knowing there is a form of communication in the first place. The proposed method provides acceptable image quality with very little deformation in the image. The main benefit of this System is to provide high security for key information exchanging. It is also useful in communications for codes self error correction. It can embed remedial audio or image data in case corruption occurs due to poor connection or transmission

6. REFERENCES

[1]Awrangjeb M (2003) An overview of reversible data hiding. ICCIT 75–79

[2]Celik MU, Sharman G, Tekalp AM & Saber E (2002) Reversible data hiding, Proceedings of IEEE 2002

International Conference on Image Processing 2, 157–160

[3]Chan CK, Cheng LM (2004) Hiding data in images by simple LSB substitution. Pattern Recognition 37:469–474

[4] Chang CC, Lin MH, Hu YC (2002) A fast and secure image hiding scheme based on LSB substitution. Int Pattern Recog 16(4):399–416

[5]GoljanM, Fredrich F & Du R (2001) Distortion-free data embedding, Proceedings of 4th Information Hiding Workshop, 27–41

[6] Huang LC, Tseng LY, Hwang MS (2013) A reversible data hiding method by histogram shifting in high quality medical images. J Syst Software 86:716–727

[7]Johnson NF & Jajodia S (1998) Exploring steganography: seeing the unseen. Comput Pract 26–34

[8]Jung KH, Yoo KY (2009) Data hiding method using image interpolation. Comput Standards Interfaces 31:465–470

[9] Artz, D., “Digital Steganography: Hiding Data within Data”, *IEEE Internet Computing Journal*, June 2001

[10] Hameed A. Younis, Dr. Turki Y. Abdalla, Dr. Abdulkareem Y. Abdalla , “ A Modified Technique For Image Encryption “,online access

[11] Simmons, G. J. The prisoners’ problem and the subliminal channel. In Advances in Cryptology: Proceedings of Crypto 83, pages 51–67. Plenum Press.

[12]Westfeld, A. (2001). F5-a steganographic algorithm: High capacity despite better steganalysis. In Proc. 4th Int’l Workshop Information Hiding, pages 289–302.2001

An Evaluation of Two-Step Techniques for Positive-Unlabeled Learning in Text Classification

Azam Kaboutari
Computer Department
Islamic Azad University,
Shabestar Branch
Shabestar, Iran

Jamshid Bagherzadeh
Computer Department
Urmia University
Urmia, Iran

Fatemeh Kheradmand
Biochemistry Department
Urmia University of Medical Sciences
Urmia, Iran

Abstract: Positive-unlabeled (PU) learning is a learning problem which uses a semi-supervised method for learning. In PU learning problem, the aim is to build an accurate binary classifier without the need to collect negative examples for training. Two-step approach is a solution for PU learning problem that consists of tow steps: (1) Identifying a set of reliable negative documents. (2) Building a classifier iteratively. In this paper we evaluate five combinations of techniques for two-step strategy. We found that using Rocchio method in step 1 and Expectation-Maximization method in step 2 is most effective combination in our experiments.

Keywords: PU Learning; positive-unlabeled learning; one-class classification; text classification; partially supervised learning

1. INTRODUCTION

In recent years, the traditional machine learning task division into supervised and unsupervised categories is blurred and a new type of learning problems has been raised due to the emergence of real-world problems. One of these partially supervised learning problems is the problem of learning from positive and unlabeled examples and called Positive-Unlabeled learning or PU learning [2]. PU learning assumes two-class classification, but there are no labeled negative examples for training. The training data is only a small set of labeled positive examples and a large set of unlabeled examples. In this paper the problem is supposed in the context of text classification and Web page classification.

The PU learning problem occurs frequently in Web and text retrieval applications, because Oftentimes the user is looking for documents related to a special subject. In this application collecting some positive documents from the Web or any other source is relatively easy. But Collecting negative training documents is especially requiring strenuous effort because (1) negative training examples must uniformly represent the universal set, excluding the positive class and (2) manually collected negative training documents could be biased because of human's unintentional prejudice, which could be detrimental to classification accuracy [6]. PU learning resolves need for manually collecting negative training examples.

In PU learning problem, learning is done from a set of positive examples and a collection of unlabeled examples. Unlabeled set indicates random samples of the universal set for which the class of each sample is arbitrary and may be positive or negative. Random sampling in Web can be done directly from the Internet or it can be done in most databases, warehouses, and search engine databases (e.g., DMOZ[1]).

Two kinds of solutions have been proposed to build PU classifiers: the two-step approach and the direct approach. In this paper, we review some techniques that are proposed for step 1 and step 2 in the two-step approach and evaluate their performance on our dataset that is collected for identifying diabetes and non-diabetes WebPages. We find that using Rocchio method in step 1 and Expectation-Maximization method in step 2 seems particularly promising for PU Learning.

The next section provides an overview of PU learning and describes the PU learning techniques considered in the evaluation - the evaluation is presented in section 3. The paper concludes with a summary and some proposals for further research in section 4.

2. POSITIVE-UNLABELED LEARNING

PU learning includes a collection of techniques for training a binary classifier on positive and unlabeled examples only. Traditional binary classifiers for text or Web pages require laborious preprocessing to collect and labeling positive and negative training examples. In text classification, the labeling is typically performed manually by reading the documents, which is a time consuming task and can be very labor intensive. PU learning does not need full supervision, and therefore is able to reduce the labeling effort.

Two sets of examples are available for training in PU learning: the positive set P and an unlabeled set U. The set U contains both positive and negative examples, but label of these examples not specified. The aim is to build an accurate binary classifier without the need to collect negative examples. [2]

To build PU classifier, two kinds of approaches have been proposed: the two-step approach that is illustrated in Figure 1 and the direct approach. In The two-step approach as its name indicates there are two steps for learning: (1) Extracting a subset of documents from the unlabeled set, as reliable negative (RN), (2) Applying a classification algorithm iteratively, building some classifiers and then selecting a good classifier. [2]

Two-step approaches include S-EM [3], PEBL [6], Roc-SVM [7] and CR-SVM [8]. Direct approaches such as biased-SVM [4] and Probability Estimation [5] also are offered to solve the problem. In this paper, we suppose some two-step approaches for review and evaluation.

[1] http://www.dmoz.org/

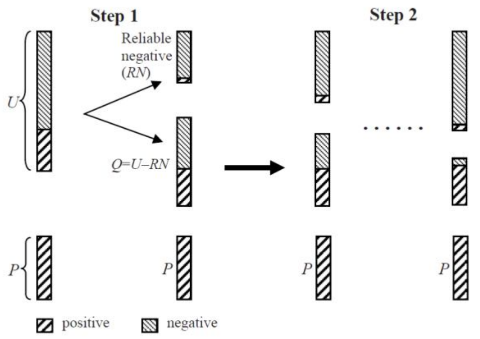


Figure 1. Two-step approach in PU learning [2].

2.1 Techniques for Step 1

For extracting a subset of documents from the unlabeled set, as reliable negative five techniques are proposed:

2.1.1 Spy

In this technique small percentage of positive documents from P are sampled randomly and put in U to act as “spies” and new sets Ps and Us are made respectively. Then the naïve Bayesian (NB) algorithm runs using the set Ps as positive and the set Us as negative. The NB classifier is then applied to assign a probabilistic class label Pr(+1|d) to each document d in Us. The probabilistic labels of the spies are used to decide which documents are most likely to be negative. S-EM [3] uses Spy technique.

