International Journal of Computer Applications Technology and Research
Volume 4– Issue 8, 584 - 591, 2015, ISSN: 2319–8656

# Enhancing Data Staging as a Mechanism for Fast Data Access

Reagan Muriithi Gatimu
School of Computer Science and Information Technology
Jomo Kenyatta University of Agriculture and Technology
(JKUAT)
Nairobi, Kenya

Wilson Cheruiyot
School of Computer Science and Information Technology
Jomo Kenyatta University of Agriculture and Technology
(JKUAT)
Nairobi, Kenya

Michael Kimwele
School of Computer Science and Information Technology
Jomo Kenyatta University of Agriculture and Technology
(JKUAT)
Nairobi, Kenya

**Abstract**: Most organizations rely on data in their daily transactions and operations. This data is retrieved from different source systems in a distributed network hence it comes in varying data types and formats. The source data is prepared and cleaned by subjecting it to algorithms and functions before transferring it to the target systems which takes more time. Moreover, there is pressure from data users within the data warehouse for data to be availed quickly for them to make appropriate decisions and forecasts. This has not been the case due to immense data explosion in millions of transactions resulting from business processes of the organizations. The current legacy systems cannot handle large data levels due to processing capabilities and customizations. This approach has failed because there lacks clear procedures to decide which data to collect or exempt. It is with this concern that performance degradation should be addressed because organizations invest a lot of resources to establish a functioning data warehouse. Data staging is a technological innovation within data warehouses where data manipulations are carried out before transfer to target systems. It carries out data integration by harmonizing the staging functions, cleansing, verification, and archiving source data. Deterministic Prioritization Approach will be employed to enhance data staging, and to clearly prove this change Experiment design is needed to test scenarios in the study. Previous studies in this field have mainly focused in the data warehouses processes as a whole but less to the specifics of data staging area.

**Keywords**: Data Staging; Source System; Deterministic Prioritization; Data Warehouse

## 1. INTRODUCTION

The growing number of business transactions in any enterprise is directly proportional to growth of data size. This data comes from variant source systems and applications and needs to be organized in a workable state so that it remains relevant and meaningful to the users. Technological development has led to the rise of Data Warehouse (DW). (Inmon, 2002) defines a data warehouse as “collection of integrated, subject-oriented databases designated to support the decision making process”. Both (Kimball and Inmon, 2002) agree that a DW has to be integrated, subject-oriented, nonvolatile and time variant. This concept of time-variance is so crucial and ultimate concern and sets the basis for this research since it focuses on improved data access. The foundations of a DW as explained by (Zineb, Esteban, Jose-Norberto, Juan, 2011) encompass integration of multiple different data sources. This allows the provision of complete and correct view of the enterprise operational data which is synthesized into a set of strategic indicators and measures that the users of the data can associate with.

DW has business intelligence implemented in three major processes used to prepare data to match user’s needs. They are commonly referred to as ETL processes namely; Extraction, Transformation and Loading. Extraction process retrieves data as is from source systems before subjecting it to any manipulations. Transformation process also referred to as transportation phase is the operational base and the most intriguing of all. Business rules and functions are some of the operations applied to the extracted data. Loading process involves moving the desired data as determined by the users to the DW. It’s important to note that the discussed flow of data is not as simple and smooth as it sounds and this is as a result of impeding performance issues raised by the following observed bottlenecks.

(El-Wessimy et al, 2013) shows the relevance of DW in decision making in today’s environment. “The best decisions are made when all the relevant data is taken into consideration. Today, the biggest challenge in any organization is to achieve better performance with least cost, and to make better decisions than competitors. That is why data warehouses are widely used within the largest and most complex businesses in the world.”

In this paper, SQL Server Integration Services tool is used to experimentally show the impact of prioritizing the data from sources as per the confidence levels and the distinctiveness of the data. The units of measure also focus on general performance of the whole ETL process after enhancement of the data staging area.

## 2. RELATED WORK

### 2.1 Extraction Stage

This is the initial stage of data migration to a data warehouse. (Kimball et al., 1998) informs that the extraction process consists of two phases, initial extraction, and changed data extraction. In the initial extraction, data from the different operational sources to be loaded into the DW is captured for the first time. This process is done only one time after building the DW to populate it with a huge amount of data from source systems. The next phase involves incremental extraction also referred to as changed data capture (CDC).

(Stephen, 2013) informs that “The staging tables usually get populated by some outside source, by either pulling or pushing the data from the source systems. This process is usually an insert only process and therefore does not rely on statistics for its successful execution.”

## 2.2 Transformation Stage

Once the data is extracted to the staging area, there are numerous potential transformations, such as cleansing the data (correcting misspellings, resolving domain conflicts, dealing with missing elements, or parsing into standard formats), combining data from multiple sources, reduplicating data, and assigning warehouse keys. These transformations are all precursors to loading the data into the data warehouse presentation area. Unfortunately, there is still considerable industry consternation about whether the data that supports or results from this process should be instantiated in physical normalized structures prior to loading into the presentation area for querying and reporting.

(Erhard and Hong, 2000) elaborate on activities within transformation phase towards clean data. These include data analysis that focus on meta-data and due to fewer integrity rules it cannot guarantee sufficient data quality of a source. Two approaches have been put across to assist in data analysis i.e. data profiling and data mining. Data profiling focuses on the instance analysis of individual attributes. It derives information such as the data type, length, value range, discrete values and their frequency, variance, uniqueness, occurrence of null values, typical string pattern providing an exact view of various quality aspects of the attribute. Data mining helps discover specific data patterns in large data sets, e.g., relationships holding between several attributes.

## 2.3 Loading Stage

Loading in the data warehouse environment usually takes the form of presenting the quality-assured dimensional tables to the bulk loading facilities of each data mart. The target data mart must then index the newly arrived data for query performance. When each data mart has been freshly loaded, indexed, supplied with appropriate aggregates, and further quality assured, the user community is notified that the new data has been published.

When it comes to moving data to DW (Stephen, 2013) informs “The biggest question for the staging area is – how do we keep the statistics up-to-date such that the statistics for a particular daily load are always available and reasonably accurate. This is actually more difficult than it sounds. If the partitions would only be analyzed in the first quarter of the month each night, going to every other night and eventually each week because of the 10% stale setting. This obviously leaves us with a problem…. In order to have the statistics available for the latest day which is loaded, the statistics would have to be gathered after the staging tables have been loaded but before the ETL process starts.

## 2.4 Data Staging

Data staging emerges as a new technological development with an attempt to handle the low performance issues noted above. Its location within the ETL process differs as (Kimball and Ross, 2002) state that data staging is available in the extraction and transformation phases of ETL framework. In some current systems a data stage exists as a location that interconnects Online Transaction Processing systems (OLTPs) to the Online Analytical processing systems (OLAPs).

## 2.5 Reasons for Data Staging Enhancement

Although data staging is not a new technology since it has been researched before, the focus has been shifted to designs and development of data staging frameworks. Little attention has been given to its operability and its significant role in speeding the ETL process.

In a production environment especially a busy organization that deals with large transactions in its daily operations, data flow to DW and storage repositories becomes an issue. Some may not be experiencing the performance problems initially but when their data levels grow they start getting intermittent performance. This should be handled early to have a maintained work path. It still remains a challenge on selection algorithms that can pre-determine the data needed at the target system. The proposed solution in data staging which forms the enhancement is to work on the pre-determining and prioritization mechanisms on the data to load.

(Aksoy et al, 2001) introduces a more workable approach to data staging concerns. They based their work on broadcast scheduling and data staging. According to them the key design considered for development of large scale on-demand broadcast server was the scheduling algorithm selection useful for selecting of items to be broadcast. However, this solution is based on assumptions that data will be available before hand which is not true due to its dynamism.

# 3. PROPOSED APPROACH IN DATA STAGING

Considering the improvements already observed from great works of other authors it’s important to appreciate their efforts in finding gaps in existing systems. Relating to the recurring problems within the data staging, the researcher identifies the following dependent and independent variables that assist in choosing deterministic prioritization approach as a probable solution.

## 3.1 Dependent and Independent Variables

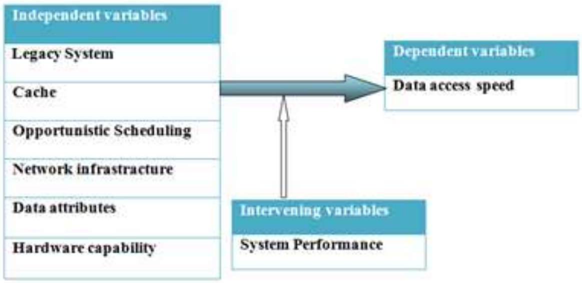


Figure 1. Dependent and Independent variables

Legacy systems use stacking, where a job is pushed into the pool of tasks and popped out of the stack sequentially when its time of execution arrives according to the queue structure. The order of execution is not highly dependent on pre-arranged structure but on the procedural mode which degrades performance.

The cache offer pre-fetching capability where data storage occurs temporarily for the most used data. The scheduled procedure looks firstly in the cache memory before checking on secondary storage locations to minimize the search time. The limitation lies on the cache size and amount of data to be maintained in cache at a time.

To achieve concurrent operation there needs to be selective algorithms to decide the priority of jobs from the source systems to the data staging area. With Opportunistic scheduling there is high probability of improving speed of data retrieval and access.

Remote connection to source systems affects the speed of retrieval and query execution is delayed by the time-lapse for distributed systems. This impact on the nature of ordering results from query execution and thus optimization should be introduced to work with stored procedures and cache facility.

The data characteristics are defined by type and formats since it comes from disparate systems. Destination requirements must be matched before data is moved to the target systems. Poor data manipulation functions result in longer time processing the data slowing down the systems. The functions for manipulating flat files are different from the ones for relational tables and databases due to underlying data formats.

## 3.2 Deterministic Prioritization (DP)

After thorough considerations of the above variables and the research gaps identified, Deterministic Prioritization approach is put forward as a solution to the data access problem. The implementation of deterministic prioritization in the data staging area expounds the relationship of the other ETL processes. This approach will tend to manipulate data immediately it is collected and availed to the staging area. Less activity is experienced in the extraction phase but the actual data work area is within the staging area. With this approach appropriate data selection is coordinated to filter out unwanted "dirty" data and assigning priority to the important "clean" data that is moved to data warehouse or data marts. The fact remains that previous activities within staging area are important and hence the approach aims to improve on the order of execution to avoid redundancy and repetition of tasks. This concept is illustrated in the following diagram that shows the relationship among the ETL processes of a data warehouse by applying the DP rules.

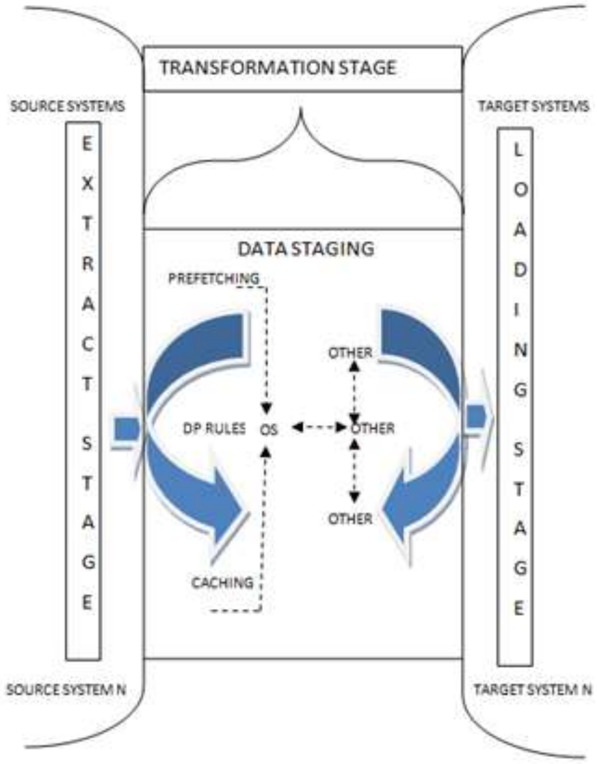


Figure 2. Proposed Conceptual Framework

This implementation is in two sections i.e. deterministic section and prioritization section as discussed below.

## 3.3 Deterministic Section

According to Macmillan dictionary, the term deterministic means "using or believing in the idea that everything is caused by another event or action and so you are not free to choose what you do". The deterministic approach is relying on data selections based on the confidence levels or behavior of data accesses in relation to previous selections. The key characteristic of determinism is that, the output remains the same for the same inputs. When applied to data staging, the selection of column values from a table is based on the distinctiveness of the data stored and frequency of previous accesses. The most distinct values (fewer NULLS) the higher the confidence levels to be loaded to the next stage in ETL. The data columns with high nullity are negated from the selection hence reducing the amount of load to the next stage.

## 3.4 Prioritization Section

According to Macmillan dictionary, the term prioritize means "to decide in what order you should do things, based on how important or urgent they are".

The main goal of implementing prioritization is to reduce the amount of active data being manipulated at a time by focusing on the minimum and meaningful details. Once the data to be loaded has been determined, then the order at which processing occurs is vital to reduce delays in data handling. The data selection from the raw tables needs to be prioritized by altering the query execution plan to give more priority on the data columns with high confidence levels. To implement this clustered indexes are introduced based on these distinct columns. These indexes alter the normal execution plan for the queries hence improving performance and efficiency. This execution plan is explained by (Grant, 2009) that "an execution plan is the result of query optimizer's attempt to calculate the most efficient way to implement the request represented by the T-SQL query..."

The query executer has an engine that optimizes query execution on its own and by altering the data selection time is reduced. Prioritization affects the logical aspects enforced by the business rules to maintain the dimensional model and giving way for faster way of retrieving data.

## 3.5 Included Improvements

Creating a new stable staging framework that is freely available to everyone and run across different hardware platforms (cross-platform) and supporting concurrent processing. This will magnify the core benefits of having intelligent data warehouses that are supportive to the top-level management systems majorly due to development of staging area.

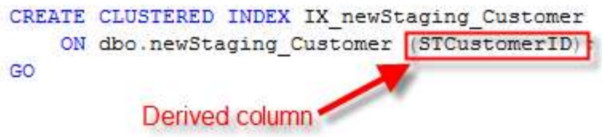
(Stephen, 2013) elaborates their approach in Oracle environment." Most ETL applications use a staging area to stage source system data before loading it into the warehouse or marts. When implemented within an oracle environment a partitioning strategy is usually employed such that data that is not required any longer can be removed from the tables with the minimum amount of effort."

The proposed framework will have forecasting and prioritization mechanisms to decide which data is necessary before transfer begins hence saving on network services and bandwidth.

Current systems such as HANA databases have high processing capability which is meaningless if there is no proper scheduling of resources. This can result to lots of losses of resources not being manned properly. Hardware is static while data is dynamic and at some point the available hardware would not be sufficient to handle the data affecting on performance. Eventually, Scheduling plays a vital role in the performance implications of any system. It represents the effect being sought is measurable to make comparison.

## 3.6 Prioritization by indexing specific columns

The newly distinct derived columns that were added to each staging table (external columns; hence do not affect the data from the source in any way to ensure consistency and integrity), are used to create indexes as shown below.



```
CREATE CLUSTERED INDEX IX_newStaging_Customer
    ON dbo.newStaging_Customer (STCustomerID)
GO
```

Figure 3. Index Creation Sample

The created index named "IX_newStaging_Customer" is prioritizing the derived unique column named "STCustomerID" on the staging table named "newStaging_Customer". The created user defined index hints the order of execution of the Data Definition Language (DDL) queries submitted to the server hence overriding the server's query execution plan. The created index is also not affected in future in case of recreations of the source data from extraction stage since its an external derived column. Prioritization by distinct columns enhances the efficiency and performance of the query execution plan by the server during query search. This results in optimized selection costs while maintaining the quality of data to the data warehouse.  The measure of improved efficiency is shown below for a selection query.

Figure 4. New Execution plan for Customer Table

## 3.7 Test scenario preparation in SSIS tool

The ETL processes of a data warehouse are demonstrated using the SSIS tool. The first scenario setup is for the current situation before enhancement and the second scenario setup is for the new situation after enhancement. The following is a demonstration of the scenario used in the experiments in run mode.