2.1.2 Cosine-Rocchio

It first computes similarities of the unlabeled documents in U with the positive documents in P using the cosine measure and extracts a set of potential negatives PN from U. Then the algorithm applies the Rocchio classification method to build a classifier f using P and PN. Those documents in U that are classified as negatives by f are regarded as the final reliable negatives and stored in set RN. This method is used in [8].

2.1.3 1DNF

It first finds the set of words W as positive words that occur in the positive documents more frequently than in the unlabeled set, then those documents from the unlabeled set that do not contain any positive words in W extracted as reliable negative and used for building set RN. This method is employed in PEBL [6].

2.1.4 Naïve Bayesian

It builds a NB classifier using the set P as positive and the set U as negative. The NB classifier is then applied to classify each document in U. Those documents that are classified as negative denoted by RN. [4]

2.1.5 Rocchio

This technique is the same as that in the previous technique except that NB is replaced with Rocchio. Roc-SVM [7] uses Rocchio technique.

2.2 Techniques for Step 2

If the set RN contains mostly negative documents and is sufficiently large, a learning algorithm such as SVM using P and RN applied in this step and it works very well and will be able to build a good classifier. But often a very small set of negative documents identified in step 1 especially with 1DNF technique, then a learning algorithm iteratively runs till it converges or some stopping criterion is met. [2]

For iteratively learning approach two techniques proposed:

2.2.1 EM-NB

This method is the combination of naïve Bayesian classification (NB) and the EM algorithm. The Expectation-Maximization (EM) algorithm is an iterative algorithm for maximum likelihood estimation in problems with missing data [1].

The EM algorithm consists of two steps, the Expectation step that fills in the missing data, and the Maximization step that estimates parameters. Estimating parameters leads to the next iteration of the algorithm. EM converges when its parameters stabilize.

In this case the documents in Q (= U−RN) regarded as having missing class. First, a NB classifier f is constructed from set P as positive and set RN as negative. Then EM iteratively runs and in Expectation step, uses f to assign a probabilistic class labels to each document in Q. In the Maximization step a new NB classifier f is learned from P, RN and Q. The classifier f from the last iteration is the result. This method is used in [3].

2.2.2 SVM Based

In this method, SVM is run iteratively using P, RN and Q. In each iteration, a new SVM classifier f is constructed from set P as positive and set RN as negative, and then f is applied to classify the documents in Q. The set of documents in Q that are classified as negative is removed from Q and added to RN. The iteration stops when no document in Q is classified as negative. The final classifier is the result. This method, called I-SVM is used in [6].

In the other similar method that is used in [7] and [4], after iterative SVM converges, either the first or the last classifier selected as the final classifier. The method, called SVM-IS.

3. EVALUATION

3.1 Data Set

We suppose the Internet as the universal set in our experiments. To collect random samples of Web pages as unlabeled set U we used DMOZ, a free open Web directory containing millions of Web pages. To construct an unbiased sample of the Internet, a random sampling of a search engine database such as DMOZ is sufficient [6].

We randomly selected 5,700 pages from DMOZ to collect unbiased unlabeled data. We also manually collected 539 Web pages about diabetes as positive set P to construct a classifier for classifying diabetes and non-diabetes Web pages. For evaluating the classifier, we manually collected 2500 non-diabetes pages and 600 diabetes page. (We collected negative data just for evaluating the classifier.)

3.2 Performance Measure

Since the F-score is a good performance measure for binary classification, we report the result of our experiments with this measure. F-score is the harmonic mean of precision and recall. Precision is defined as the number of correct positive predictions divided by number of positive predictions. Recall is defined as the number of correct positive predictions divided by number of positive data.

3.3 Experimental Results

We present the experimental results in this subsection. We extracted features from normal text of the content of Web pages, and then we perform stopwording, lowercasing and stemming. Finally, we get a set of about 176,000 words. We used document frequency (DF), one of the simple unsupervised feature selection methods for vocabulary and vector dimensionality reduction [9].

The document frequency of a word is the number of documents containing the word in the training set, in our case in P∪U. Then we create a ranked list of features, and returns the i highest ranked features as selected features, which i is in {200, 400, 600, 1000, 2000, 3000, 5000, 10000}.

As discussed in Section 2, we studied 5 techniques for Step 1 and 3 techniques for Step 2 (EM-NB, I-SVM and SVM-IS). Clearly, each technique for first step can be combined with each technique for the second step. In this paper, we will empirically evaluate only the 5 possible combinations of methods of Step 1 and Step 2 that available in the LPU[2], a text learning or classification system, which learns from a set of positive documents and a set of unlabeled documents.

These combinations are S-SVM which is Spy combined with SVM-IS, Roc-SVM is Rocchio combined with SVM-IS, Roc-EM is Rocchio+EM-NB, NB-SVM is Naïve Bayesian+ SVM-IS and NB-EM is Naïve Bayesian+ EM-NB.

In our experiments, each document is represented by a vector of selected features, using a bag-of-words representation and term frequency (TF) weighting method which the value of each feature in each document is the number of times (frequency count) that the feature (word) appeared in the document. When running SVM in Step 2, the feature counts are automatically converted to normalized tf-idf values by LPU. The F-score is shown in Figure 2.

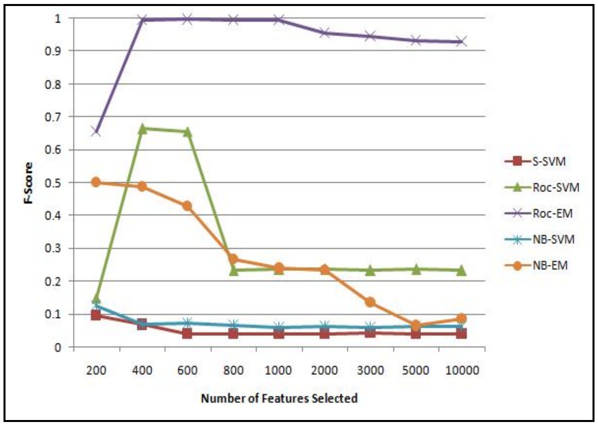


Figure 2. Results of LPU using DF feature selection method.

As Figure 2 shows, very poor results are obtained in S-SVM which Spy is used in Step 1 and SVM-IS is used in Step 2. Since we obtain better results in other combinations that SVM-IS is used in Step 2, we conduct that Spy in not a good technique for Step 1 in our experiments. By using NB in step 2, results are improved and best results we have obtained in our experiments when using Rocchio technique in Step 1. Figure 2 also shows that how using EM-NB instead of SVM-IS in Step 2 can improve results significantly.

The average of all F-score in each combination of techniques of Step 1 and Step 2 are shown in Table 1. As seen in Table 1 and Figure 2 Roc-EM is the best combination in our experiments which Rocchio technique is used in Step 1 and EM-NB is used in Step 2.

Table 1. Comparison of two-step approaches results.

	S-SVM	Roc-SVM	Roc-EM	NB-SVM	NB-EM
Average F-score	0.0489	0.3191	0.9332	0.0698	0.2713

4. CONCLUSIONS

In this paper, we discussed some methods for learning a classifier from positive and unlabeled documents using the two-step strategy. An evaluation of 5 combinations of techniques of Step 1 and Step 2 that available in the LPU system was conducted to compare the performance of each combination, which enables us to draw some important conclusions. Our results show that in the general Rocchio technique in step 1 outperforms other techniques. Also, we found that using EM for the second step performs better than SVM. Finally, we observed best combination for LPU in our experiments is R-EM, which is Rocchio, combined with EM-NB.

In our future studies, we plan to evaluate other combinations for Step 1 and Step 2 for Positive-Unlabeled Learning.

5. REFERENCES

[1] Dempster, N. Laird and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society. Series B (Methodological), 1977, 39(1): pp. 1-38.

[2] Liu and W. Lee, “Partially supervised learning”, In “Web data mining”, 2nd ed., Springer Berlin Heidelberg, 2011, pp. 171-208.

[3] Liu, W. Lee, P. Yu and X. Li, “Partially supervised classification of text documents,” In Proceedings of International Conference on Machine Learning(ICML-2002), 2002.

[4] B. Liu, Y. Dai, X. Li, W. Lee and Ph. Yu, “Building text classifiers using positive and unlabeled examples,” In Proceedings of IEEE International Conference on Data Mining (ICDM-2003), 2003.

[5] Elkan and K. Noto, “Learning classifiers from only positive and unlabeled data,” In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2008), 2008.

[6] H. Yu, J. Han and K. Chang, “PEBL: Web page classification without negative examples”, Knowledge and Data Engineering, IEEE Transactions on , vol.16, no.1, pp. 70- 81, Jan. 2004.

[7] X. Li and B. Liu. "Learning to classify texts using positive and unlabeled data". In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-2003), 2003.

[8] X. Li, B. Liu and S. Ng, “Negative Training Data can be Harmful to Text Classification,” In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010), 2010.

[9] X. Qi and B. Davison, “Web page classification: Features and algorithms,” ACM Comput. Surv., 41(2): pp 1–31, 2009.

[2] http://www.cs.uic.edu/~liub/LPU/LPU-download.html

International Journal of Computer Applications Technology and Research
Volume 3– Issue 9, 595 - 599, 2014, ISSN: 2319–8656

An Evaluation of Feature Selection Methods for Positive-Unlabeled Learning in Text Classification

Azam Kaboutari
Computer Department
Islamic Azad University,
Shabestar Branch
Shabestar, Iran

Jamshid Bagherzadeh
Computer Department
Urmia University
Urmia, Iran

Fatemeh Kheradmand
Biochemistry Department
Urmia University of Medical Sciences
Urmia, Iran

Abstract: Feature Selection is important in the processing of data in domains such as text because such data can be of very high dimension. Because in positive-unlabeled (PU) learning problems, there are no labeled negative data for training, we need unsupervised feature selection methods that do not use the class information in the training documents when selecting features for the classifier. There are few feature selection methods that are available for use in document classification with PU learning. In this paper we evaluate four unsupervised methods including, collection frequency (CF), document frequency (DF), collection frequency-inverse document frequency (CF-IDF) and term frequency-document frequency (TF-DF). We found DF most effective in our experiments.

Keywords: feature selection; unsupervised feature selection; positive-unlabeled learning; PU learning; document classification

1. INTRODUCTION

Feature selection for classification is the process of selecting a subset of relevant features among many input features and to remove any redundant or irrelevant one. The default in classifying text documents is to use terms as features. Feature selection reduces the dimensionality of the feature space, which leads to a reduction in computational burden. Furthermore, in some cases, classification can be more accurate in the reduced space. [12]

Many methods for feature selection have been presented. Most of these methods are supervised that use the class information in the training data when selecting features for the classifier. Hence, for supervised methods to be usable, a pre-classified set of documents must be available.

In recent years, a new type of learning problems has been raised due to the emergence of real-world problems that blurred traditional machine learning tasks division into supervised and unsupervised categories. These are partially supervised learning problems that do not need full supervision. One of these problems is the problem of learning from positive and unlabeled examples. This problem, called Positive-Unlabeled learning or PU learning [2], assumes two-class classification. However, the training data only has a small set of labeled positive examples and a large set of unlabeled examples, but no labeled negative examples. We suppose this problem in the context of text classification and Web page classification.

So, supervised feature selection methods cannot be applied for the feature selection of the PU learning problem when there are no available training data for the second class. However, there are few feature selection methods that are unsupervised and available for use in partially supervised learning problems. In Unsupervised feature selection methods, the training data does not need to be manually classified. All that is needed is a fixed set of documents the classifier is to be used on. Hence, these methods are handy for PU learning problem.

In Web and text retrieval applications, the PU learning problem occurs frequently, because most of the time the user is only interested in documents of a particular topic. In this application positive documents are usually available or collecting some from the Web or any other source is relatively easy. But Collecting negative training documents is especially delicate and arduous because (1) negative training examples must uniformly represent the universal set excluding the positive class and (2) manually collected negative training documents could be biased because of human's unintentional prejudice, which could be detrimental to classification accuracy [8]. PU learning eliminates the need for manually collecting negative training documents.

PU learns from a set of positive data as well as a collection of unlabeled data. Unlabeled data indicates random samples of the universal set for which the class of each sample is arbitrary and uncorrelated. Random sampling can be done in most databases, warehouses, and search engine databases (e.g., DMOZ[1]) or it can be done independently directly from the Internet. So the dimensions of feature space that contains the terms appearing in the training (positive and unlabeled) documents will be very high and need for effective methods for feature selection is essential.

In this paper we review some unsupervised feature selection methods and evaluate their performance on a number of PU learning techniques. We find that feature selection based on document frequency seems particularly promising for PU Learning.

In the next section we review some related works that focused on evaluation of feature selection methods for text classification. Section 3 provide an overview of PU learning and describe the PU learning techniques included in the evaluation. In section 4 we describe some unsupervised feature selection methods considered in the evaluation - the evaluation is presented in section 5. The paper concludes with a summary and some proposals for further research in section 6.

2. RELATED WORK

Previous feature selection studies for text domain consider the problem of selecting one set of features for multi-class classification. These problems are traditional classification

[1] http://www.dmoz.org/

problems that labeled examples for each class are available for use in training and often supervised methods are applied for feature selection.

For example a review of traditional feature selection methods used in text classification can be found in [14]. This study considered five feature selection metrics, including document frequency (DF), information gain (IG), mutual information (MI), χ^2-test (CHI) and term strength (TS) and found that IG and CHI are most effective in their experiments.

Another work [6] presents an empirical comparison of twelve feature selection methods. In addition, a new feature selection method, called bi-normal separation, is shown to outperform other commonly known methods in some circumstances.

In other study [7], ten feature selection methods including a new feature selection method, called the GU metric were evaluated. The experiments were performed on the 20 Newsgroups data sets with the Naive Probabilistic Classifier. The results show that the GU metric obtained best F-score.