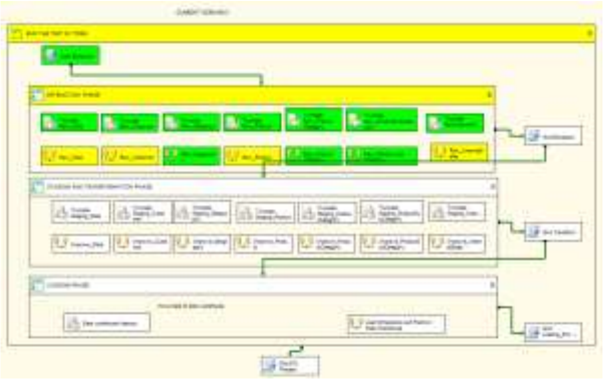


Figure 5. Test Scenario Setup.

## 4. RESULTS

The experiments are performed and collecting results of time variables in a Ms Excel file. This file is generated automatically from the scripts written in Visual Studio 2008 and C# programming language when the scenario is run. The experiment is run for fifteen times and for each cycle it records the time change for the different variables to the file. Finally, the comparison is made based on these results from both scenarios to note the impact of the change introduced as shown below.

**Table 1. Before Enhancement ETL Processes results**

| TestRunNumber | prevExtraction Time | prevStagingTime | prevLoading Time |
|---|---|---|---|
| 1 | 63368.28 | 68717.21 | 92851.52 |
| 2 | 97940.12 | 79815.68 | 132989.3 |
| 3 | 52449.29 | 45874.92 | 62384.48 |
| 4 | 54358.59 | 87338.59 | 49553.42 |
| 5 | 49600.26 | 50350.14 | 45086.72 |
| 6 | 48285.48 | 54800.62 | 45312.21 |
| 7 | 52569.22 | 48670 | 48858.79 |
| 8 | 53581.86 | 51044.78 | 66801.6 |
| 9 | 88992.69 | 96697.22 | 60973.71 |
| 10 | 72884.88 | 56306.1 | 52252.09 |
| 11 | 58839.45 | 54556.57 | 59913 |
| 12 | 69740.99 | 58111.31 | 53562.38 |
| 13 | 59401.77 | 50840.68 | 51099.61 |
| 14 | 65250.34 | 55357.5 | 52560.51 |
| 15 | 56115.72 | 71712.82 | 58035.14 |

**Table2. After Enhancement ETL Processes results**

| TestRunNumber | newExtractionTime | newStagingTime | newLoadingTime | newTotalTime |
|---|---|---|---|---|
| 1 | 62261.37 | 50691.27 | 4824.541 | 117777.2 |
| 2 | 59639.95 | 53591.02 | 4299.265 | 117530.2 |
| 3 | 55766.46 | 54468.57 | 3652.78 | 113887.8 |
| 4 | 54012.87 | 53544.93 | 3552.135 | 111109.9 |
| 5 | 61811.67 | 54836.81 | 3340.999 | 119989.5 |
| 6 | 59963.43 | 57082.69 | 3374.887 | 120421 |
| 7 | 63116.5 | 69659.48 | 4147.719 | 136923.7 |
| 8 | 61813.26 | 61603.78 | 3455.976 | 126873 |
| 9 | 69365.94 | 54521.06 | 3623.27 | 127510.3 |
| 10 | 56561.51 | 52100.49 | 5690.811 | 114352.8 |
| 11 | 54106.91 | 59241.29 | 3701.506 | 117049.7 |
| 12 | 59291.2 | 86575.51 | 3493.712 | 213390.8 |
| 13 | 66104.49 | 89692.62 | 4130.86 | 159928 |
| 14 | 72926.03 | 56987.29 | 5984.634 | 135898 |
| 15 | 57962.95 | 50812.86 | 3926.081 | 112701.9 |

## 5. DISCUSSION

The following discussion is based on the comparison of the results obtained from the tests. Each of the ETL stage is compared separately for the two situations and graphically shown in the following figures to have a clear distinction of the two situations.

### 5.1 Comparison of Extraction stage

The following is an illustration of the individual comparison per stage of the ETL processes to bring out a clear view of the improvement made. Explanation for each comparison follows for every illustration. The negative sign indicates that it is in the reducing direction thus showing the enhancement has taken place by reducing the particular running time per stage.

**Table 3. Extraction stage results analysis**

| Test | prevExtractionTime | newExtractionTime |
|---|---|---|
| Average | 62891.92931 | 60980.30203 |
| Change | | -1911.627273 |
| Change % | | -3.039543061 |

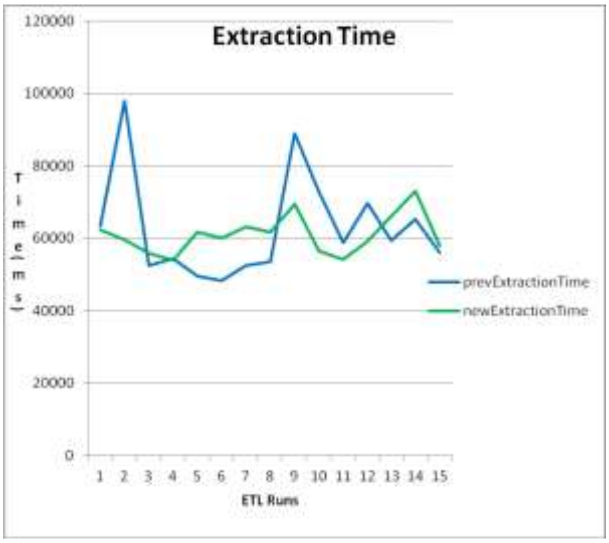


Figure 6. Extraction stage comparison

The above illustrations show the time taken for extraction using Deterministic Prioritization approach has reduced by 3.04% compared to the previous extraction time.

### 5.2 Comparison of Staging stage

**Table 4. Staging stage results analysis**

| Test | prevStagingTime | newStagingTime |
|---|---|---|
| Average | 62012.94289 | 60360.64505 |
| Change | | -1652.29784 |
| Change % | | -2.664440297 |

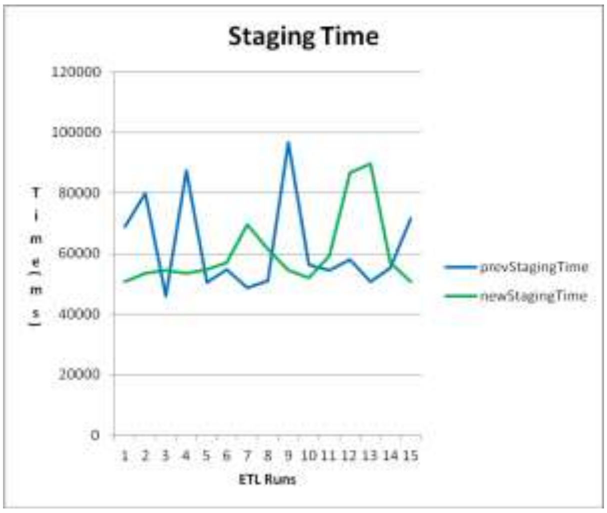


Figure 7. Staging Stage Comparison

The above illustrations show the time taken in staging area using Deterministic Prioritization approach has reduced by 2.66% compared to the previous extraction time.

## 5.3 Comparison of Loading stage

**Table 5. Loading stage results analysis**

| Test | prevLoadingTime | newLoadingTime |
|---|---|---|
| **Average** | 62148.96424 | 4079.94504 |
| **Change** | | -58069.0192 |
| **Change %** | | -93.43521635 |

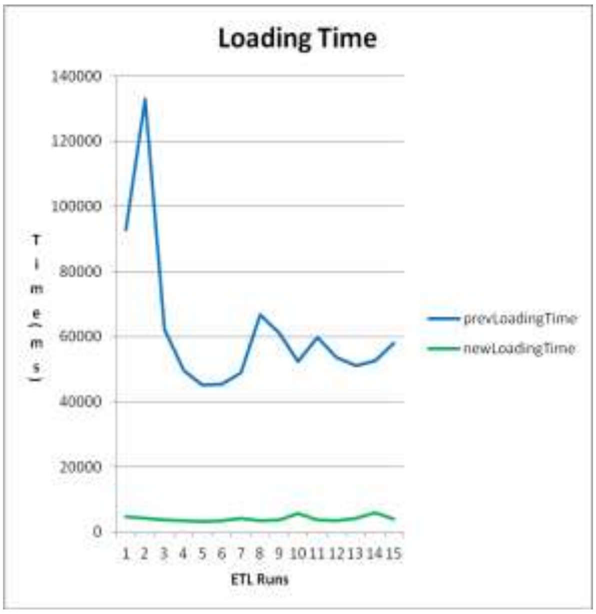


Figure 8. Loading Stage Comparison

The above illustrations show the time taken in loading stage using Deterministic Prioritization approach has immensely reduced by 93.44% compared to the previous loading time. This is a great change and is attributed to the fact that the number of operations talking place at this stage being minimal and less traffic to the destination targets of the data i.e. the data warehouses. However with increased traffic of data especially over the network, the loading time percentage change may further reduce. The loading process has benefited a lot from changes in previous stages.

## 6. SUMMARY OF FINDINGS

There has been a significant change for every stage of the ETL processes after implementation of deterministic prioritization approach. Moreover, the subsequent impact of this improvement has been depicted at the loading stage where the rate of transfer improved by a large margin of 93.44%. This shows that any change occurring in the staging area will definitely affect the entire process flow. However, if more time is taken in proper design of the data staging area, then there are high expectations for further improvements on performance beyond the ones achieved by the researcher.

Based on the limitation of resources and the specifications of the workstation used to run the tests, the performance levels are considered acceptable for an organization with fast growing data sets. However, the setup is limited by the resources at run time but with increased disk space as provided by the owners as data size grows, then the total data access time in the data warehouse will be highly improved. The choice of data prioritization mechanism using indexes for columns is supported by (Cecilia and Mihai, 2011), where they state the use of Indexes on database queries improves the performance of the whole system. Clustered indexes perform better than nonclustered indexes when the expected returned records are many and should be set for the most unique column of a table. This proposition goes in line with the research with the use of clustered indexes in the staging area mainly due to the need to fetch and process large data.

(Costel G et al, 2014) did a research on query execution and optimization in the MSSQL Server and put across the missing of indexes as a contributor to low performance of query execution. They inform that when a table misses indexes, the search engine has to parse through the entire table step by step to find the searched value. The resources spent on this process are enormous and considerably increases time to execute the queries.

(Grant, F. 2012) explains about execution plan management done by query optimizer. The database relational engine performs logical reads within the cache memory while the storage engine performs physical reads directly from disk. Improvements are highly realized mostly for data manipulation language statements since the engine needs to parse the query for correctness. The SQL server generates statistics against the indexes and sends them to the optimizer to determine the execution plan.

(El-Wessimy et al, 2013). Similarly did an enhancement in the data warehouse staging area by using different techniques (FIFO, MC,RR time and record rotation) targeting the loading phase. The tests ran captured the time taken to transfer data in each stage of the ETL process and suggest the most suitable technique. They did a comparison amongst all techniques and noted that FIFO performed better for less data set while Record Limit Based Round Robin was best for large data sets. Their research supported further reduction of overall time taken to deliver data from source to destination. The uniqueness of this study is the ability to handle large data sets from the beginning as well as newer inclusions of data from the sources without any readjustments of the system structure setup.

## 7. CONCLUSION

The adoption of Deterministic Prioritization approach in the staging area has shown promising results and the users at the presentation level of the data warehouse are rest assured of fast data access and retrievals. They can timely make decisive conclusions and reports based on current data that is made available in a timely fashion similar to real time systems. They will also benefit to wide range of data availed since the ETL processes aim to denormalize the data comprehensively before its delivered to users. This forms an association that is deterministic in nature for future priority loads. The researcher was keen to avoid compromising data quality for high performance gains and this resulted in a more balanced system setup where the data still meet the qualitative assurance defined by the users' requirements.

## 8. RECOMENDATIONS AND FUTURE WORK

In view of the results and findings of the experiments undertaken in this research, the researcher recommends the incorporation of the deterministic prioritization approach in the design and development phase of the data staging frameworks. This is so because it is cross-platform to all database management systems that support the SQL language in the market today. The change gives room for further customization since it only happens at the design and before query execution. Notably for well formed queries the performance will be even better.

The data staging area is an area which has not been researched exhaustively and the impact of high resources should be considered as a next check on performance gain over cost. The setup experiments are carried based on same format of data sources and further studies should be carried out on variant data sources and using different staging framework other than the SSIS tool used here. The scenario in this test is demonstrated in a local workstation and it would be essential to note the performance levels in a distributed system with both sources and targets widely separated by networks.

## 9. REFERENCES

[1] Abbasi, H., Wolf, M., Eisenhauer, G., Klasky, S., Schwan, K., & Zheng, F. (2010). Datastager: scalable data staging services for petascale applications. *Cluster Computing*, *13*(3), 277-290.

[2] Akkaoui, Z. E., Munoz, E. Z. J.-N., and Trujillo .J. A. (2011). *Model-Driven Framework for ETL Process Development*. In Proceedings of the international workshop on Data Warehousing and OLAP. pp. 45–52 Glasgow, Scotland, UK.

[3] Aksoy, D., Franklin, M. J., & Zdonik, S. (2001). Data staging for on-demand broadcast. In *VLDB* (Vol. 1, pp. 571-580).

[4] B´ezivin, J. (2005). On the unification power of models.Software and System Modeling, 4(2):171–188

[5] Cecilia, C., Mihai, G. (2011). Increasing Database Performance using Indexes. *Database Systems Journal vol. II, no. 2/2011*. Economic Informatics Department, Academy of Economic Studies Bucharest, Romania.

[6] Costel, G.C., Marius, M. L., Valentina, L., Octavian, T. P.(2014). Query Optimization Techniques in Microsoft SQL Server. *Database Systems Journal vol. V, no. 2/2014*. University of Economic Studies, Bucharest, Romania.

[7] Da Silva, M.S.,Times, V.C., Kwakye, M.M. (2012). Journal of Information and Data Management.3 (3).

[8] Deterministic. (2009-2015). In Macmillan Dictionary. Macmillan Publishers Limited: Accessed from: www.macmillandictionary.com on 20th July 2015.

[9] Eckerson, W., & White, C. (2003). Evaluating ETL and data integration platforms. *Seattle: The DW Institute*.

[10] El-Wessimy, M., Mokhtar, H. M., & Hegazy, O. (2013). ENHANCEMENT TECHNIQUES FOR DATA WAREHOUSE STAGING AREA. *International Journal of Data Mining & Knowledge Management Process*, *3*(6).

[11] Erhard, R., and Hong, H.D. (2000). *Data Cleaning*: Problems and Current Approaches. Journal IEEE Data Eng. Bull.23 (4), 3-13.

[12] Grant, F.(2009).The Art of High Performance SQL Code: SQL Server Execution Plans. Simple-Talk Publishing. ISBN 978-1-906434-02-1

[13] Grant, F. (2012). SQL Server Execution Plans. Second Ed. ISBN: 978-1-906434-92-2. Simple Talk Publishing.

[14] El-Wessimy, M., Mokhtar, H. M., & Hegazy, O. (2013). ENHANCEMENT TECHNIQUES FOR DATA WAREHOUSE STAGING AREA. *International Journal of Data Mining & Knowledge Management Process*, *3*(6).

[15] Firestone, J. M. (1998). Dimensional modeling and ER modeling in the data warehouse. *White Paper No, Eight June*, *22*.

[16] Flinn, J., Sinnamohideen, S., Tolia, N., & Satyanarayanan, M. (2003). Data Staging on Untrusted Surrogates. In *FAST* (Vol. 3, pp. 15-28).

[17] Inmon, W. H. (2002). *Building the Data Warehouse*. Wiley.

[18] Inmon, W. H. (2005). *Building the data warehouse*. John wiley & sons.

[19] Kimball, R., & Caserta, J. (2004). *The data warehouse ETL toolkit*. John Wiley & Sons.

[20] Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (2008). The Data Warehouse Lifecycle Toolkit, 2nd ed. Practical Techniques for Building Data Warehouse and Business Intelligence Systems.

[21] Muller, P. A., Studer, P., Fondement, F., & Bézivin, J. (2005). Platform independent Web application modeling and development with Netsilon. *Software & Systems Modeling*, *4*(4), 424-442.

[22] Per-Åke, L., Cipri, C., Campbell, F., Eric, N. H., Mostafa, M., Michal, N., Vassilis, P., Susan, L. P., Srikumar, R., Remus, R., Mayukh, S.(2013).Enhancements to SQL Server Column Stores. ACM 978-1 -4503-2037-5/13/06. New York, USA.

[23] Prioritize. (2009-2015). In Macmillan Dictionary. Macmillan Publishers Limited: Accessed from: www.macmillandictionary.com on 20th July 2015.

[24] Ralph, K. and Margy, R. (2002). *The Data WarehouseToolkit*: The Complete Guide to Dimensional Modeling. Second Edition. Published by John Wiley and Sons, Inc. Canada.

[25] Russom, P. (2012). BI Experts: Big Data and Your Data Warehouse's Data Staging Area. *TDWI Best Practices Report, Fourth Quarter*. Retrieved from http://tdwi.org/articles/2012/07/10/big-data-staging-area.aspx

[26] SAP AG. (2002). Business Information Warehouse – Data Staging Retrieved from http://scn.sap.com/docs/DOC-8100.

[27] Stephen, B. (2013). Staging, Statistics & Common Sense: Oracle Statistics Maintenance

[28] Strategy in an ETL environment Retrieved from http://www.seethehippo.com/

[29] Vassiliadis, P. (2009). A survey of Extract–transform–Load technology. *International Journal of Data Warehousing and Mining (IJDWM)*, *5*(3), 1-27.

[30] Zineb, A., Esteban, Z., Jose-Norberto.M., Juan, T. (2011). *A Model-Driven Framework for ETL Process Development*. In DOLAP 11 Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP Pages 45-52. ACM New York, NY, USA. ISBN: 978-1-4503-0963-9

International Journal of Computer Applications Technology and Research
Volume 4– Issue 8, 592 - 598, 2015, ISSN: 2319–8656

# Evaluation of Iris Recognition System on Multiple Feature Extraction Algorithms and its Combinations

Ashwini M B
Department of Computer Science,
MMK & SDM Mahila Maha Vidyalaya, K M Puram, Mysore,
Karnataka, India

Mohammad Imran
Department of CCSIT,
King Faisal University
Al Ahsa, KSA

Fawaz Alsaade
Department of CCSIT,
King Faisal University
Al Ahsa, KSA

**Abstract:** Multi-algorithmic approach to enhancing the accuracy of iris recognition system is proposed and investigated. In this system, features are extracted from the iris using various feature extraction algorithms, namely LPQ, LBP, Gabor Filter, Haar, Db8 and Db16. Based on the experimental results, it is demonstrated that Mutli-algorithms Iris Recognition System is performing better than the unimodal system. The accuracy improvement offered by the proposed approach also showed that using more than two feature extraction algorithms in extracting the iris system might decrease the system performance. This is due to redundant features. The paper presents a detailed description of the experiments and provides an analysis of the performance of the proposed method.

**Keywods**: Biometrics, Iris, Unimodal, Multiple Algorithms, Feature Level Fusion, Performance of Algorithms

## 1. INTRODUCTION

Biometrics is the science of recognizing a person on the basis of physical or behavioral characteristics. Psychological characteristics are characteristics that are gathered or learned during time (e. g. signature, gait, typing dynamics, voice features etc.). Physical characteristics are characteristics that are genetically implied and possibly influenced by the environment (e. g. face, iris, retina, finger, vascular structure etc.). Iris biometric is employed in this work because of its distinctive and uniqueness in each person including identical twins and each eye. There are number of properties that enhance the suitability of the iris for use in automatic identification. One important feature is that the iris is inherently isolated and protected from the external environment. Moreover, being an internal organ of the eye, behind the cornea and the aqueous humor; it is almost impossible to surgically modify the iris without unacceptable risk of damage. Also, its physiological response to light provides a natural test against fake irises [11].

Iris provides many interlacing minute characteristics such as freckles, coronas, stripes, furrows, crypts and so on. It is essentially stable over a person’s life. Iris-based personal identification systems are noninvasive to their users. The technology of Iris recognition has some advantages and disadvantages which make it appropriate to use in some application and not to other applications[12]. Some of the advantages of iris biometrics are as follows: a) Iris is very accurate biometric. It has low false acceptance rate, which is important in security aspect. Therefore iris might be a good biometric for identification applications. b) The sensing of the biometric is without physical contact and it is convenient for the users because the iris pattern acquisition process uses unnoticeable and distant cameras. c)Iris recognition system has low cost training. d) Iris technology has received little negative press and so is more generally accepted biometric identifier[2].

There are various unimodal iris recognition systems exists in real world, unimodal system means rely on single source of information for verification/identification. However, Multibiometric system relies on the evidence presented by multiple sources of biometric information. Based on the nature of these sources, a Multibiometric system can be classified into one of the following six categories: multi-sensor, multi-algorithm, multi-instance, multi-sample, multimodal and hybrid systems [ ]. In some cases, invoking multiple feature extraction and/or matching algorithms on the same biometric data can result in improved matching performance. Multi-algorithm system consolidates the output of multiple feature extraction algorithms, or that of multiple matchers operating on the same feature set. These systems do not necessitate the deployment of new sensors and, hence, are cost-effective compared to other types of multibiometric systems. On the other hand, the introduction of new feature extraction and matching modules can increase the computational complexity of these systems [3].

The work presented in this paper aims to enhancing the security of iris recognition system using multi-algorithmic approach. This is achieved by fusing the data acquired at the feature level and applying the K-Nearest Neighbor classifier(K-NN). The rest of the paper is structured as follows. Section 2 presents related works. Section 3 introduces the proposed approach for multi-algorithm iris recognition system. Section 4 describes the experimental investigations, and the overall conclusions are presented in Section 5.

## 2. RELATED WORKS

Karen et al [1] states that, genetically identical irises have texture similarity that is not detected by iris biometrics. By performing experimental studies on left and right irises of the same person, it results that are as different as irises of unrelated people. Similarly, in terms of iris biometric matching, the eyes of identical twins are as different as irises of unrelated people. The experiments done based on texture feature extraction methods, this highlight the difference between automated biometric technology and human viewers. This work suggests that human examination of pairs of iris images for forensic purposes may be feasible and results of experiments suggest that development of different approaches to automated iris image analysis could be used to improve current iris biometric technology. Vijaya et al [4] states that generally algorithm is divided into four steps, such as. localization, Normalization, Feature Extraction and Matching. Doughman‘s approach is there with highest accuracy of 99.9% and Kaushik Rai‘s approach is also promising in state of the art, only thing that can be done is to decrease the computational time and number of features to obtain the same efficiency.

Libor et al [5] presents an automatic segmentation algorithm that remove iris region from an eye image and isolate eyelid, eyelash and reflection areas. The iris region was normalized into a rectangular block with constant dimensions to account for imaging inconsistencies. However, the phase data from 1D Log-Gabor filters was extracted and quantized to four levels to encode the unique pattern of the iris into a bit-wise biometric template. The Hamming distance was employed for classification of iris templates, and two templates were found to match if a test of statistical independence was failed. The system performed with perfect recognition on a set of 75 eye images; however, tests on another set of 624 images resulted in false accept and false reject rates of 0.005% and 0.238% respectively. Therefore, iris recognition is shown to be a reliable and accurate biometric technology. Somnath et al [6] present a novel and efficient approach to extract iris features and matching technique to compare iris features. This approach uses Daubechies wavelet transform with four coefficients. Daubechies wavelet transform is easy to compute and fast compared to the other methods on texture analysis. Further, Daubechies wavelet transform allows to keep the count of feature vectors into a significantly lesser numbers.

Fenghua et al [7] presents a multi-algorithmic fusion approach for iris recognition, which combines phase based algorithm and zero-crossing based algorithm. The two algorithms are fused at the matching score level using SVM fusion strategy. The experimental results on CASIA and UBIRIS iris database show that the proposed approach can improve the recognition performance compared with the individual recognition algorithm and SVM based fusion strategy can give the better performance than the traditional fusion strategies.

# 3. METHODS AND MATERIALS

In this section, we emphasis on different iris recognition methods which are explored in state of the art approaches. Brief details of different iris biometric technology methods as well as feature extraction algorithms are explored here.

## 3.1 Iris Recognition Methods

a. Phase-based method: It recognizes iris patterns based on phase information which is independent of image contrast and illumination. The first complete, commercially available phase-based iris recognition system was designed and patented in 1994 by J.Daugman [15].

b. Texture-analysis based method: Wildes proposed iris recognition based on texture analysis. High quality iris images were captured using silicon intensified target camera coupled with a standard frame grabber and resolution of 512x480 pixels[12].

c. Zero-Crossing representation method: This method was developed by Boles. It represents features of the iris at different resolution levels based on the wavelet transform zero-crossing. The algorithm is translation, rotation and scale invariant. The input images are processed to obtain a set of 1D signals and its zero crossing representation based on its dyadic wavelet transform[9].

d. Intensity variations: An Iris recognition system developed by Li Ma and characterized by local intensity variations. The sharp variation points of iris patterns are recorded as features[9].

e. Independent Component Analysis: The iris recognition system developed by Ya-Ping Huang adopts Independent Component Analysis (ICA) to extract iris texture features[11].

**f.** Iris authentication based on Continuous Dynamic Programming: The technique proposed by Radhika authenticates iris based on kinematic characteristics, acceleration. Pupil extraction begins by identifying the highest peak from the histogram which provides the threshold for lower intensity values of the eye image [12].

## 3.2 Feature extraction algorithms

In this paper, five different feature extraction algorithms have been used to extract the features prior to feature level fusion for the iris biometric. These are:

### a) Local Binary Pattern (LBP)

Local Binary Pattern is an efficient method used for feature extraction and texture classification**.** **LBP** was first introduced by Ojala[8]. The LBP operator was introduced as a complementary measure for local image contrast, and it was developed as a grayscale invariant pattern measure adding complementary information to the “amount” of texture in images. LBP is ideally suited for applications requiring fast feature extraction and texture classification[13]. Due to its discriminative power and computational simplicity, the LBP texture operator has become a popular approach in various applications, including visual inspection, image retrieval, remote sensing, biomedical image analysis, motion analysis, environmental modeling, and outdoor scene analysis[3] .

## b) Local Phase Quantization(LPQ)

Ojansivu et al. [8] proposed a new descriptor for texture classification named as Local phase quantization (LPQ) that is robust to image blurring. The descriptor utilizes phase information computed locally in a window for every image position. The phases of the four low-frequency coefficients are computed. These coefficients are decorrelated and uniformly quantized in an eight-dimensional space. The LPQ is in frequency domain, which contained partial contrast information of the image. However that is not enough to describe the texture features without contrast information in spatial domain. In LPQ algorithm, it firstly transformed the image from spatial domain to frequency domain to find the phase information coefficient for every single point from the image. Then the covariance between the adjacent pixels is computed.

## c) Haar Wavelet

Haar functions have been used from 1910 when they were introduced by the Hungarian mathematician Alfred Haar[9]. Haar wavelet is discontinuous, and resembles a step function. It represents the same wavelets Daubechiesdb1. In mathematics, the Haar wavelet is a certain sequence of functions. Haar used these functions to give an example of a countable orthonormal system for the space of square-integrable functions on the real line. The study of wavelets, and even the term "wavelet", did not come until much later. The technical disadvantage lies in the discontinuity of Haar wavelet, and therefore it is not differentiable. This property can, however, be an advantage for the analysis of signals with sudden transitions, such as monitoring of tool failure in machines.

## d) Gabor Filter

In image processing, a Gabor filter, named after Dennis Gabor, is a linear filter used for edge detection[10]. Frequency and orientation representations of Gabor filters are similar to those of the human visual system, and they have been found to be particularly appropriate for texture representation and discrimination. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave. The Gabor filters are self-similar: all filters can be generated from one mother wavelet by dilation and rotation.J. G. Daugman discovered that simple cells in the visual cortex of mammalian brains can be modeled by Gabor functions. Thus, image analysis by the Gabor functions is similar to perception in the human visual system.

## e) Daubechies(db) wavelet

Daubechies constructed the first wavelet family of scale functions that are orthogonal and have finite vanishing moments, i.e., compact support [9]. This property insures that the number of non-zero coefficients in the associated filter is finite. This is very useful for local analysis. The Haar wavelet is the basis of the simplest wavelet transform. It is also the only symmetric wavelet in the Daubechies family and the only one that has an explicit expression in discrete form. Haar wavelets are related to a mathematical operation called Haar transform, which serves as a prototype for all other wavelet transforms.

For the Daubechies wavelet transforms, the scaling signals and wavelets have slightly longer supports, i.e., they produce averages and differences using just a few more values from the signal. This slight change, however, provides a tremendous improvement in the capabilities of these new transforms. They provide us with a set of powerful tools for performing basic signal processing tasks. This family of wavelets with one parameter, due to Daubechies is the first one to make it possible to handle orthogonal wavelets with compact support and arbitrary regularity.

## 4. RESULTS AND DISCUSSION

This section, deals with the investigation consequences of combining different biometric feature extraction algorithm. Specifically at feature level fusion, Min-Max normalization rule to measure the performance of multi-algorithms system. In all the experiments, performance is measured in terms of Recognition rate, Recall, Precision and F-measure. First the performance of a single modality biometric system is measured; later the results for multi-algorithmic biometric system are evaluated.

| Feature extraction | LPQ | Gabor | Haar | Db8 | LBP |
|---|---|---|---|---|---|
| Recognition rate | 89 | 68 | 66.5 | 73 | 61 |
| Recall | 89 | 68 | 66.5 | 73 | 61 |
| Precision | 90.66 | 66.91 | 68.32 | 75.08 | 63.23 |
| F-measure | 89.82 | 67.45 | 67.39 | 74.02 | 62.09 |

CASIA Iris Image Database Version 1.0 (CASIA-IrisV1) is used for experimentations, it includes 756 iris images from 108 eyes[14]. For each eye, 7 images are captured in two sessions with our self-developed device CASIA close-up iris camera, where three samples are collected in the first session and four in the second session.

According to our experimental study, In table:1, it is found that as individual algorithm LPQ, Gabor Filter, Haar, Db8 and LBP independently results in 89%, 68%, 66.5%, 73% and 61% accuracy respectively. Performance of LPQ is the superior compare to other feature extraction methods.

By analyzing the results in Table-2, it can be observed that; The fusion of two feature extraction algorithms at feature level gives higher accuracy than single feature extraction method. We can also infer that, combination of LPB and LPQ algorithms outperforms the other existing combinations set.

However, Db8 and Haar algorithms combination is the least performance even compare to a single algorithm of LPQ. From Table -3, we can observe that there is no such impact in performance. Compare to its previous two algorithm fusion, however one can find the reduction in accuracy by introducing third feature extraction algorithm. Hence, choosing best combination plays a major role than increase the feature extraction algorithms.