3. POSITIVE-UNLABELED LEARNING

One of the difficulties of supervised learning algorithms is that a large number of labeled examples are needed in order to learn accurately. In text classification, the labeling is typically performed manually by reading the documents, which is a time consuming task and can be very labor intensive. Partially supervised learning problems such as PU learning do not need full supervision, and therefore are able to reduce the labeling effort.

PU learning is a collection of techniques for training binary classifier on positive and unlabeled examples only. Traditional binary classifiers for text or Web pages require laborious preprocessing to collect positive and negative training examples.

In PU learning [2], two sets of examples are available for training: the positive set P and an unlabeled set U, which is assumed to contain both positive and negative examples, but without these being labeled as such. The aim is to build an accurate binary classifier without the need to collect negative examples.

Two kinds of approaches have been suggested to build PU classifiers: the two-step approach and the direct approach. The two-step approach as its name indicates consists of two steps: (1) extracting some reliable negative (RN) documents from the unlabeled set, (2) Constructing a set of classifiers by using a classification algorithm iteratively and then selecting a good classifier from the set. These approaches include S-EM [3], PEBL [8], Roc-SVM [10] and CR-SVM [11]. Direct approaches such as biased-SVM [4] and Probability Estimation [5] also are offered to solve the problem.

3.1 Techniques for Step 1

In two-step approaches five techniques proposed for step 1:

3.1.1 Spy

It randomly samples small percentage of positive documents from P and put them in U to act as “spies”. Thus new sets Ps and Us are made respectively. Then runs the naïve Bayesian (NB) algorithm using the set Ps as positive and the set Us as negative. The NB classifier is then applied to assign each document d in Us a probabilistic class label Pr(+1|d). It uses the probabilistic labels of the spies to decide which documents are most likely to be negative. S-EM [3] uses Spy technique.

3.1.2 Cosine-Rocchio

It first extracts a set of potential negatives PN from U by computing similarities of the unlabeled documents in U with the positive documents in P using the cosine measure. To extract the final reliable negatives, the algorithm applies the Rocchio classification method to build a classifier f using P and PN. Those documents in U that are classified as negatives by f are regarded as the final reliable negatives and stored in set RN. This method is used in [11].

3.1.3 1DNF

It first find the set of words W that occur in the positive documents more frequently than in the unlabeled set, then extract those documents from unlabeled set that do not contain any word in W. These documents form the reliable negative documents. This method is employed in PEBL [8].

3.1.4 Naïve Bayesian

It runs the naïve Bayesian (NB) algorithm using the set P as positive and the set U as negative. The NB classifier is then applied to classify each document in U. Those documents that are classified as negative documents denoted by RN. This method is employed in [4].

3.1.5 Rocchio

The algorithm is the same as that in previous technique except that NB is replaced with Rocchio. This method is used in Roc-SVM [10].

3.2 Techniques for Step 2

If the reliable negative set RN is sufficiently large and contains mostly negative documents, a learning algorithm such as SVM using P and RN used in this step and it works very well. But if a very small set of negative documents identified in step 1, then running a learning algorithm will not be able to build a good classifier, rather a learning algorithm iteratively till it converges or some stopping criterion is met. For iteratively learning approach two techniques proposed, which are based on EM and SVM respectively.

3.2.1 EM-NB

This method is based on naïve Bayesian classification (NB) and the EM algorithm. The Expectation-Maximization (EM) algorithm is an iterative algorithm for maximum likelihood estimation in problems with missing data [1]. The EM algorithm consists of two steps, the Expectation step that fills in the missing data, and the Maximization step that estimates parameters. Estimating parameters leads to the next iteration of the algorithm. EM converges when its parameters stabilize. In this case the documents in Q (= U−RN) regarded as having missing class. First, a NB classifier f is constructed from set P as positive and set RN as negative. Then EM iteratively runs and in Expectation step, uses f to assign a probabilistic class labels to each document in Q. In the Maximization step a new NB classifier f is learned from P, RN and Q. The classifier f from the last iteration is the result. This method is used in [3].

3.2.2 SVM Based

In this method, SVM is run iteratively using P, RN and Q (= U-RN). In each iteration, a new SVM classifier f is constructed from set P as positive and set RN as negative, and then f is applied to classify the documents in Q. The set of documents in Q that are classified as negative is removed from Q and added to RN. The iteration stops when no document in Q is classified as negative. The final classifier is the result. This method, called I-SVM is used in [8]. In the other similar method that is used in [10] and [4], after iterative

SVM converges, either the first or the last classifier selected as the final classifier. The method, called SVM-IS.

4. FEATURE SELECTION METHODS

There are two main categories of feature selection methods: filters and wrappers. In filter methods feature scoring metrics are used on each feature for measure feature relevance and ranking features. Wrapper methods perform a search algorithm like greedy hill-climbing over the space of all feature subsets, repeatedly calling the same induction algorithm that is later used for building the classifier, as a subroutine to evaluate subsets of features. Where filter methods evaluate each feature independently, wrappers evaluate feature sets as a whole, which would avoid redundant features and lead to better results. However, wrapper methods are often impractical and very computationally intensive for large datasets, and are also more prone to overfitting, so filter methods are more commonly used.

Unsupervised feature selection methods [9] are methods that do not use the class information in the training data when selecting features for the classifier. It means that the training data does not need to be manually pre-classified. All that is needed is a fixed set of documents from the collection the classifier is to be used on. Hence, these methods are handy if there is no pre-classified training data available, and if there is no time to create such data. So these methods are suitable for PU learning. However, pre-classified documents are of course needed for evaluation of the classifier's performance.

In the current study we choose four unsupervised filter methods for feature selection in PU Learning:

4.1 Collection Frequency (CF)

The collection frequency [9] of a feature is the total number of instances of the feature in the collection, in our case in $P \cup U$. It does not look at which documents or categories the feature occurs in, it is simply a count.

4.2 Document Frequency (DF)

One of the simplest methods of vocabulary reduction and vector dimensionality reduction is the document frequency [12]. The document frequency of a feature is the number of documents containing a feature in the training set, in our case in $P \cup U$.

4.3 Collection Frequency-Inverse Document Frequency (CF-IDF)

The CF-IDF [9] is computed by weighting the collection frequency values by the inverse document frequency for feature:

$$\text{CF}-\text{IDF}\ (w) = \text{CF}\ (w) \times \log_2 (N / \text{DF}\ (w)) \qquad (1)$$

Where w denoted feature and N is the total number of documents in the training data, in our case $N= |P \cup U|$.

4.4 Term Frequency-Document Frequency (TF-DF)

In [13], a method based on the term frequency combined with the document frequency is presented. They call it Term Frequency-Document Frequency, and prove it better than DF measure. TF-DF for feature w is computed as follows:

$$\text{TF}-\text{DF}\ (w) = (n_0 \times n_1 + c\ (n_0 \times n_2 + n_1 \times n_2)) \quad (2)$$

Where $c \geq 1$ is a constant, n0 is the number of documents in the training data without the feature; n1 is the number of documents where the feature occurs exactly once, n2 is the number of documents where the feature occurs twice or more. As the value of c increases, we give more weight for multiple occurrences of a term. The authors of [13] use c=10 in their experiments, and we follow this decision in our experiments.

5. EVALUATION

5.1 Data Set

In our experiments the universal set is the Internet. We used DMOZ, which is a free open directory of the Web containing millions of Web pages, to collect random samples of Internet pages as unlabeled set U. To construct an unbiased sample of the Internet, a random sampling of a search engine database such as DMOZ is sufficient [8]. We randomly selected 5,700 pages from DMOZ to collect unbiased unlabeled data. We also manually collected 539 Web page about diabetes as positive set P to construct a classifier for classify diabetes and non diabetes Web pages. For evaluating the classifier, we manually collected 2500 non-diabetes pages and 600 diabetes page. (We collected negative data just for evaluating the classifier we construct.)

5.2 Performance Measure

We report the result with F-score, a good performance measure for binary classification. F-score is the harmonic mean of precision and recall. Precision is defined as number of correct positive predictions division by number of positive predictions. Recall is defined as number of correct positive predictions division by number of positive data.

5.3 Experimental Results

We now present the experimental results. We extracted features from normal text of the content of Web pages, and then we perform stopwording, lowercasing and stemming. Finally we get a set of about 176,000 words. We used four methods which is discussed briefly in Section IV in our evaluation and create a ranked list of features, and returns the i highest ranked features as selected features, which i is in {200, 400, 600, 1000, 2000, 3000, 5000, 10000}.

As discussed in Section III, we studied 5 techniques for Step 1 and 3 techniques for Step 2 (EM-NB, I-SVM and SVM-IS). Clearly, each technique for first step can be combined with each technique for second step. In this paper, we will empirically evaluate only the 5 possible combinations of methods of Step 1 and Step 2 that available in the LPU[2], a text learning or classification system, which learns from a set of positive documents and a set of unlabeled documents. These combinations are S-SVM which is Spy combined with SVM-IS, Roc-SVM is Rocchio combined with SVM-IS, Roc-EM is Rocchio+EM-NB, NB-SVM is Naïve Bayesian+ SVM-IS and NB-EM is Naïve Bayesian+ EM-NB.

In our experiments each document is represented by a vector of selected features, using a bag-of-words representation and term frequency (TF) weighting method which the value of each feature in each document is the number of times (frequency count) that the feature (word) appeared in the document. When running SVM in Step 2, the feature counts are automatically converted to normalized tf-idf values by LPU. The F-score for 5 combinations of methods of Step 1 and Step 2 are shown in Figure 1 to 5. In each combination we perform an evaluation of 4 feature selection methods.

[2] http://www.cs.uic.edu/~liub/LPU/LPU-download.html

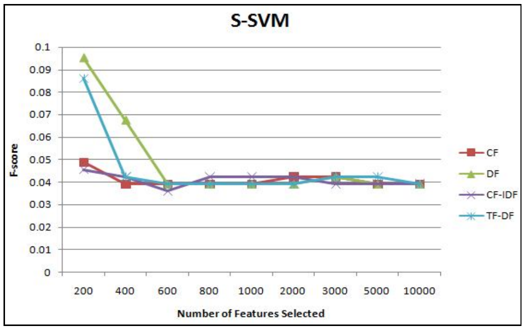


Figure 1. Results of LPU (Spy in Step 1 and SVM-IS in step 2) using 4 feature selection methods.

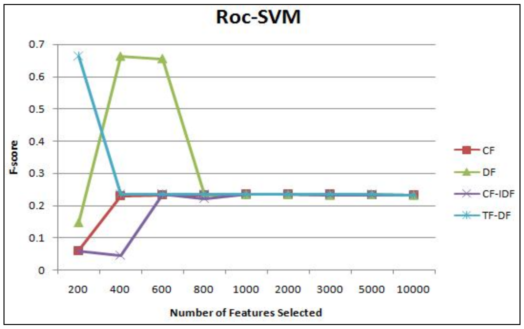


Figure 2. Results of LPU (Rocchio in Step 1 and SVM-IS in step 2) using 4 feature selection methods.

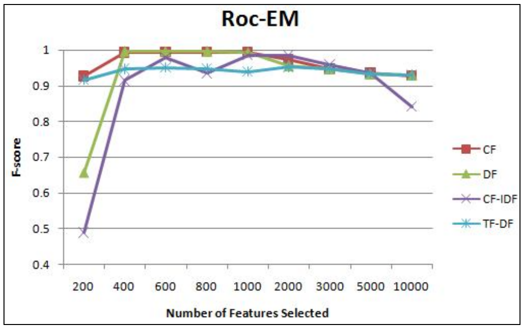


Figure 3. Results of LPU (Rocchio in Step 1 and EM-NB in step 2) using 4 feature selection methods

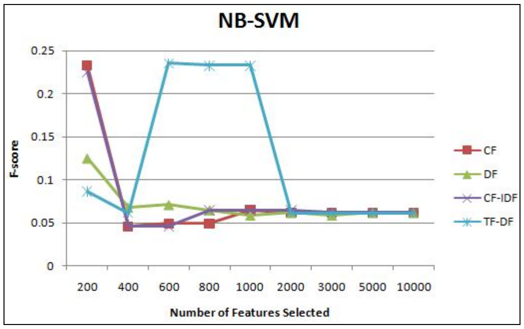


Figure 4. Results of LPU (Naïve Bayesian in Step 1 and SVM-IS in step 2) using 4 feature selection methods.

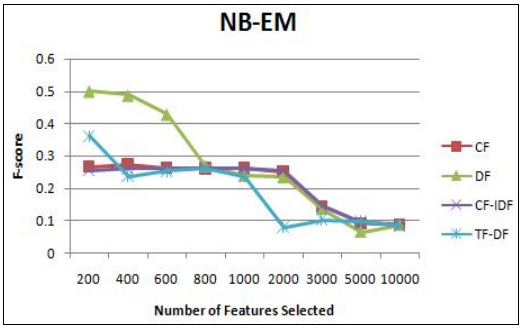


Figure 5. Results of LPU (Naïve Bayesian in Step 1 and EM-NB in step 2) using 4 feature selection methods

As Figure 1 shows, very poor results are obtained using feature selection methods in S-SVM which Spy is used in Step 1 and SVM-IS is used in Step 2. Since we obtain better results in other combinations that SVM-IS is used in Step 2, we conduct that Spy in not good technique for Step 1 in our experiments.

Figure 2 shows that when using Rocchio technique in Step 1, better results can be achieved using all feature selection methods. In this case, DF method in average is better than other feature selection methods.

Figure 3 shows the best results we have obtained in our experiments. As can be seen in Figure 3, when number of feature is 400 and more, all 4 feature selection methods can achieve good results, but CF method results in average is better than others. Figure 3 also shows that how using EM-NB instead of SVM-IS in Step 2 can improve results of all feature selection methods significantly.

Figure 4 shows results of 4 feature selection methods when Naïve Bayesian is used for Step 1 and SVM-IS for Step 2. In this case also we have obtained poor results. Best result in average is obtained from TF-DF method that is 0.122. When using EM-NB instead of SVM-IS in Step 2, results are improved. These results are shown in Figure 5. In this case, with increasing the dimension of feature space, the results are worse. Best result in average is obtained from DF method.