**Table-1**: Performance of Single feature extraction algorithm

**Table-2**: Performance of two feature extraction algorithm

| Algorithm 1 | Algorithm 2 | Recognition | Recall | Precision | F-measure |
|---|---|---|---|---|---|
| Db8 | Haar | 85.5 | 85.5 | 87.7 | 86.6 |
| Db8 | LPQ | 92.0 | 92.0 | 90.6 | 91.3 |
| Db8 | LBP | 91.5 | 91.5 | 91.1 | 91.3 |
| Db8 | Gabor | 84.5 | 84.5 | 84.8 | 84.6 |
| Haar | LPQ | 92.0 | 92.0 | 90.5 | 91.2 |
| Haar | Gabor | 85.0 | 85.0 | 84.7 | 84.8 |
| Haar | LBP | 90.5 | 90.5 | 90.2 | 90.3 |
| LPQ | LBP | 95.0 | 95.0 | 93.8 | 94.4 |
| LPQ | Gabor | 89.0 | 89.0 | 89.1 | 89.0 |
| Gabor | LBP | 88.0 | 88.0 | 86.5 | 87.2 |

**Table-3**: Performance of three feature extraction algorithm

| Algorithm 1 | Algorithm 2 | Algorithm 3 | Recognition | Recall | Precision | F-measure |
|---|---|---|---|---|---|---|
| LBP | LPQ | Db8 | 93.5 | 93.5 | 92 | 92.8 |
| Db8 | Haar | LPQ | 90.5 | 90.5 | 89.8 | 90.14 |
| Db8 | Haar | LBP | 90 | 90 | 90.4 | 90.2 |
| Haar | Gabor | LPQ | 90.5 | 90.5 | 89.8 | 90.1 |
| Haar | Gabor | LBP | 88.5 | 88.5 | 88.2 | 88.3 |
| LPQ | Gabor | LBP | 94.5 | 94.5 | 93.7 | 94.1 |

## 5. CONCLUSION

From the analysis of experimental results and observations, it can be concluded that; The performance of fusion of three feature extraction algorithms, at feature level on feature normalization of Min-Max rule. Does not have expected performance improvement over the two feature extraction algorithms. Fusion of three algorithms in iris recognition, the performance is same or decreased when compared to the performance of fusion of two algorithms. One such case in our experiment is fusion of LBP and LPQ has 95% accuracy, when we introduce one more feature extraction algorithm to this combination, the performance is decreased drastically. The only exception is when fusing three algorithms of combination Db8 Haar and LPQ, which gained some good performance when compared to fusion of two algorithms. The performance of fusion of three algorithms, does not have performance improvement over the fusion of two which are the same performance in the best case. This is due to fusion of multiple redundant features at feature level, which also causes performance degradation of biometric system.

# REFERENCES

[ 1] Karen Hollingsworth , Kevin W. Bowyer, Stephen Lagree, Samuel P. Fenker, Patrick J. Flynn “Genetically identical irises have texture similarity that is not detected by iris biometrics” Computer Science and Engineering Department, University of Notre Dame, Notre Dame, IN 46556, United States.

[ 2] R. Wheeler, S. Aitken, “Multiple algorithms for fraud detection”, Artificial Intelligence Applications Institute, The University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland, UK.

[ 3] Chowhan.S.S and G.N.Shinde, “Iris Biometrics Recognition Application in Security Management”, *COCSIT, Ambajoagai Road, Latur413512, (M.S.) India,Indira Gandhi College, CIDCO, Nanded431602,(M.S.) India.

[ 4] S V Sheela, P A Vijaya, “Iris Recognition Methods – Survey”, International Journal of Computer Applications (0975 – 8887) Volume 3 – No.5, June 2010.

[ 5] Libor Masek, “ Recognition of Human Iris Patterns for Biometric Identification”, The University of Western Australia, 2003.

[ 6] SomnathDey and DebasisSamanta, “Improved Feature Processing for Iris Biometric Authentication System”, International Journal of Electrical and Electronics Engineering 4:2 2010

[ 7] Fenghua Wang, Jiuqiang Han, Xianghua Yao, “iris recognition based on multialgorithmic fusion”, WSEAS TRANSACTIONS ON INFORMATION SCIENCE & APPLICATIONS manuscript received may 20, 2007; revised august 10, 2007.

[ 8] Ojansivu, V., Heikkilä, J.: Blur Insensitive Texture Classification Using Local Phase Quantization. In Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) ICISP 2008 2008, LNCS, vol. 5099, pp. 236–243. Springer, Heidelberg (2008).

[ 9] Ayra Panganiban, Noel Linsangan and FelicitoCaluyo, “Wavelet-based Feature Extraction Algorithm for an Iris Recognition System”, Journal of Information Processing Systems, Vol.7, No.3, September 2011.

[ 10] C. Sanchez-Avila, R. Sanchez-Reillo, “Two different approaches for iris recognition using Gabor filters and multiscale zero-crossing representation”, Pattern Recognition Society, July 2004 (231 – 240).

[ 11] S V Sheela, P A Vijaya, “Iris Recognition Methods – Survey”, International Journal of Computer Applications (0975 – 8887) Volume 3 – No.5, June 2010.

[ 12] Hugo Proenca and Luis A. Alexandre, “Iris Recognition: An Analysis of the Aliasing Problem in the Iris Normalization Stage”, IEEE Proceedings of the 2006 International Conference on Computational Intelligence and Security - CIS 2006, Guangzhou, China, November 3-6, 2006, vol. 2, page. 1771-1774.

[ 13] Ahonen, T.; Hadid, A.; Pietikainen, M., "Face Description with Local Binary Patterns: Application to Face Recognition," Pattern Analysis and Machine Intelligence,

IEEE Transactions on , vol.28, no.12, pp.2037,2041, Dec. 2006

[ 14] Casia iris database. http://http://www.cbsr.ia.ac.cn/IrisDatabase.html

[ 15] John Daugman:Demodulation by Complex-Valued Wavelets for Stochastic Pattern Recognition. IJWMIP 1(1): 1-17 (2003)

# Route Update Overhead Reduction in MANETS Using Node Clustering

Anoop Abraham Eapen
Dept. of C.S.E,
Mangalam College of Engineering
Kottayam, Kerala, India

Vinodh P Vijayan
Dept. of C.S.E,
Mangalam College of Engineering
Kottayam, Kerala, India

**Abstract**: Most of the routing protocols used in Mobile Adhoc Networks (MANET) require update the route information to neighbour or any other nodes. These route update overhead degrade the performance of routing algorithms as there is a significant routing overhead. Proposed is a technique to reduce route update overhead through the minimization of the routing delay. The aim is to reduce the network congestion and minimize the complexities that are commonly faced by Mobile Adhoc Networks. Here, a clustering mechanism is introduced, which clusters the arriving nodes in the Mobile Adhoc Network. The election strategy is required to elect a particular node from the group of nodes in the cluster to act as the cluster head, based on the resources that are possessed by that node. Experimental results indicate reduced time and routing packets in this scheme

**Keywords**: MANET, PSR, Clustering, Routing, Neighborhood Trimming.

## 1. INTRODUCTION

The Mobile Adhoc Network or MANETS, is commonly used in different areas such as military communications. The concept of MANET was introduced in the year 1972. At that time they were introduced as Packet Radio Networks. Later, the second generation of MANETS was known as Survivable Adaptive Networks. They are used where speed and ease of deployment are a concern and where there is fewer infrastructures. Although there were many advantages that were associated with MANETS, they often posed different problems. Some of these issues include:

1. Routing
   The nodes within a wireless network can be mobile. They are able to change their location any time. So the process of sending packet from one node to another node within the network poses an issue, as the receiver may already have moved out of that particular network or might have failed.

2. Power:
   The mobile nodes that are present within the network has limited transmission power. So, we cannot expect all nodes to be active all the time. Any node within the network can fail at any time.

3. Security:
   In a wireless network there exists a problem of security. The packets that are transmitted from one node to another may get dropped due to network congestion or due to node failure. It is a serious issue as far as a network is concerned.

4. Quality of Service:
   Quality of service is never a fixed measure as far as MANETS are concerned. There are different nodes with different capabilities in the network. Some may fail during the operation or might leave the network. So, QoS is a variable measure.

The adhoc networks, i.e, MANETS are self-organizing and adaptive and they are able to adapt to changes within the network. The network performance can be further improved by localized grouping/clustering, which is explained. The paper is divided into four parts. The related works, the mechanisms used, evaluation and results.

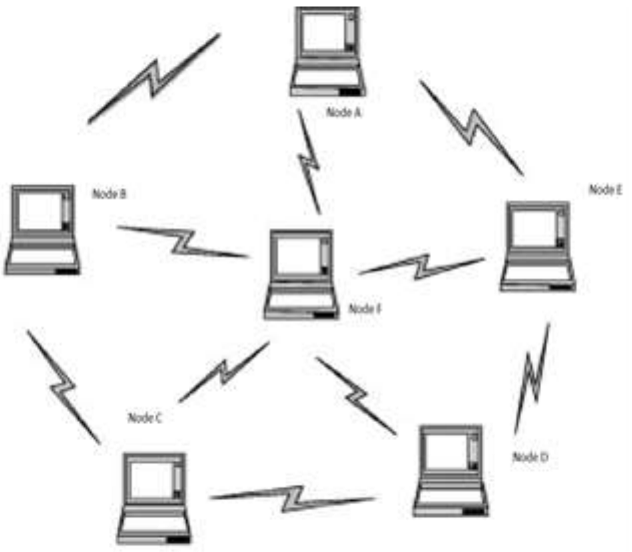


**Figure 1: Typical MANET architecture**

## 2. RELATED WORK

There have been different works associated with routing in MANETS. Due to the dynamic nature of MANETS, implementation of any static schemes was not feasible. Technologies that enable the operation of MANETS were also studied in different approaches [2]. To address the concept of mobility in MANETS, a special mechanism was introduced called VANETS. VANETS introduced the concept of geographic location based routing in vehicular networks, which helped to locate nodes by GPS [3]. Another routing scheme, ExOR [5] also focused on routing of packets in multi hop environments. There proactive as well as reactive routing protocols, i.e, table driven or dynamic in nature. Different routing protocols such as the The route path update was one of the major issues, which were associated with routing. Any node may arrive or leave to or from a network. The route path update is essential as one node present in the network may send a packet to another node, which might have already left the network earlier. PSR [1] proposed a route update scheme, where each node that is present within the network broadcasted the information to every other node that is present within the network. PSR was compared with other routing protocols such as the OLSR [6] and DSR [7]. For successful transmission of packets, each node should have an idea regarding the node to which it is transmitting the packet. In order to make this possible, route update messages need to be sent to the other nodes. This can have a negative impact on the network itself. In a large network, when all the nodes broadcast their route path updates, it may lead to network congestion. Periodic updates can be used, where the nodes periodically update their route paths. This can lead to much congestion within the network during that period. Another way is to carry out differential updates. i.e, the nodes only send the route update message, only when a significant update occurs, which was proposed along with PSR. These ideas were proposed earlier, so as to reduce network congestion, so as to reduce the network overhead. However, this can be further improved by adopting the clustering mechanism which is proposed here.

## 3. EXISTING SYSTEM

The existing system focuses on Proactive Source Routing protocol. Here, a breadth first spanning tree is maintained regarding the nodes in the network. The main focus of this scheme is to reduce the routing overhead in MANETS. The tree structure is periodically updated and broadcasted in each periodic update. Opportunistic data forwarding is also used here, by which the best neighboring node is allowed to forward the packets to the destination. PSR functions on the basis of a timer driven approach, where the information is broadcasted periodically among the nodes. There are different mobile nodes present within a network. Whenever a node wants to transmit a packet, it is forwarded to the destination with the use of the PSR. Besides the opportunistic data forwarding strategy, the PSR introduces the concepts of route update, neighborhood trimming and streamlined differential updates. The route update is done on a periodic basis. The neighborhood trimming is carried out so as to remove the unnecessary or failed nodes form the tree structure. Hello messages are usually broadcasted between a node and its neighbors. If a neighbor does not respond, then it can be

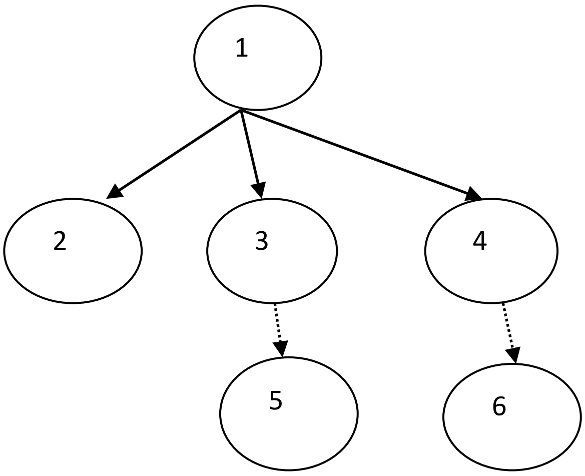


**Figure 2: Breadth first spanning tree**

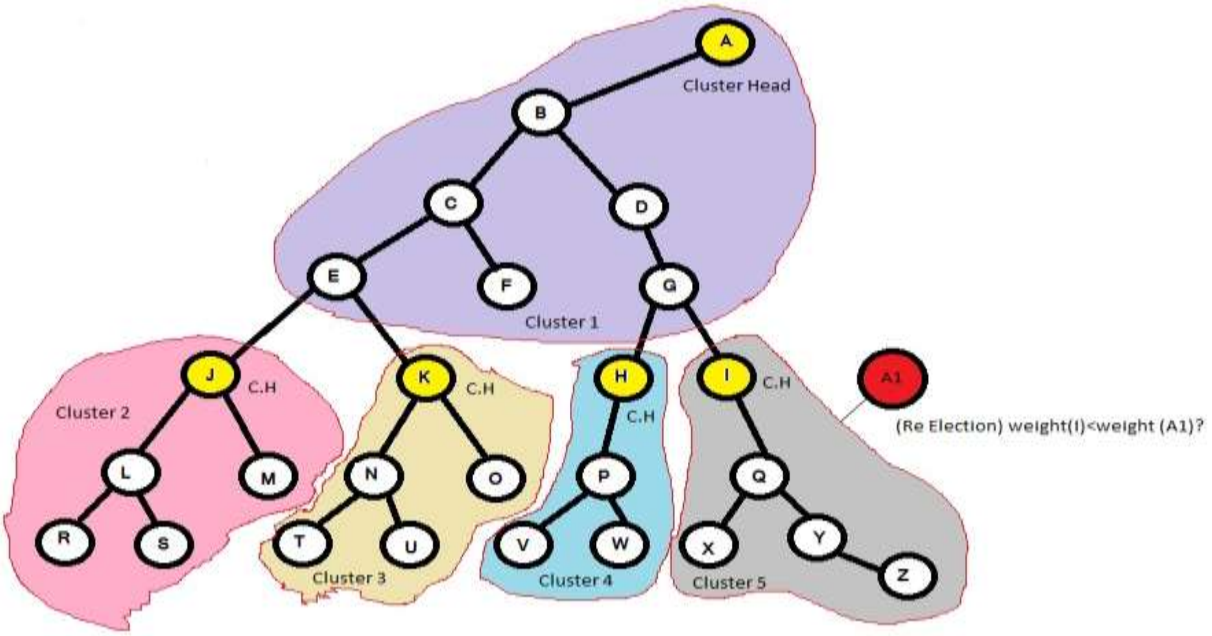


**Figure 3: Cluster & Cluster head**

deemed to be lost.

The routing information is stored in the routing table. When a node leaves a network or if a particular node fails, supposing that the node is the receiver for the message, the sender will send assuming that the receiver is still online. So solve this problem, periodic route update broadcast messages are transmitted between the nodes. So, if hundred nodes are present in the network, then each node will have to transmit ninety nine route update broadcast messages, so as to ensure that all the nodes are aware of the status of every other node. This is done on a periodic basis. During that time, the route path update messages that are transmitted will produce increased network traffic, leading to network congestion. Differential updates are performed, so as to reduce the number of messages transmitted. i.e, depending on the change, the message will be broadcasted. This overhead can be further reduced, if nodes are made to manage themselves as small groups.

.

# 4. PROPOSED SYSTEM

## 4.1 Clustering with PSR

Here, a clustering mechanism is proposed, which is used to further reduce the problem of network congestion that was prevalent within the existing PSR. i.e, the route path update messages were broadcasted on a periodic basis in the previous scheme so as to maintain the network information. Here, groups/clusters are formed so as to reduce the route path update overhead. Each cluster consists of the following:

## 4.2 Cluster Head:

The role of the cluster head is to maintain the cluster. The cluster head is elected by using the BBCMS[9] election algorithm. The cluster changes are locally informed to all the nodes within the cluster, so that the failure of a cluster head will not render the cluster ineffective. The election algorithm considers several parameters based on which the cluster head is elected.

## 4.3 Sub Nodes:

The remaining nodes within the cluster constitute the sub nodes. They broadcast their updates, if any (due to differential update), to the local group only. Hence it is assured the network cannot get congested easily with update broadcast messages as in earlier setup.

## 4.4 BBCMS

The cluster head has to be elected from the available nodes. Any node cannot be declared as the cluster head. The cluster head is selected on the basis of certain parameters that are specified within the algorithm. They include:

### *4.4.1 Belief Value(B):*

It is defined as how much a node is trusted by its neighboring nodes. The belief value is a way of measuring how stable, a particular node is.

### *4.4.2 Connectivity(C):*

Defined as the number of neighbors of a node within a 2d hop.

### *4.4.3 Battery Power(b):*

The cluster head should have fairly enough battery power to carry out its activities. If the battery power of a node is too low, it may go offline or may fail, as a cluster head. So, battery power is considered for the cluster head election criteria.

### *4.4.4 Max Value(M):*

Max Value is defined as the total number of nodes that can exist within a cluster

### *4.4.5 Stability:*

The stability of a node is calculated on the basis of the following parameters:

- Distance:
  The distance between the two nodes can be found out by using the distance formula.

- Average distance:
  The average distance between a node and its neighbors in the cluster.

- Mobility:
  It's the difference between value of average distances between two points

$$\text{MTA}= \text{A}D_t - AD_{t-1}$$

- Weight Factor:
  The weight factor is the value that is assigned to each parameter, based on which the global weight is calculated.

- Global Weight:
  Global weight of the node is the weight that is calculated by considering all the above parameters, which will be used in the cluster head election process. It is calculated as:

$$W_G[i]=(W_B[i]*F_B[i])+(W_C[i]*F_C[i])+(W_b[i]*F_b[i])+(W_M[i]*F_M[i])+(W_S[i]*F_S[i])$$

$W_B$[i]: Partial weight factor for belief value
$W_b$[i]: Partial weight factor for battery
$W_C$[i]: Partial weight factor for node connectivity
$W_M$[i]: Partial weight factor for Max Value
$W_S$[i]: Partial weight factor for Stability
$F_B$[i]: Belief Value
$F_C$[i]: Connectivity
$F_b$[i]: Battery Power

### *4.4.6 Steps in BBCMS algorithm:*

1. Random number generation:
   A random number generator is used so as to assign random values to the nodes that are present within the MANET. Random values are assigned to the nodes so as to simulate different characteristics of the nodes.

2. Cluster creation:
   The cluster is created based on the availability of the nodes. Initially, the one node available is considered, and clusters are formed mainly considering the max value set. If it exceeds the max value, then a new cluster is added.

3. Cluster head election:
   Cluster head election is carried out by considering the weight that is calculated for the nodes. The node with the minimum weight factor is made as the cluster head.

4. New node arrival:
   When a new node arrives, the weight of the node is compared with the existing cluster head. If the weight is more, then it is made as the cluster head. In this approach, however, we change the cluster head only if it fails or leaves the cluster so as to reduce the unnecessary complexity associated with the election and reelection.

5. Battery threshold:
   The battery power of the cluster head is compared with the threshold value. If it is found to be lower, then the cluster head is reelected.

6. Certificate revocation:
   In this scenario a new security certificate is issued to the newly arriving nodes joining the cluster.

## 5. EXPERIMENTAL RESULTS

The experimentation was carried out in order to carry out the comparative study of this approach with PSR in reducing the route update overhead. This section deals with the system and tools used as well as the experimental methodologies adopted for this evaluation.

### 5.1 System & Tools used

The simulation of the network was developed in java. The implementation was simulated with the help of provision of random values assigned to simulate the network environment. The mobile node parameters such as battery power, mobility etc were assigned by the use of a random function. The database was created using MySQL and was deployed with the help of Wamp Server. Netbeans was the development platform used. The system used was running Windows 7 64 bit os with an Intel core i5 processor and 12GB of RAM.

### 5.2 Evaluation

The system was evaluated by comparing with the existing PSR approach. In the existing PSR approach, the nodes were distributed randomly in the network and the message passing overhead was high. A test bed was developed in order to simulate this environment. In the first approach, the nodes were distributed in the network. If at all any node had to leave the network, it would have to send the route path update broadcast message to all the remaining nodes that was present in the entire network. The time to send the update message was calculated in milliseconds. It may vary depending upon the performance of the simulated system. In an actual scenario, it will depend upon network performance. The number of route path update broadcast messages was also calculated.

In the second simulation, the PSR with cluster mechanism was employed. Based on the random values that were assigned to the nodes by the random generator function, the BBCMS clustering algorithm was used to create clusters of nodes. The BFST structure of the PSR was maintained here. The cluster heads are also elected, on the basis of the parameters specified by the algorithm. Whenever a node leaves a cluster, the information is broadcasted to the corresponding cluster heads as well as the nodes within the cluster.

The cluster head is then reelected on the basis of the parameters specified by the algorithm. Once the election is done, the information is broadcasted to all the remaining cluster heads that are present within the network. The failure of a cluster head will thus not affect the topology of the network, as the information is broadcasted to all the cluster head nodes on a periodic basis. The node that has left the cluster may rejoin any other cluster any time. When that happens, the information will be broadcasted to the cluster heads.

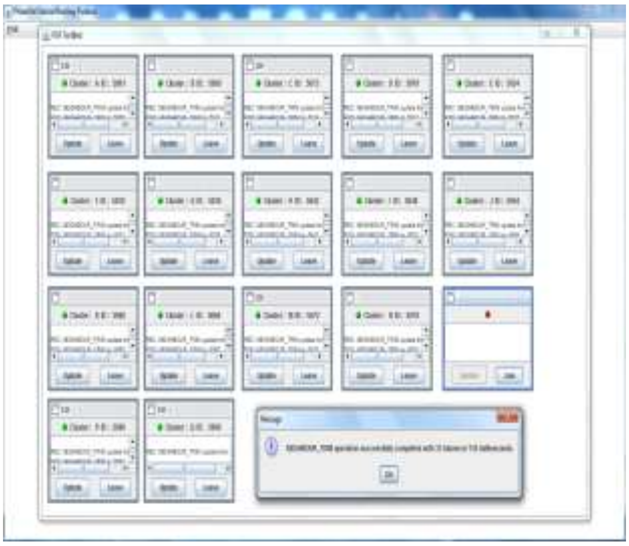


**Figure 4: Node population**

### 5.3 Results

The evaluation yielded that the time required to perform the update operation was reduced from the previous PSR approach, by using the clustering mechanism, as the total number of messages to be transmitted was reduced. Since the exchange of messages was primarily between the cluster head nodes, the total number of packets that was necessary for the transmission of broadcast information was reduced. Since less number of packets were to be transmitted, the route path update operation also completed faster in the PSR with cluster approach.

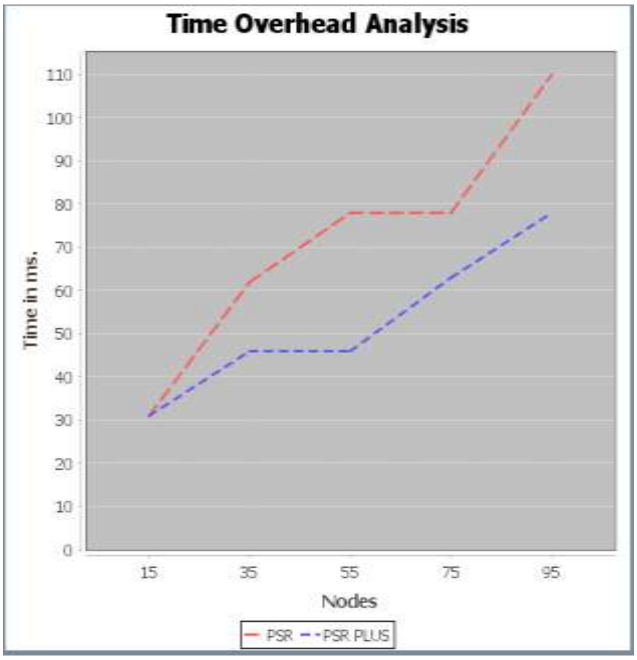


**Figure 5: Time taken to send packets for PSR and PSR with cluster**

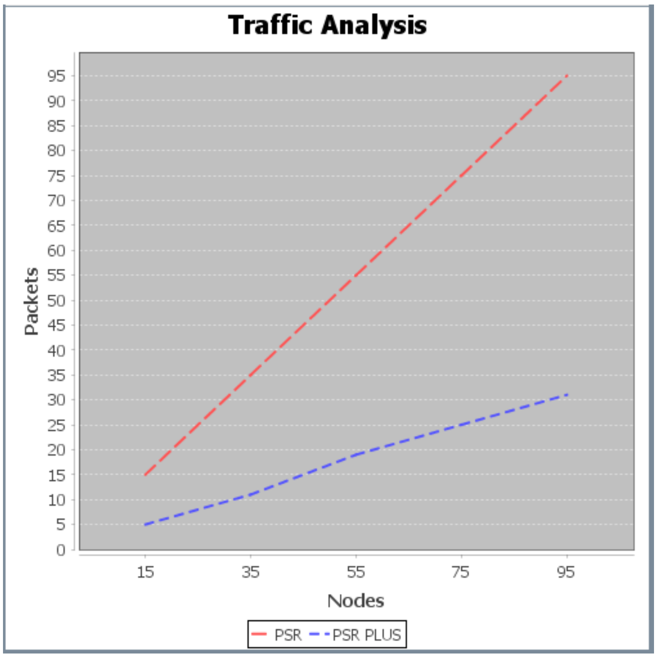


**Figure 6: No of broadcast packets for PSR and PSR with cluster**

## 6. CONCLUSION

The route path update overhead in MANETS is of great importance, as it degrades network performance. The nodes that are present within the network are arranged in a hierarchical form to be aware of other nodes that are present in the network so as to communicate effectively between the nodes. The clustering applied in the hierarchical network improves the network management by limiting the route update packet with in the cluster. The performance of the Network is improved however overall system is dependent on the frequency of cluster head failures. But in the case of a large network consisting of many nodes, this mechanism will ensure that the route path update overhead will be minimized and network congestion problems can be avoided.

## 7. FUTURE WORK

The system has proposed clustering mechanism in network to reduce the route update overhead. The cluster head failure is one of the major problems that are associated with this approach. To mitigate this, a live node monitoring approach can be employed, so as to detect any chances of failure corresponding to any head nodes, which are present in the cluster. The cluster heads can also be assigned more responsibilities, other than the transmission of the update packets.

## 8. REFERENCES

[1] Zehua Wang, Yuanzhu Chen, and Cheng Li, “PSR: A light weight Proactive Source Routing Protocol for Mobile Adhoc Networks, IEEE 2014

[2] I. Chlamtac, M. Conti, and J.-N. Liu, “Mobile ad hoc networking: Imperatives and challenges,” Ad Hoc Netw., vol. 1, no. 1, pp. 13–64, Jul. 2003.

[3] M. Al-Rabayah and R. Malaney, “A new scalable hybrid routing protocol for VANETs,” IEEE Trans. Veh. Technol., vol. 61, no. 6, pp. 2625–2635, Jul. 2012.

[4] Y. P. Chen, J. Zhang, and I. Marsic, “Link-layer-and-above diversity in multi-hop wireless networks,” IEEE Commun. Mag., vol. 47, no. 2, pp. 118–124, Feb. 2009.

[5] S. Biswas and R. Morris, “ExOR: Opportunistic multi-hop routing for wireless networks,” in Proc. ACM Conf. SIGCOMM, Philadelphia, PA, USA, Aug. 2005, pp. 133–144.

[6] T. Clausen and P. Jacquet, “Optimized Link State Routing Protocol (OLSR),” RFC 3626, Oct. 2003.

[7] D. B. Johnson, Y.-C. Hu, and D. A. Maltz, “On The Dynamic Source
Routing Protocol (DSR) for mobile ad hoc networks for IPv4,” RFC 4728,Feb. 2007.

[8] C. E. Perkins and P. Bhagwat, “Highly dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for mobile computers,” Comput. Commun. Rev., vol. 24, pp. 234–244, Oct. 1994.

[9] Heenavarsheney, Pradeep Kumar, “Secure Communication Architecture Based On “BBCMS” Clustering Algorithm for Mobile Adhoc Network” IJITEE vol.3 Isuue 2, July 2013.

# An Improved Energy Efficient Wireless Sensor Networks Through Clustering In Cross Layer Network Operations

Neethu Krishna
Department of CSE
Mangalam College of Engineering
Kottayam, Kerala, India-686631

Vinodh P Vijayan
Department of CSE
Mangalam College of Engineering
Kottayam, Kerala, India-686631

**Abstract**: One of the major reason for performance degradation in Wireless sensor network is the overhead due to control packet and packet delivery degradation. Clustering in cross layer network operation is an efficient way manage control packet overhead and which ultimately improve the lifetime of a network. All these overheads are crucial in a scalable networks. But the clustering always suffer from the cluster head failure which need to be solved effectively in a large network. As the focus is to improve the average lifetime of sensor network the cluster head is selected based on the battery life of nodes. The cross-layer operation model optimize the overheads in multiple layer and ultimately the use of clustering will reduce the major overheads identified and their by the energy consumption and throughput of wireless sensor network is improved. The proposed model operates on two layers of network ie., Network Layer and Transport Layer and Clustering is applied in the network layer . The simulation result shows that the integration of two layers reduces the energy consumption and increases the throughput of the wireless sensor networks.

**Keywords**: Cross layer design, energy efficiency, wireless sensor networks, clustering

## 1. INTRODUCTION

Wireless sensor network consists of spatially distributed autonomous sensors to monitor environmental conditions such as sound, temperature, pressure etc. Sensor nodes can sense and detect events in the region and communicate data back to the Base Station (BS). Wireless Sensor Network have become most interesting area of research. Sensor nodes are equipped with small batteries that can store at most I J. Limiting the transmission range and power consumption are the important constraints offered for communication, and hence it is advantageous to put in order the sensors into clusters.

Clustering in cross layer networks is one of the important mechanism to improve the energy consumption of sensor network and thereby increase the network lifetime. In clustering, whole sensor network is divided into group of clusters. Cluster head is selected based on the battery life of a node. Cluster head gather and aggregate the data and send it back to the BS.

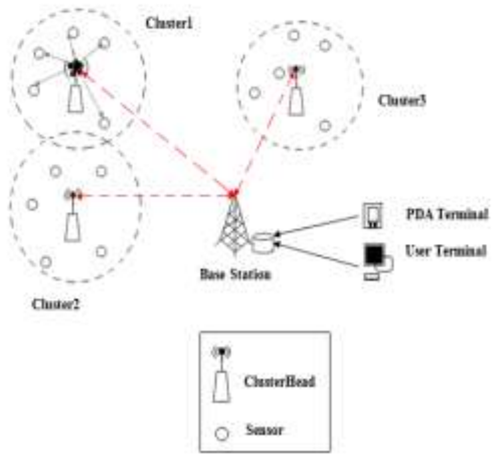


**Figure 1. Clusters with single cluster head**

Some of the clustering goals in wireless sensor networks are as follows;

- Data aggregation and limits data transmission
- Facilitating the reusability of the resources
- CHs and gateway nodes can form a virtual backbone for inter-cluster communications
- Cluster structure gives the impression of a smaller and more stable network
- Improve network lifetime: reducing network traffic and the contention for the channel
- Grouping of similar objects or sensors
- Topology control by load balancing and network scalability

The advantage of cluster is to collect data from neighboring node is operationally more convenient then observing units spread over a region. Clustering technique is done in the network layer. In transport layer, nodes are scheduled on the basis of how much time they are active.

The remainder of this paper is organized as follows. Section 2 describes related work on cross layer network operations. Section 3 describes about proposed work and implementation. Section 4 describes experimental evaluation. Section 5 concludes this paper with a discussion of future approach.

## 2. RELATED WORK

Lifetime extension of wireless sensor network [2] uses two cluster heads and hierarchical routing. In this paper an algorithm Two Cluster Head Energy efficient Wireless Sensor Network (TCHE-WSN) is proposed to improve the lifetime. The use of two cluster heads analogy reduces the overhead of single cluster head, avoids packet collision and improves reliable data transmission.

A clustering based routing protocol called base station controlled dynamic clustering protocol, utilizes a high energy base station o setup cluster heads and perform other energy efficient tasks and thereby increasing the lifetime of a network. A cross layer network operation mechanism [5] which considers the physical and MAC layers to maximize the lifetime of a network. The model assumes that the problem of network is convex where G (P, h(ni )) is the network graph , P is the set of nodes deployed and h(ni ) is the amount of data needed from node i to indicate the sensed event in the deployment area. The deployed nodes are static and this model has not been tested for wireless sensor networks with mobility characteristics.