The average results of 4 feature selection methods in each combination of techniques of Step 1 and Step 2 are shown in Table 1. Last column indicate the method that achieved best result among other methods.

Table 1. Comparison of feature selection methods.

Methods	CF	DF	CF-IDF	TF-DF	Best
S-SVM	0.041	0.049	0.041	0.045	DF
Roc-SVM	0.214	0.319	0.192	0.282	DF
Roc-EM	0.964	0.933	0.891	0.94	CF
NB-SVM	0.076	0.07	0.077	0.122	TF-DF
NB-EM	0.212	0.271	0.208	0.191	DF

6. CONCLUSIONS

In this paper, we discussed the 4 unsupervised methods for feature selection in learning a classifier from positive and unlabeled documents using the two-step strategy. An evaluation of 5 combinations of techniques of Step 1 and Step

2 that available in the LPU system was conducted to compare the performance of each feature selection method in each combination, which enables us to draw some important conclusions. Our results show that in general Document Frequency method outperforms other methods in most case. Also we found that best combination for LPU in our experiments is R-EM, which is Rocchio, combined with EM-NB. In this combination best results are obtained by the Collection Frequency method.

In our future studies, we plan to evaluate other combinations for Step 1 and Step 2 and other unsupervised feature selection methods for Positive-Unlabeled Learning.

7. REFERENCES

[1] A. Dempster, N. Laird and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society. Series B (Methodological), 1977, 39(1): p. 1-38.

[2] B. Liu and W. Lee, “Partially supervised learning”, In “Web data mining”, 2nd ed., Springer Berlin Heidelberg, 2011, pp. 171-208.

[3] B. Liu, W. Lee, P. Yu and X. Li, “Partially supervised classification of text documents,” In Proceedings of International Conference on Machine Learning(ICML-2002), 2002.

[4] B. Liu, Y. Dai, X. Li, W. Lee and Ph. Yu, “Building text classifiers using positive and unlabeled examples,” In Proceedings of IEEE International Conference on Data Mining (ICDM-2003), 2003.

[5] C. Elkan and K. Noto, “Learning classifiers from only positive and unlabeled data,” In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2008), 2008.

[6] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” The Journal of Machine Learning Research, 3, 3/1/2003.

[7] G. Uchyigit, “Experimental evaluation of feature selection methods for text classification,” Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on , vol., no., pp.1294,1298, 29-31 May 2012.

[8] H. Yu, J. Han and K. Chang, “PEBL: Web page classification without negative examples”, Knowledge and Data Engineering, IEEE Transactions on , vol.16, no.1, pp. 70- 81, Jan. 2004.

[9] Ø. Garnes, “Feature selection for text categorisation,” Master's thesis, Norwegian University of Science and Technology, 2009.

[10] X. Li and B. Liu. "Learning to classify texts using positive and unlabeled data". In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-2003), 2003.

[11] X. Li, B. Liu and S. Ng, “Negative Training Data can be Harmful to Text Classification,” In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010), 2010.

[12] X. Qi and B. Davison, “Web page classification: Features and algorithms,” ACM Comput. Surv., 41(2):1–31, 2009.

[13] Y. Xu, B. Wang, J. Li and H. Jing, “An extended document frequency metric for feature selection in text categorization,” Proceedings of the 4th Asia information retrieval conference on Information retrieval technology, January 15-18, 2008, Harbin, China.

[14] Y. Yang, J. Pedersen, “A comparative study on feature selection in text categorization”. In Proceedings of the Fourteenth International Conference on Machine Learning (ICML). Morgan Kaufmann, San Francisco, CA, 412–420,1997.

The Impact of Mobility Models on the Performance of AODV, DSR and LAR Routing Protocols

Veena Garg
Samsung Research Institute
Noida, India

Poonam Mittal
Department of Computer Engineering
YMCA University
Faridabad, India

Abstract: MANETs are the collection of wireless nodes that can dynamically form a network anytime and anywhere to exchange information without using any pre-existing infrastructure. There are some challenges that make the design of mobile ad hoc network routing protocols a tough task. Firstly, in mobile ad hoc networks, node mobility causes frequent topology changes and network partitions. Secondly, because of the variable and unpredictable capacity of wireless links, packet losses may happen frequently. Moreover, the broadcast nature of wireless medium introduces the hidden terminal and exposed terminal problems. Additionally, mobile nodes have restricted power, computing and bandwidth resources and require effective routing schemes. The highly dynamic nature of MANET coupled with limited bandwidth and battery power imposes severe restrictions on routing protocols especially on achieving the routing stability. Due to all these constraints, designing of a routing protocol is still a challenging task for researchers. In this paper an attempt has been made to evaluate and compare the impact of different mobility models on the performance of three most commonly used on-demands routing protocols named as AODV, DSR and LAR. The performance of these routing protocols has been simulated using QualNet 5.0 simulator.

Keywords: MANET, Ad hoc networks, Routing Protocols, Network simulation, Mobility models

1. INTRODUCTION

A Mobile ad hoc network [1][2] is a group of wireless mobile computers (or nodes); in which nodes collaborate by forwarding packets for each other to allow them to communicate outside range of direct wireless transmission. Ad hoc networks require no centralized administration or fixed network infrastructure such as base stations or access points.

Traditional table-driven routing approach was used in which tables are created at each node and when a node wishes to communicate with a distant node that is not within its vicinity the node consults its routing table and routes the packet accordingly. The protocols based on the above mechanism such as DSDV and CGSR consumes large memory and significant control overhead is consumed in maintaining tables which can be bearable in wired network but in case of wireless networks like MANETs this approach is not feasible due to above mentioned constraints.

The second method of routing is on demand. These protocols start to set up routes on-demand. The routing protocol will try to establish such a route, whenever any node wants to initiate communication with another node to which it has no route. This kind of protocols is usually based on flooding the network with Route Request (RREQ) and Route reply (RREP) messages. By the help of Route request message the route is discovered from source to target node; and as the target node gets a RREQ message it send RREP message for the confirmation that the route has been established. The three prominent on-demand routing protocols are AODV [5] [6] and DSR [7] [8] and LAR.

In order to thoroughly simulate a protocol for an ad hoc network, it is imperative to use a mobility model that accurately represents the mobile nodes (MNs) that will eventually utilize the given protocol. Currently, there are two types of mobility models used in the simulation of networks: traces and synthetic models. Traces provide accurate information, especially when they involve a large number of participants and an appropriately long observation period. New network environments are not easily modeled if traces have not yet been created. In this situation it is necessary to use synthetic models. Synthetic model attempt to realistically represent the behavior of mobile nodes without traces. Synthetic models can be: group mobility model or entity mobility models. This paper considers three routing protocols and compares them using QualNet 5.0 simulator [14] on different parameters. The rest of the paper is organized as follows: Section 2 describes literature survey of AODV, DSR and LAR routing protocols. Section 3 discusses the results, comparisons and simulation. Finally, we present the conclusion.