Load balancing and clustering in Hybrid Sensor network with mobile Cluster Nodes [8] proposed an algorithm which works on the position of mobile cluster heads balancing of traffic load in sensor network that consist of mobile and static nodes.

Low-energy adaptive clustering hierarchy (LEACH) [9] is a clustering-based protocol which utilizes randomized rotation of local CHs to evenly distribute the energy load across the network. Compared with other ordinary routing protocols like DD, it can prolong the network lifetime up to 8 times. However, the 5% of CHs are randomly selected and CHs transmit data directly to SN. Reference [3] proposed an Energy Efficient and QoS aware multipath routing protocol (EQSR) has been proposed for WSNs. This protocol is mainly used to find out the best path from the multiple path from source to destination. This protocol chooses its routing path based on the physical layer elements of the next hop. Those elements are the nodes residual energy interface buffer availability and the connection signal-to-noise ratio between two neighbour nodes. This protocol is an example of the tight cross layer of information between the physical layer and the network layer.

In Energy Efficient Hierarchical Clustering Algorithm [7] a distributed, randomized clustering algorithm is proposed. The algorithm generates hierarchy of cluster heads. It has been observed that the energy savings increases with the number of levels in the hierarchy and thereby increases the lifetime of a network.

SPEED [11] is another QoS based routing protocol that provides soft real-time end-to-end guarantees. Each sensor node maintains information about its neighbors and exploits geographic forwarding to find the paths. To ensure packet delivery within the required time limits, SPEED enables the application to compute the end-to-end delay by dividing the distance to the sink by the speed of packet delivery before making any admission decision. In addition to that, SPEED can provide congestion avoidance when the network is congested.

In order to suit the periodical data gathering applications an Energy Efficient Clustering scheme [4] a novel scheme (EECS) for single-hop wireless sensor networks. This paper dealt with an approach to elect cluster heads with more residual energy in an autonomous manner using local radio communication. It produce good cluster head distribution and balances the load among cluster heads using this novel scheme.

## 3. PROPOSED APPROACH

In this paper, Cross layer network operation is done in two layers which are Network Layer and Transport Layer .In Network Layer clustering technique(LEACH) is used and in the Transport layer energy can be improved by sensing the distance of nodes in which how long they are far from the antenna. We can analyze the performance by adjusting the antenna range. One of the important factors to improve lifetime of wireless sensor network is the design of network. LEACH(Low-Energy Adaptive Clustering Hierarchy) protocol is used for clustering.

### 3.1 Set-up phase

In LEACH, nodes take autonomous decisions to form clusters by using a distributed algorithm with out any centralized control. Here no long-distance communication with the base station is required and distributed cluster formation can be done without knowing the exact location of any of the nodes in the network. In addition, no global communication is needed to set up the clusters. The cluster formation algorithm should be designed such that nodes are cluster-heads approximately the same number of time, assuming all the nodes start with the same amount of energy. Finally, the cluster-head nodes should be spread throughout the network, as this will minimize the distance the non-cluster-head nodes need to send their data. A sensor node chooses a random number, r, between 0 and 1. Let a threshold value be T(n)

$$T(n)= p/1-p \times ( r \bmod p-1 )$$

If this random number is less than a threshold value, T(n), the node becomes a cluster-head for the current round. The threshold value is calculated based on the above given equation that incorporates the desired percentage to become a cluster-head, the current round, and the set of nodes that have not been selected as a cluster-head in the last (1/P) rounds, p is cluster head probability. After the nodes have elected themselves to be cluster-heads, it broadcasts an advertisement message (ADV). This message is a small message containing the node's ID and a header that distinguishes this message as an announcement message. Cluster head is selected based on the battery life. First sense the energy of whole network and then select each node’s sensing power and processing power in which node has high battery life, it would be the cluster head. the energy of node which has less than 5J it is disabled from the cluster.. The cluster-heads in LEACH act as local control centers to co-ordinate the data transmissions in their cluster [9]. The cluster-head node sets up a TDMA schedule and transmits this schedule to the nodes in the cluster. This ensures that there are no collisions among data messages and also allows the radio components of each non cluster-head node to be turned off at all times except during their transmit time, thus minimizing the energy dissipated by the individual.

### 3.2 Steady-State Phase

The steady-state operation is broken into frames where nodes send their data to the cluster-head at most once per frame during their allocated transmission slot. The set-up phase does not guarantee that nodes are evenly distributed among the cluster head nodes. Therefore, the number of nodes per cluster is highly variable in LEACH, and the amount of data each node can send to the cluster-head varies depending on the number of nodes in the cluster. To reduce energy dissipation, each non-cluster-head node uses power control to set the amount of transmits power based on the received strength of the cluster-head advertisement. The radio of each non-cluster-head node is turned off until its allocated transmission time. Since all the nodes have data to send to the cluster-head and the total bandwidth is fixed, using a TDMA schedule is efficient use of bandwidth and represents a low latency approach, in addition to being energy-efficient. The cluster-head must keep its receiver on to receive all the data from the nodes in the cluster. Once the cluster-head receives all the data, it can operate on the data and then the resultant data are sent from the cluster-head to the base station.

## 4. EXPERIMENTAL RESULTS

An experiment set up has done using Network Simulator 2 version 2.29 (ns-2).The Energy constraint is an important factor for Wireless sensor networks, Leach Protocol is used for the simulation. NS-2 is a tool that provide rich environment for simulation of wireless sensor network at different layers. Following are details of the experimental setup and collected result.

### 4.1 Experimental Setup

Simulation is done on Mannasim simulator for finding out the energy effectiveness of network. Here clustering technique is used on the basis of LEACH protocol. Cluster head is selected based on the battery life of node. It senses the sending energy power and processing power of each node with time. If the energy of the node is less than 5 J , it is disabled from the cluster which it belongs. So that energy can be improved and cluster can send the data to base station easily without losing so much of power and thereby increasing the lifetime of a network.

### 4.2 Result

Simulations are carried out and results are obtained. Results obtained are compared with the AODV protocol. Fig. 2 shows the sensing energy power with time. By using LEACH protocol network life time can be increased more than that of using AODV protocol.

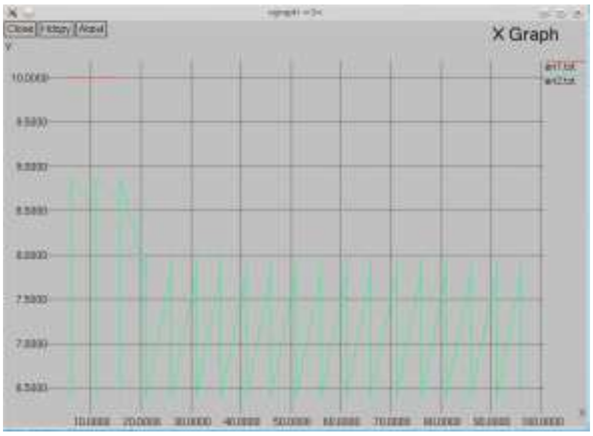


**Figure 2. Energy Vs Time**

Figure 3. shows the residual energy. Lifetime of wireless sensor network is evaluated in terms of alive sensor nodes over the time period and residual energy of sensor network. From the figure it is observed that the proposed clustering algorithm achieves better network life time as compared to the AODV protocol.

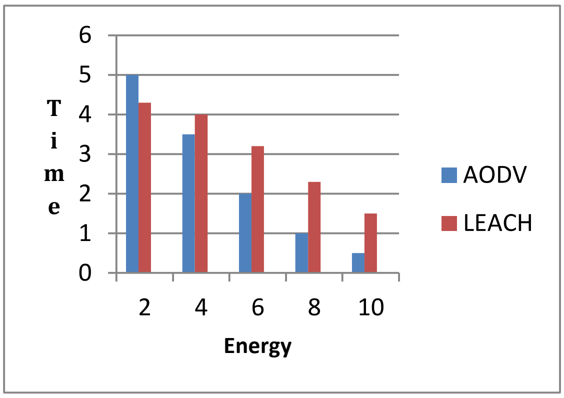


**Figure 3. Residual energy**

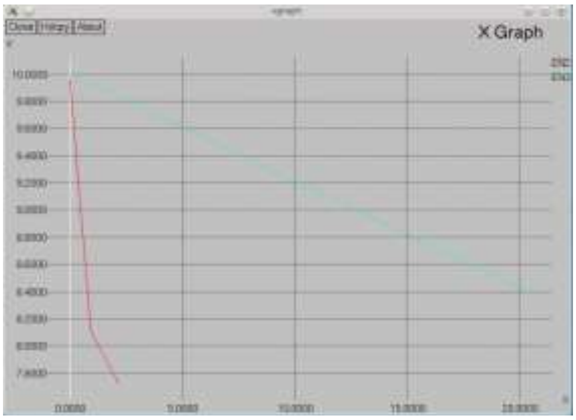


**Figure 4. Energy consumption of specific node**

The Figure 4. shows the energy versus time graph of a specific node. The red line shows the energy depletion of sensor node that runs on AODV protocol, whereas the green line shows the energy depletion of sensor node running on LEACH protocol. For a lifetime of the network, a node can be elected as a cluster head based on the battery life and maintain the network life time easily.

## 5. CONCLUSION

Packet delivery degradation and control packet overhead are the main issues occur in cross layer network operation model of wireless sensor networks. In order to reduce these issues clustering mechanism is proposed in cross layer network operations and thereby increases the lifetime and throughput of network. Here, cluster head selection is based on the battery life of nodes. Simulation results demonstrate that proposed clustering algorithm using high energy cluster heads is more effective in prolonging the lifetime of sensor network than using AODV algorithm. The reserved energy in the sensor nodes leads to the extended life time of entire wireless sensor networks.

## 6. FUTURE WORK

In future, algorithm can enhanced to incorporate load balancing among the cluster heads based on the parameter like traffic. More energy preservation could be done by studying the placement of sensor nodes in network which opens us a scope for future research.

## 7. REFERENCES

[1] Marwan -Jemeli, Fawnizu A. Hussin, Marwan Al-Jemeli, and Fawnizu A. Hussin," An Energy Efficient Cross- Layer Network Operation Model for IEEE 802.15.4- Based Mobile Wireless Sensor Networks," ," *IEEE SENSORS JOURNAL*, VOL. 15, NO. 2, FEBRUARY 2015.

[2] ] B. Meenakshi, P. Anandhakumar, "Lifetime extension of wireless sensor network by selecting two cluster heads and hierarchical routing", *IEEE International Conference on Advances in Computing, Communications and Informatics, 2012.*

[3] ] H.-W. Tseng, S.-C. Yang, P.-C. Yeh, and A.-C. Pang, "A cross-layer scheme for solving hidden device problem in IEEE 802.15.4 wireless sensor networks," *IEEE Sensors J.*, vol. 11, no. 2, pp. 493–504, Feb. 2011.

[4] Mao YE, Cengafa LI, Guihai Chen, Jie WU, "Energy Efficient Clustering Scheme in Wireless Sensor Networks", *IEEE International Conference on Performance, Computing and Communications*, 2005.

[5] S. He, J. Chen, D. K. Y. Yau, and Y. Sun, "Cross-layer optimization of correlated data gathering in wireless sensor networks," in *Proc. 7th Annu. IEEE Commun. Soc.Conf. Sensor Mesh Ad Hoc Commun. Network( SECON)*, Jun. 2010, pp. 1–9.

[6] J. Ben-Othman and B. Yahya, "Energy efficient and QoS based routing protocol for wireless sensor networks," *J.Parallel Distrib. Comput.*vol. 70, no. 8, pp. 849–857, Aug. 2010

[7] Seema Bandyopadhyay, Edward J. Coyle, "An Energy Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks", Twenty-Second Annual Joint Conference of the IEEE Computer and Communications, 2003.

[8] Gaurav Gupta, Mohamed Younis, "Performance Evaluation of Load- Balanced Clustering of Wireless Sensor Networks", IEEE International Conference on Telecommunications, 2003.

[9] J. Ben-Othman and B. Yahya, "Energy efficient and QoS based routing protocol for wireless sensor networks," *J. Parallel Distrib. Comput.* vol. 70, no. 8, pp. 849–857, Aug. 2010.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 8, 608 - 613, 2015, ISSN: 2319–8656

# Utilization of Support Vector Machine for Efficient CBVR and Classification of Video Database using Gabor Features from Multiple Frames

Mohd. Aasif Ansari
Engineering and Technology
Shri Venkateshwara University
Gajraula, India

Hemlata Vasishtha
Shri Venkateshwara University
Gajraula, India

**Abstract**: Content Based Video Retrieval (CBVR) systems are used for retrieval of desired videos from a large collection on the basis of features extracted from videos. The extracted features are used to index, classify and retrieve desired and relevant videos while filtering out undesired ones. Videos can be represented by their audio, texts, faces and objects in their frames. An individual video possesses unique motion features, color histograms, motion histograms, text features, audio features, features extracted from faces and objects existing in its frames. Videos containing useful information and occupying significant space in the databases are under-utilized unless exist CBVR systems capable of retrieving desired videos by sharply selecting relevant while filtering out undesired videos. Results have shown performance improvement when features suitable to particular types of videos are utilized wisely. Various combinations of these features can also be used to achieve desired performance. Many researchers have an opinion that result is poor when images are used as a query for video retrieval. Here, instead of using a single image or key frames, multiple frames of the video clip being searched are used. Also, instead of using Euclidean Distance to measure similarity Support Vector Machine (SVM) is used. This method used for CBVR system shown in this paper yields an enhanced and higher retrieval results. Also, multiple frames based classification and retrieval yields significantly higher results without the complexity of finding key frames to represent a shot. The system is implemented using MATLAB. Performance of the system is assessed using a database containing 1000 video clips of 20 different categories with each category having 50 clips. The performance is tested using features extracted using Gabor filters as these are most frequently used to represent texture features.

**Keywords**: CBVR; Multiple Frames; Gabor; SVM; MATLAB

## 1. INTRODUCTION

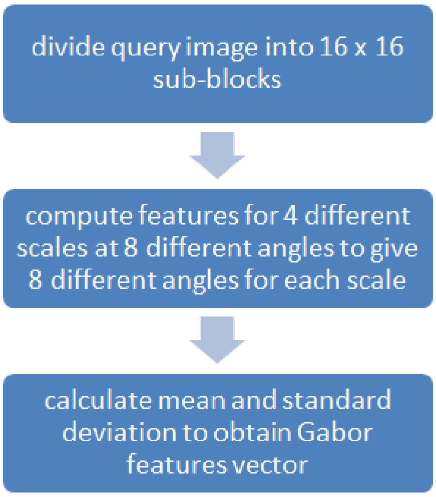
With lack of satisfaction from textual based video retrieval, the idea of content based video retrieval has been the attention for researchers since long time. In the beginning of content based video retrieval, they tried to retrieve videos using an image. However, video retrieval using query by image is not successful as it cannot represent a video. A video is a sequence of images and audio. A query video provides rich content information than that provided by a query image. Finding the relevant video by sequentially comparing the low level visual features of key frames of the query video with those of key frames of videos in database provide long pending solution to yield better result [6] of video retrieval. Finding similarity measure requires key frames matching and hence computing key frame features including color histogram, texture and edge features, etc., to calculate distance parameter. These huge computations cause long response time to the users and thus, the problem of high computation cost in computing visual features of videos is persistent. Apart from this, considerations for motion features, temporal, sequence and duration of shots in a video pose a challenge for the research area [5]. The structural and content attributes obtained through content analysis, segmentation, video parsing, abstraction processes and the attributes entered manually are referred to as metadata. Video is indexed on a table using the metadata using clustering process which categorizes video clips or shots. Clustering process categorizes video clips or shots using metadata to form an index table of videos into different visual categories. Researchers have developed various tools and schemes to index, enquire, browse, search and retrieve videos from large databases but effective and robust tools are still lacking to test with large databases [6]. Due to these limitations [5], [6] a majority of video searches and retrievals still relies on keyword or text attributions. Face detection is assessed for image and video analysis. It was experimented in a commercial system [15]. It was found that accuracy of face recognition in video collection of the type mentioned in the system [8] was too poor to prove to be useful. Overall a large number of queries do not yield satisfactory results as mentioned [8] about one third of the queries were unanswerable by any of the automatic systems participating in the video retrieval track [16]. No system or method was able to provide relevant results. An integrated video retrieval system is proposed [2] where a video shot is represented not by key frame only but by all frames to extract more visual features of a shot. Color and motion features are integrated to fully exploit the spatio-temporal information contained in a video [29]. To overcome these drawbacks, i.e. considering lower efficiency of CBVR systems using a single image and very high computational cost of CBVR systems using key frames and the problem of availability of effective tools for CBVR systems using clustering process and to strike a balance between the efficiency and computational cost, visual features from multiple frames of a video clip are used in the system proposed here instead of a single frame or key frames or all frames of a clip. Also, it is learnt from the evaluation of video information retrieval that good image retrieval leads to good performance of video retrieval system when query is an image or an image from the query video [8]. Computational cost point of view, the system proposed in this paper is cost effective along with acceptable as well as significantly higher results.

In section 2 features and features extraction algorithms are discussed; section 3 discuses about similarity measure; section 4shows the methodology to calculate result parameters in the proposed CBVR system. Proposed CBVR system is elaborated in section 5 and the result charts are shown in section 6; problems and challenges posed to this CBVR system are discussed in section 7 and the conclusion is presented in section 8.

## 2. FEATURES AND FEATURES EXTRACTION

### 2.1 Extraction of Gabor Features

For effective video indexing, classification and retrieval visual features embedded in video data is exploited. Three primary features to be extracted are color, texture and motion for effective video indexing. These features are represented by color histogram, Gabor texture features and motion histogram respectively [4]. Edge histogram and texture features are one of the most reliable data for effective video retrieval application. Gabor filters can also be used to obtain textural properties of texts which are distinct and distinguish them from its background in the image [7]. Extraction of Gabor features involves finding local energy of the signal i.e., localized frequency parameters are obtained. Gabor filter consists of multiple wavelets obtaining energy in multiple orientations with multiple frequencies with each of them tuned to a particular direction and frequency. Thus, texture features are obtained. The texture features are used to find images or regions inside the images having similar textures. The filters of a Gabor filter bank are designed to detect different frequencies and orientations [30]. They can be used to extract features on key points detected by interest operators [17]. From each filtered image, Gabor features are calculated and used to retrieve images. The algorithm for extracting the Gabor feature vector is shown in fig. 1 and the related equations (1 - 4) are also shown below [18], [20].



divide query image into 16 x 16 sub-blocks

compute features for 4 different scales at 8 different angles to give 8 different angles for each scale

calculate mean and standard deviation to obtain Gabor features vector

Figure 1. Gabor Filter Algorithm

For a given image The discrete Gabor wavelet transform is given by a convolution using equation (1) for an image I(r,c) where, r = 0,1,2,..R and c = 0,1,2,..C.

$$W_{uv} = \sum_{p}\sum_{q} I(r-p, c-q) G_{uv}{}^{*}(p,q) \quad -(1)$$

where, $G_{uv}{}^{*}$ is complex conjugate of $G_{uv}$ . $G_{uv}$ is generated by some morphological operations on mother wavelet. P X Q is the size of filter mask, u and v are scale and orientations. Gabor filters are applied on the image with different orientations and different scales to find a set of magnitudes U(u, v)containing the energy distribution in the image in different orientations and scales.

$$E(u,v) = \sum_{r}\sum_{c} |W_{uv}(r,c)| \quad - \quad (2)$$

Since we are interested to obtain texture features Standard deviation σ and mean is calculated using equations (3) and (4) respectively

$$\text{Standard Deviation, } \sigma_{uv} = \sqrt{\frac{\sum_r \sum_c (|W_{uv}(r,c)| - \mu_{uv})^2}{R\ X\ C}} \quad - \quad (3)$$

$$\text{Mean, } \mu_{uv} = \frac{E(u,v)}{R\ X\ C} - (4)$$

Texture features vector F is formed by a set of feature components [19], [14] i.e., different values of $\sigma_{uv}$ and $\mu_{uv}$ calculated by varying u and v as shown in equation (5).

$$f = [\sigma_{u0v0}, \sigma_{u1}\sigma_{v1} \dots \sigma_{uUvV}] - (5)$$

$$f_{Gabor} = \frac{f - \mu}{\sigma} \quad - \quad (6)$$

### 2.2 Classification of features using Support Vector Machine

Use of Support Vector Machine (SVM) can be of great help for video classification. The frames from a video or a key frame representing a shot can be used to represent a video. It can also be represented by other components such as shots, scenes or events. Features are extracted from these video components. Corresponding features of videos from different categories are labeled to train SVM. Once the SVM is trained for these classes, it can be used to classify another group of videos having features extracted similarly. It is a big achievement towards automatic classification of videos [21]. Enhanced results can be obtained to classify a group of videos into their corresponding categories as it has been already obtained for features representing images. It has been observed that SVM can improve the results for CBIR problems [11]. SVMs are kernel based techniques used for classification. They can perform linear as well as non-linear classification as per the kernel design. The training process of a SVM is shown according to equations mentioned below.

Let's have a data (which may be feature vectors) $V_K$ of m points spreaded over a d dimensional plane is used to train a SVM.

$$X = \left\{(V_K, C_K) \mid V_K \in R^d, C_K \in \{-1, +1\}\right\}_{K=1}^{m} - (7)$$

$X$is termed as the training data. The data $V_K$ is to be classified among two different categories as denoted by $C_K \in \{-1, +1\}$ and $V_K$ is a d dimensional real vector.

We need to find a hyper plane separating the data $V_K$ as shown in the fig. 2

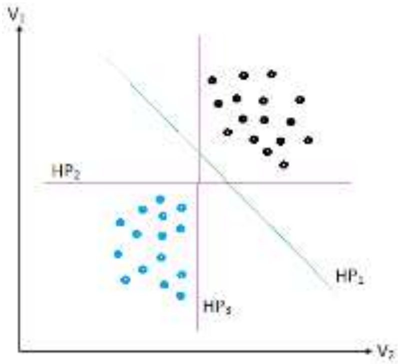


Figure 2. Hyper planes and two classes

Fig. 2 shows three hyper planes separating the two classes of variables. It can be observed that hyper planes HP$_2$ and HP$_3$ are separating the two classes but the margins are very less as they are very closed to some of the variables while the hyper plane HP$_1$ separates the two classes with a good margin. So HP$_1$ is selected while training the SVM. The hyper plane is shown by equation 8.

$$R \cdot V - q = 0 \quad - \quad (8)$$

Where, $R$ is a normal vector to the hyperplane, $\cdot$ denotes dot product and a variable $\frac{q}{\| R \|}$ is used to find the offset of the hyperplane from the origin along the normal vector $R$. The given classification is linear classification. Linear classification is always not possible. In such cases, non-linear classification is required using non-linear equations for the kernel used for SVM training.

## 3. SIMILARITY MEASURE

Queries are classified by categories sorted out according to type of features used or type of example data. The query is found out by calculating similarity between feature vector [9], [10] stored in the database and the features of the query videos. The similarity is obtained by classification of videos using these features. Measuring similarity by using features is most convenient and direct method [1]. It is found by obtaining groups of videos classified by an SVM using their features. In query by example frames like the one used in the system shown in this paper similarity measure to find relevant and similar videos usually low level feature matching is used. Video similarity can be measured at different levels of resolution or granularity [13]. A video clip is retrieved by finding most similar video from the group of videos classified by the SVM. Furthermore, the most similar video can be obtained using the frames separated out from the enquired clip with those of the videos stored in the database. Video retrieval result depends greatly on video similarity measures. The videos are retrieved by finding similarity between the features extracted from multiple frames associated with query video and videos from the database.

## 4. RESULT EVALUATION METHOD

The performance of video retrieval is evaluated with the same parameters as it is evaluated in image retrieval [11]. Recall and precision are the two parameters [2] as given in equations (9) and (10).

$$Recall = \frac{DC}{DB} \quad -(9)$$

$$Precision = \frac{DC}{DT} \quad -(10)$$

$DC = number\ of\ similar\ clips\ detected\ correctly$

$DB = number\ of\ similar\ clips\ in\ the\ database$

$DT = total\ number\ of\ detected\ clips$

Crossover points are calculated using the above mentioned two parameters to find the performance of the proposed system.

## 5. PROPOSED CBVR SYSTEM

A CBVR system is proposed in this paper in which multiple frames are obtained for the query videos and the videos' database instead of using single frame or key frames or all frames [2]. Features are extracted from these frames. The similar and most relevant videos are obtained from the output directory containing videos of that category. Significantly higher results have been obtained using this system. A typical methodology is used in this system where a video is retrieved from its category. Here, database is processed offline. The videos are represented by features extracted from their multiple frames. Features are then labelled and stored in the features database. An SVM is trained for the categories registered in the system using the labelled features stored in the database. Variables are obtained from the trained SVM. Features from the query videos are used for classification using SVM variables already saved. Videos obtained in the output folder are the videos of the desired category. For a query clip, videos stored in the given category can be ranked according to the distance measures and most similar videos are retrieved. The proposed system is shown in fig. 3. As mentioned above, multiple frames based classification and retrieval yields acceptable results without the complexity of finding key frames to represent a shot. A process flow of the proposed CBVR system is shown in fig. 3. Multiple frames are obtained during segmentation. Features are then extracted for each frame and stored in features database. Features are labelled for the pre-decided categories. SVM is trained and its variables are stored. This process is done offline. The query videos are separated into the categories based on stored SVM variables using features of the query videos. Videos obtained for different categories are stored with different categories in the output database.

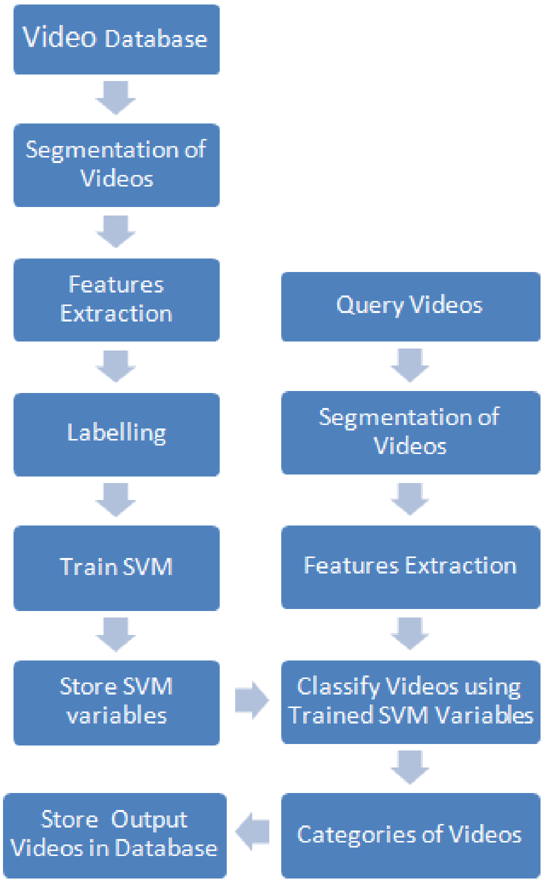


Figure 3. Proposed CBVR system

# 6. RESULTS

## 6.1 Database

The technique using multiple frames with Gabor features using SVM is applied to a video database having 1000 videos with 20 categories of 50 videos each as shown in fig. 4. Videos similar to the query video are stored in output folder after classification using SVM classifier. The precision and recall values are computed by grouping the number of classified videos belonging to the category of query video.

Figure 4. Video database of 1000 videos with 20 categories

## 6.2 Results

The charts shown below in fig.5 and fig.6 represent the retrieval results obtained for retrieving and classification of video clips from ten different categories. These categories are among the 20 categories of video clips from the video database of 1000 videos. The results obtained are much appreciable for all the categories but these ten categories of them are demonstrated here. The results are obtained using SVM based on features extracted using Gabor wavelet transform from multiple frames of video clips.

### *6.2.1 Results(Precision Values) for the video clips*

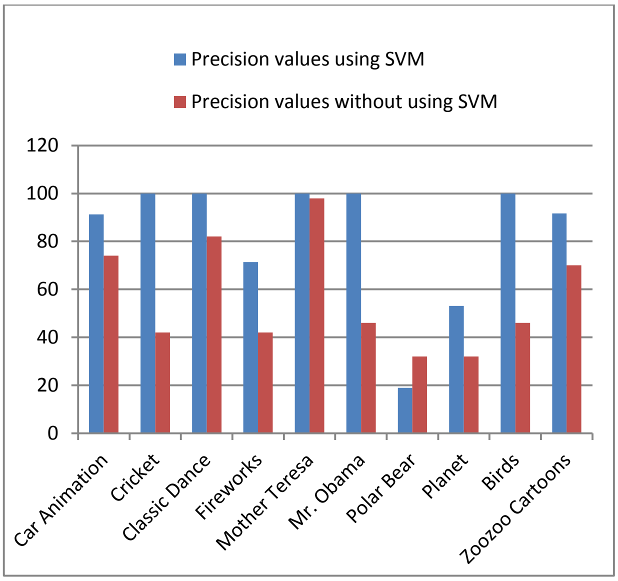


Figure 5. Comparison of Precision values shown for ten categories of videos using SVM and without using SVM using Gabor features

Fig.5 shows results (precision values) obtained by CBVR system based on Gabor features extracted from multiple frames using SVM. There is significant improvement in results using SVM as compared to results obtained without using SVM.

### *6.2.2 Results(Recall Values) for the video clips*

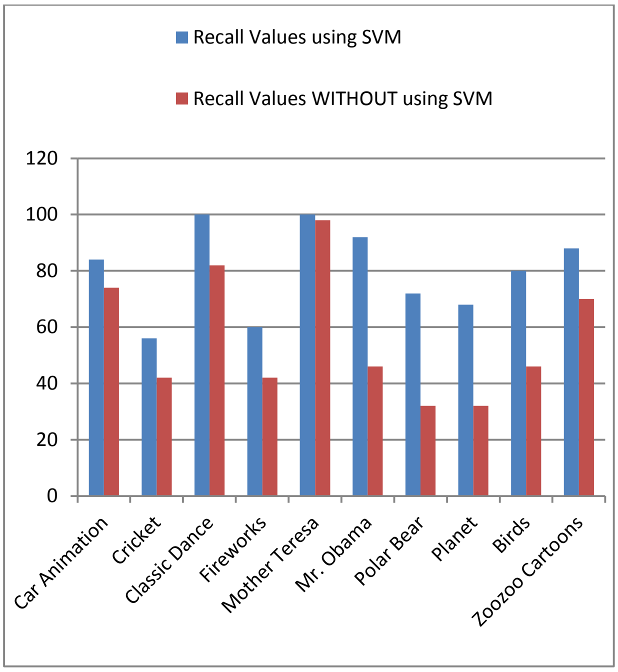


Figure 6. Comparison of Recall values shown for ten categories of videos using SVM and without using SVM using Gabor features

Fig.6 shows results (recall values) obtained by CBVR system based on Gabor features extracted from multiple frames using SVM. There is significant improvement in results using SVM as compared to results obtained without using SVM.

## 7. PROBLEMS AND CHALLENGES

The video content is represented by spatial and temporal characteristics of videos. In spatial domain, features are obtained from frames to form feature vectors from different parts of the frames. In temporal domain, video is segmented into its elements like frames, shots, scenes and video clips and features like histograms, moments, textures and motion vectors represent the information content of these video segments [7]. Drawback of techniques employing key frames matching is that temporal information and the related information between the key frames in a shot is lost. Content based video retrieval systems using query by image or query by clips using images or frames is implemented with low level features present in these images. Because of this, different objects present against similar backgrounds in frames belonging to different videos can yield confusing or false retrievals. Also, the low level features of the frames belonging to different videos can also yield false retrievals due to their corresponding low level features matching though use of SVM enhances result to a significant level.