2. LITERATURE SURVEY

Ad Hoc On-Demand Distance-Vector Routing Protocol (AODV) is a reactive unicast routing protocol for mobile ad hoc networks. As a reactive routing protocol, AODV only needs to maintain the routing information about the active paths. In AODV, routing information is maintained in routing tables at nodes. Every mobile node keeps a next-hop routing table, which contains the destinations to which it currently has a route. A routing table entry expires if it has not been used or reactivated for a pre-specified expiration time. Moreover, AODV adopts the destination sequence number technique used by DSDV in an on-demand way.

In AODV, when a source node wants to send packets to the destination but no route is available, it initiates a route discovery operation. In the route discovery operation, the source broadcasts route request (RREQ) packets. A RREQ includes addresses of the source and the destination, the broadcast ID, which is used as its identifier, the last seen sequence number of the destination as well as the source node's sequence number. Sequence numbers are important to ensure loop-free and up-to-date routes. To reduce the flooding overhead, a node discards RREQs that it has seen before and the expanding ring search algorithm is used in route discovery operation. The RREQ starts with a small TTL (Time-To-Live) value. If the destination is not found, the TTL is increased in following RREQs. In AODV, each node maintains a cache to keep track of RREQs it has received. The cache also stores the path back to each RREQ originator. When the destination or a node that has a route to the destination receives the RREQ, it checks the destination sequence numbers it currently knows and the one specified in the RREQ. To guarantee the freshness of the routing information, a route reply (RREP) packet is created and forwarded back to the source only if the destination sequence number is equal to or greater than the one specified in RREQ. AODV uses only symmetric links and a RREP follows the reverse path of the respective RREP. Upon receiving the RREP packet, each intermediate node along the route updates its next-hop table entries with respect to the destination node. The redundant RREP packets or RREP packets with lower destination sequence number will be dropped.

In AODV, a node uses hello messages to notify its existence to its neighbors. Therefore, the link status to the next hop in an active route can be monitored. When a node discovers a link disconnection, it broadcasts a route error (RERR) packet to its neighbors, which in turn propagates the RERR packet towards nodes whose routes may be affected by the disconnected link. Then, the affected source can re-initiate a route discovery operation if the route is still needed.

Dynamic Source Routing Protocol (DSR) was proposed for routing in MANET by Broch, Johnson and Maltz [7]. In DSR, each mobile node is required to maintain a route cache that contains the source routes of which the mobile node is aware. The node updates entries in the route cache as and when it learns about new routes. The protocol consists of two phases:

The Route Discovery process initiates whenever the source node wants to send a packet to some destination. Firstly, the node consults its route cache to determine whether it already has a route to the destination or not. If it finds that an unexpired route to the destination exists, it makes use of this route to send the packet. On the other hand, if the node does not have such a route, it initiates route discovery by broadcasting a Route Request (RREQ) packet. The Route Request (RREQ) packet contains the address of the source and the destination, and a unique identification number as well. Each intermediate node that receives the packet checks whether it knows of a route to the destination. If it does not, it appends its own address to the route record of the packet and forwards the packet along to its neighbors. However, in case it finds a route, a Route Reply (RREP) packet containing the optimal path is transmitted back to the source node through the shortest route. To limit the number of route requests propagated, a node processes the Route Request (RREQ) packet only if it has not already seen the packet and its address is not present in the route record of the packet. A Route Reply (RREP) is generated when either the destination or an intermediate node with current information about the destination receives the Route Request (RREQ) packet. As the Route Request (RREQ) packet propagates through the network, the route record is formed. If the Route Reply (RREP) is generated by the destination then it places the route record from Route Request (RREQ) packet into the Route Reply (RREP) packet. The Route Reply (RREP) packet is sent by the destination itself.

In Route maintenance Phase, when a node encounters a fatal transmission problem at its data link layer, it generates a Route Error (RERR) packet. When a node receives a route error packet, it removes the hop in error from its route cache. All routes that contain the hop in error are truncated at that point. Acknowledgement (ACK) packets are used to verify the correct operation of the route links. This also includes passive acknowledgements in which a node hears the next hop forwarding the packet along the route.

The Location Aided Routing (LAR) is a reactive unicast routing scheme. LAR exploits position information and is proposed to improve the efficiency of the route discovery procedure by limiting the scope of route request flooding.

In LAR, a source node estimates the current location range of the destination based on information of the last reported location and mobility pattern of the destination. In LAR, an expected zone is defined as a region that is expected to hold the current location of the destination node. During route discovery procedure, the route request flooding is limited to a request zone, which contains the expected zone and location of the sender node. The source node calculates the expected zone and defines a request zone in request packets, and then initiates a route discovery. Receiving the route request, a node forwards the request if it falls inside the request zone; otherwise it discards the request. When the destination receives the request, it replies with a route reply that contains its current location, time and average speed. The size of a request zone can be adjusted according to the mobility pattern of the destination. When speed of the destination is low, the request zone is small; and when it moves fast, the request zone is large.

3. RESULTS AND SIMULATION

Various researchers have evaluated the performance of on demand routing protocols [10][11][12][13] on different simulators such as NS2,MATLAB but in our case we used QualNet 5.0 simulator[14] as it is a network modelling software that predicts performance of networks through simulation and emulation.For the purpose of simulation different scenarios were created for different number of nodes (15, 20, 25 and 30). The following parameters were configured as shown in Table 1.

Table 1. Configured Parameters

Parameter	Description
Size of Region	1500*1500
Shape of Region	Square
Mobility Model Used	File, RWP, Group Mobility
No. of Nodes Deployed	40
Battery Model	Linear model
Placement of Nodes	Random
No. Of Iterations	25
Energy model	Mica Motes
Antenna	Omni Direction
Total Bytes Sent	12288
Total Packet Sent	24
Throughput	4274

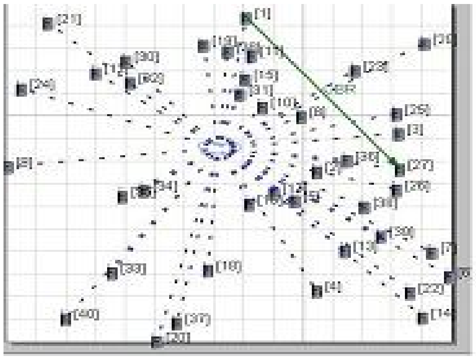


Figure 1. A Scenario for AODV, DSR and LAR routing protocols (on 40 Nodes)

In Figure 1, a scenario with 40 nodes is shown. The nodes were randomly distributed in 1500 X 1500 unit area. The node1 (Source) and the nodes 3,4,5,7,8,9,11,13,15,16,17,19, 21,22,23,25,27,29,31,32,33,35,37,38,39 (Destination) were connected and 1kb data was transmitted. The simulation was run for 30 seconds. The routing protocols taken were AODV, DSR, LAR and a comparison of the following parameters have been done.

3.1 Average Jitter

In case of AODV, avg. jitter is more when RWP mobility model is used, but it works well in group and file mobility model in compare with LAR, but DSR works best as shown in Figure 2.

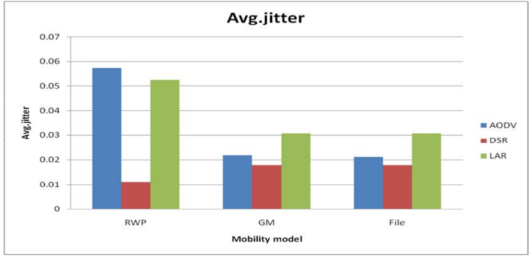


Figure 2. Average Jitter in AODV, DSR and LAR

3.2 First packet received

In case of DSR, result is same and best in all 3 mobility model in comparison to AODV, LAR as shown in Figure 3.

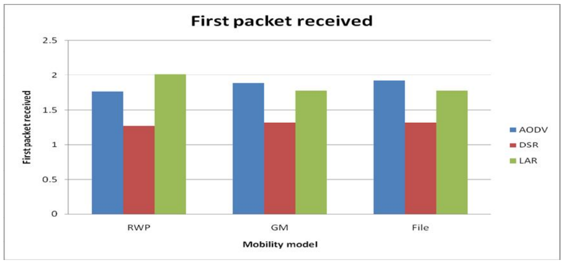


Figure 3. First packet received in AODV, DSR and LAR

3.3 Total packet received

In case of DSR, total packets received are more in comparison to AODV and LAR, as shown in Figure 4.

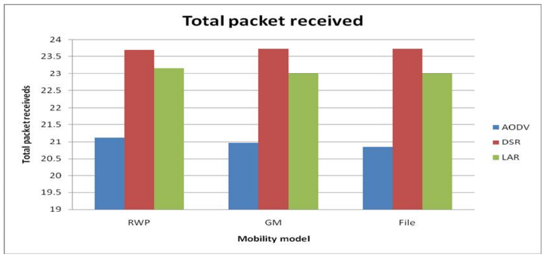


Figure 4. Total Packets received in AODV, DSR and LAR

3.4 Last packet received

In case of LAR, last packet receives faster in comparison to AODV and DSR, as shown in Figure 5.

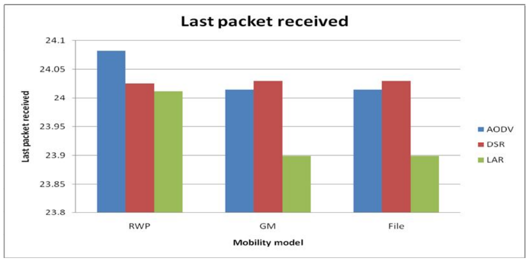


Figure 5. Last Packet received in AODV, DSR and LAR

3.5 Throughput

In case of DSR, numbers of hop counts are very high which indicates that congestion will be quite more in DSR in comparison to AODV, as shown in Figure 6.

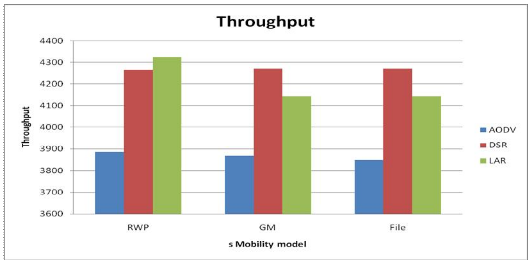


Figure 6. Throughput in AODV, DSR and LAR

From the above graphs which are generated on different parameters, we can see the comparison of AODV, DSR and LAR routing protocols (Table 2).

Table 2. Comparison of AODV, DSR and LAR Routing Protocols (On 40-Nodes Placement)

Parameter	AODV RWP N	DSR RWP	LAR RWP	AODV GM	DSR GM	LAR GM	AODV File	DSR File	LAR File
Average jitter	Very High	Low	High	Low	Very Low	High	Low	Very Low	High
First packet received	High	Low	Very High	Very High	Low	High	Very High	Low	High
Total packet received	Less	Very High	High	Less	Very High	High	Less	Very High	High
Last packet received	More	Less	Very Less	Less	More	Very Less	Less	More	Very Less
Throughput	Low	High	Very high	Low	Very High	High	Very Less	Very High	High

4. CONCLUSION

In this paper, the comparison of routing protocols AODV, DSR and LAR has been presented after their simulation on the QualNet 5.0 simulator. The following conclusions were drawn:

- The average jitter (uneven delay) will be more in case of LAR, but it is very less in DSR.AODV shows higher jitter in case of random waypoint mobility model in comparison to LAR.
- The first packet received earliest in DSR in comparison to AODV and LAR.
- The total packet received is highest in DSR in comparison to AODV and LAR.LAR results better than AODV.
- The last packet received earlier in LAR in comparison to AODV and DSR.
- The throughput is more in DSR in comparison to AODV and LAR.LAR results better than AODV.

5. REFERENCES

[1] Joseph Macker and Scott Corson, “Mobile Ad Hoc networks(MANET)”,http://www.ietf.org/proceedings/01 dec/183.htm, December 2001.

[2] M. Abolhasan, T. Wysocki and E. Dutkiewicz, “A review of Routing Protocols for Mobile Ad Hoc Networks”, Ad Hoc Networks, vol.2, issue1, pp.122, Jan.2004.

[3] Amit Goel and A. K. Sharma, “A Comparative Study of Unicast Routing Protocols for Mobile Ad Hoc Network”, Proc. of National Conference on Communication, Information and Telemetric- An Indian Scenario (CITEL-2005), KCT, Coimbatore, Mar. 2005.

[4] C. Sivaram Murthy and B. S. Manoj, “Ad Hoc Wireless Networks”, Architecture and Protocols, Pearson Education, Fourth Impression, 2009.

[5] C. Perkins and E. M. Royer, “Ad Hoc on Demand Distance Vector (AODV) routing”, In Proceeding of 2nd IEEE workshop on Mobile computing systems and Applications, pp. 90-100, February 1999.

[6] C. E. Perkins, E. M. Royer, S. R. Das, “Ad Hoc On demand Distance Vector (AODV) Routing”, IETF MANET Internetdraft (2003).

[7] Josh Broch, D. B. Johnson, D. A. Maltz, “The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks”, Internet Draft, IETF MANET Working Group, March 2nd 2001.

[8] Johnson, D. A. Maltz, and Y.-C. Hu, “The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)", IETF Draft, April 2003, work in progress. [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-09.txt .

[9] Z. Alexander, “Performance Evaluation of AODV Routing Protocol: Real-Life Measurements”, SCC, June 2003.

[10] J. Macker and S. Corson, RFC2501, “Mobile Ad Hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations”, IETF 1999.

[11] M. A. Bhagyaveni and S. Shanmugavel, “Performance of Ad Hoc Network Routing Protocols with Environment Awareness”, Information Technology Journal 4 (1): 1-5, 2005.

[12] Qian Feng, Zhongmin Cai, Jin Yang, Xunchao Hu, “A Performance Comparison of Ad Hoc Network Protocols”, Second International Workshop on Computer Science and Engineering, 2009 IEEE.

[13] Nor Surayati Mohamad Usop, Azizol Abdullah and Ahmad Faisa Amri Abidin, “Performance Evaluation of AODV, DSDV and DSR Routing Protocol in Grid Environment”, IJCSNS International Journal of Computer Science and Network Security, Vol.9 No.7, July 2009.

[14] The Network Simulator QualNet 5.0, [Online], Available http://www.scalable-network.com/products/qualnet/