## 8. CONCLUSION

It can be concluded from discussion in the previous sections that encouraging results are obtained and comparatively higher efficiency is achieved by using features in support vector machines from multiple frames instead of single frame or key frames representing a shot. Also, computational cost is lower for the system proposed here than that when using key frames to represent shots of a video. Query by example image is popular for content based image retrieval. Low level features are used for retrieval. The retrieval performance and the usefulness of these systems is restricted to the queries having distinct low level visual features but they do not address to the problems of video retrievals using semantic information for the query. Also, an efficient solution is needed to address the problems for the queries having similar backgrounds and showing confusing results. Automatic retrieval systems should be the focus and it requires more attention from researchers for improved retrieval results. A trend to reduce computational cost is needed to project commercialized systems for video indexing, classification and retrieval to facilitate the availability of low cost, fast and efficient CBVR systems. Capability of these systems can be magnified by reaching huge video databases that exist and are accessible on the web. The accessible databases should empower the users with options to accurately select the desired videos only while filtering out the relevant but undesired as well as irrelevant videos so that valuable, moral, ethical and informative data becomes accessible efficiently, quickly and at low cost.

## 9. REFERENCES

[1] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, Maybank S., "A Survey on Visual Content-Based Video Indexing and Retrieval", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41-6,797-819, 11/2011

[2] Liang-Hua Chen, Kuo-Hao Chin, Hong-Yuan Liao, "An integrated approach to video retrieval", Proceedings of the nineteenth conference on Australasian database-Volume 75, 49–55, 2008

[3] Dengsheng Zhang, Aylwin Wong, Maria Indrawan, Guojun Lu,"Content-based Image Retrieval Using Gabor Texture Features",IEEE Transactions PAMI,pages 13-15, vol. 12.

[4] Yining Deng, B.S. Manjunath, "Content-based Search of Video Using Color, Texture, and Motion", IEEE, pg 534-537, 1997

[5] Ja-Hwung Su, Yu-Ting Huang, Hsin-Ho Yeh, Vincent S. Tseng , "Expert Systems with Applications", 37, pg 5068-5085, 2010

[6] Nicu Sebe, Michael S. Lew, Arnold W.M. Smeulders, "Video retrieval and summarization", Computer Vision and Image Understanding, vol. 92, no. 2-3, pg 141-146, 2003

[7] C. V. Jawahar, Balakrishna Chennupati, Balamanohar Paluri, Nataraj Jammalamadaka, "Video Retrieval Based on Textual Queries", Proceedings of the Thirteenth InternationalConference on Advanced Computing and Communications, Coimbatore, Citeseer, 2005

[8] Alexander G. Hauptmann, Rong Jin, and Tobun D. Ng, "Video Retrieval using Speech and Image Information", Electronic Imaging Conference (EI'03), Storage Retrieval forMultimedia Databases, Santa Clara, CA, January 20-24, 2003.

[9] Swain M.J. and Ballard, B.H. "Color Indexing," Int'l J. Computer Vision, vol. 7, no. 1, pp. 11-32, 1991.

[10] Hafner, J. Sawhney, H.S. Equitz, W. Flickner, M. and Niblack, W. "Efficient Color Histogram Indexing for Quadratic Form Distance," IEEE Trans. Pattern Analysis and Machine Intelligence, 17(7), pp. 729-736, July, 1995.

[11] Aljahdali, S., Ansari, A., Hundewale, N., "Classification of Image Database Using SVM With Gabor Magnitude", International Conference on Multimedia Computing and Systems (ICMCS), 2012 , vol., no., pp.126,132, 10-12 May 2012

[12] Aasif Ansari, Muzammil H. Mohammed,“Content Based Video Retrieval Systems - Methods, Techniques, Trends and Challenges”, International Journal of Computer Applications (ISBN : 973-93-80885-36-9),Volume 112 – No. 7, February 2015

[13] R. Lienhart, “A System For Effortless Content Annotation To Unfold The Semantics In Videos,” in Proc. IEEE Workshop Content-Based Access Image Video Libraries, pp. 45–49, Jun. 2000.

[14] C. Faloutsos, R. Barber, M. Flicker, J. Hafner, W. Niblack, D. Petkovic and W. Equitz, "Efficient and effective querying by image content", J. lntelL Inf. Systems 3, 231-262, 1994.

[15] Visionics Corporate Web Site, FaceIt Developer Kit Software, http://www.visionics.com, 2002.

[16] The TREC Video Retrieval Track Home Page, http://www-nlpir.nist.gov/projects/trecvid/

[17] Arti Khaparde, B. L. Deekshatulu, M. Madhavilath, Zakira Farheen, Sandhya Kumari V, "Content Based Image Retrieval Using Independent Component Analysis”, IJCSNS International Journal ofComputer Science and Network Security, VOL.8 No.4, April 2000.

[18] H.B. Kekre, V.A. Bharadi, S.D. Thepade, B.K. Mishra, S.E. Ghosalkar, S.M. Sawant, "Content Based Image Retreival Using Fusion of Gabor Magnitude and Modified Block Truncation Coding," icetet, pp.140-145, 2010 3rd International Conference on Emerging Trends in Engineering and Technology, 2010.

[19] Flickner M. et al, “Query by image and video content: the QBIC system”, IEEE Computer 1995, Volume 28, Number 9, pp 23-32, 1995.

[20] Sinora Banker Ghosalkar, Vinayak A.Bharadi, Sanjay Sharma, Asif Ansari, "Feature Extraction using Overlap Blocks for Content based Image Retreival" International Journal of Computer Applications (0975-8887), Volume 28-No.7, August 2011.

[21] Markos Zampoglou, Theophilos Papadimitriou, IEEE Member, and Konstantinos I. Diamantaras, IEEE Member, "Support Vector Machines Content-Based Video Retrieval basedsolely on Motion Information", IEEE, ISSN : 1551-2541, Print ISBN: 978-1-4244-1566-3, 2007.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 8, 614 - 617, 2015, ISSN: 2319–8656

# Review of the Introduction and Use of RFID

Fariba Ghorbany Beram
Sama Technical and Vocational Training College Islamic Azad University, Shoushtar Branch, Shoushtar, Iran

Mojtaba Khayat
Sama Technical and Vocational Training College Islamic Azad University, Shoushtar Branch, Shoushtar, Iran

Sajjad Ghorbany Beram
Sama Technical and Vocational Training College Islamic Azad University, Shoushtar Branch, Shoushtar, Iran

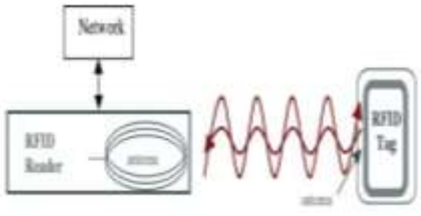
**Abstract**: We live in an age where technology is an integral part of human daily life. The aim of the technology is secure, accurate, correct use of time for mankind and nature brings it may also have disadvantages and challenges. In this paper we describe the RFID technology. Today, this technology in hospitals, shops, airports, tracking birds are used. It can be used in hospital intensive care unit for monitoring and remote patient care, which causes it to print patients and physicians are not required to comply with very close distance, the result of which can cause safety Patients and physicians should be. The security technology implementation is a challenge. This paper introduces the applications, the challenges we have this technology.

**Keywords:** Radio Frequency identification; tag; network; challenge

## 1. INTRODUCTION

Radio Frequency identification (RFID) is the popular wireless induction system [1-7]. RFID tags were initially developed as very small electronic hardware components having as their main function to broadcast a unique identifying number upon request. The simplest types of RFID tags are passive devices that not have an internal power source and are incapable of autonomous activity. They are powered by the reader’s radio waves, with their antenna doubling as a source of inductive power. While admittedly a new technology, the low-cost and high convenience value of RFID tags gives them the potential for massive deployment, for business automation applications and as smart, mass-market, embedded devices that support ubiquitous applications. However, current RFID protocols are designed to optimize performance, with lesser attention paid to resilience and security. Consequently, most RFID systems are inherently insecure. The general design of a simple RFID system is displayed through the following figure:

Figure 1: Diagram describing operation of the RFID system.



## 2. RFID TECHNOLOGY

A typical deployment of an RFID system involves three types of legitimate entities, namely tags, readers and back-end servers. The tags are attached to, or embedded in, objects to be identified. They consist of a transponder and an RF coupling element. The coupling element has an antenna coil to capture RF power, clock pulses and data from the RFID reader. The readers typically contain a transceiver, a control unit, and a coupling element, to interrogate tags. They implement a radio interface to the tags and also a high level interface to a backend server that processes captured data. The back-servers are trusted entities that maintain a database containing the information needed to identify tags, including their identification numbers. Since the integrity of an RFID system is entirely dependent on the proper behavior of the server, it is assumed that the server is physically secure and not attackable. It is certainly legitimate to consider privacy mechanisms that reduce the trust on the back-end server—for instance, to mitigate the ability of the server to collect user-behavior information, or to make the server function auditable[8]. In this paper, however, we shall not investigate such privacy attacks. Here we shall consider the servers to be entirely trusted following:

- Signal strength limited to a required

distance;

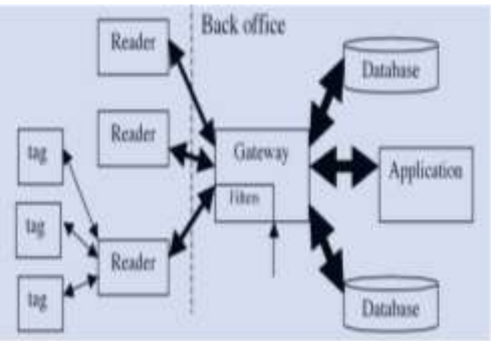
- Radio frequency unable to work in certain

geographical areas;

- Electromagnetic field prone to interruption from solar and electrical storms.

The medical and healthcare sectors are using this technology as a means to keep track of medical equipment and the delivery of pharmaceuticals that proved to be costly in the past when it came to tracking them down. But now this technology is being applied as a means of protecting online medical information systems from those who are involved in perpetrating the criminal activity of medical identity theft. Because the identity of every individual is centred around recognising specific features, personality, knowledge and traits that define who we are requires having the implementation of a robust identity management system that can be used in determining our individual specific features and characteristics through the provision of unique identifiers in recognising us[9,10]. A general RFID architecture is depicted in Figure 2.

Figure 2: A general RFID architecture



### 2.1 Classification of technologies

around RFID We identify three classes to order the relation

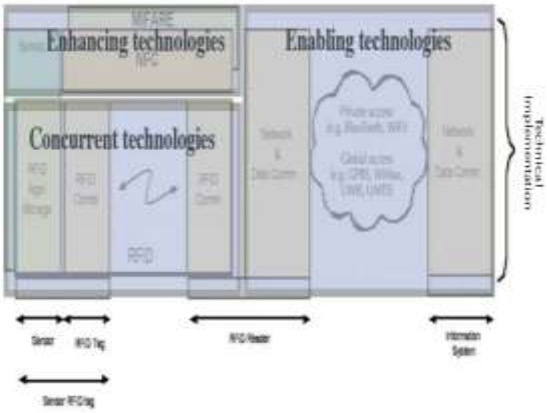
between RFID and network technologies:

- Enabling technology
- Enhancing technology
- Concurrent technology

In Figure 3 these classes are mapped onto

the RFID will be used to characterize the different technologies.

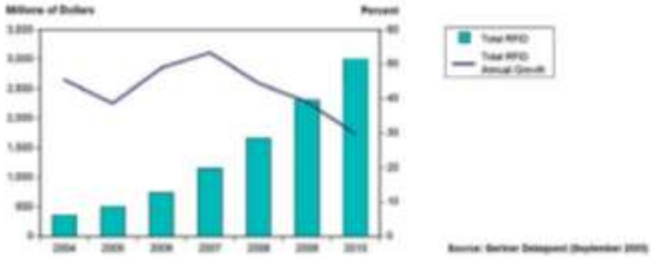
Figure 3: Classification of technologies around RFID



## 3. APPLICATIONS

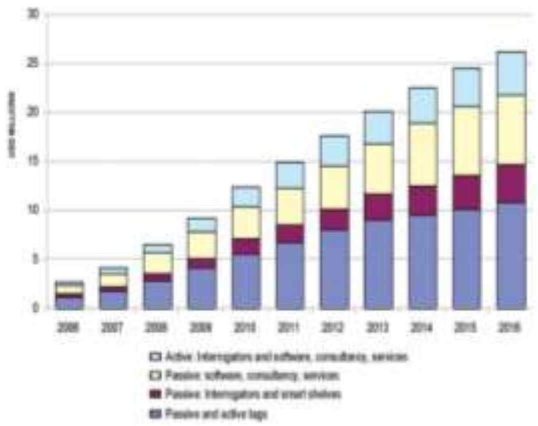
The deployment of a pervasive RFID-based infrastructure in an everyday environment holds the promise of enabling new classes of applications that go beyond tracking and monitoring. Such a system could, for example, support logging and analysis of individual tag movements over time, allowing a user to ask questions such as ”how often do I get interrupted in my office on an average day?”. Current and historical data on groups of tags could also be used to identify and analyze aggregate phenomena such as the impact of seminars on improving communication between researchers. Real-time streams of tag reads could be used in “find my object” applications, reminding services [11,12], and to actuate devices. Enabling these classes of application, however, presents a significant challenge to the system design. First, the system must be able to consistently and accurately read tags and it must do so at a granularity sufficient for the intended applications. The system must support archiving and retrieval of tag reads and real-time reporting of new reads. To enable the monitoring of large spaces with a possibly large number of tags, the system must scale to handle highvolume streams of tag reads. Finally, because such a system will manage large amounts of (potentially sensitive) personal data from multiple users, it must be secure and support an appropriate privacy model. In the following section, we present our preliminary system architecture and discuss how these application requirements affected its design. As an example Gartner (Gartner, 2005b) claims that the RFID market will experience an annual growth rate somewhere between 30% and 50% in the years 2004-2010 ending up at 3 billion USD in 2010 (see Figure 5-7). They say this refers to “the use of RFID technologies within a supply chain environment to improve the visibility, management and security of cargo shipments or supply chain assets, such as conveyances or valuable mobile assets. The applications and hardware that are used outside of the above environment were excluded. These would include consumer uses, such as contactless smart cards.”

Figure 4: RFID, worldwide size and growth, 2004-2010



The IDTechEx report provides some more data as depicted in Figure 5.

Figure 5: Total RFID market projections 2006-2016



## 4. PRIVACY CHALLENGES

Pervasive RFID-based deployments raise privacy concerns because they can enable the tracking of people and personal objects by parties that would otherwise be unable or unauthorized to do so. These concerns involve: the physical security of the communication between tags and readers, the security of the data stored in and processed by the system, and controlled access to the data. In this study, we focused on the latter problem and studied in-situ many of the privacy concerns experienced by the participants. Chief among the participants' concerns was the perceived ease with which one's activities could be inferred from the data (e.g., time of day, direction of movement, and set of tags seen). We were able to validate this concern by writing a simple script that could detect lunch breaks with better than 75% accuracy for three of the participants, showing that participant P1 took 29 minute lunch breaks on average, P2 took about 32 minutes, and P3 took the longest (40 minute) breaks. Similarly, it was easy to infer potentially more sensitive information such as when and how many times a participant used the restroom in a day. Our initial approach to addressing these privacy concerns was to allow display and deletion of one's personal data via the web interface. The limitation of this approach is that a user can still see another's data before that data is deleted. A more appropriate option would be for users to specify high-level rules that describe which TREs should be accessible to which users and which TREs should be dropped automatically (e.g., all trips from my office to the restroom shorter than 2 minutes)[13]. For example, a query on a colleague's location could return approximate information (e.g., 4th floor) by default, or more exact information (e.g., room 490) only a few times per day. We are currently exploring the suitability of such techniques for various applications. Finally, to protect the privacy of non-participants, each Node Server automatically discards any tag reads for a tag that is not registered in our database[14].

## 5. CONCLUSION

In this paper, we motivated the benefits of RFID. As mentioned, this technology has many applications is. Studies show that in the near future this technology will be increasingly used and welcomed. By examining the current challenges in the technology and resolving the problems with reliability Khatrbyshtr can use this technology.

## 6. REFERENCES

[1] Ajay Malik, "RTLS for Dummies," publishing by Wiley publishing. Inc., USA, Indianapolis, Indiana, (2009) ISBN: 978-0-470-39868-5.

[2] J. Zhou, and J. Shi "RFID localization algorithms and applications a review," The International Journal of Intelligent Manufacturing, Vol. 20 (6), Springer Netherlands, pp. 695-707, 2008.

[3] R. Want, "An Introduction to RFID Technology," IEEE Pervasive Computing, Vol. 5, pp. 25-33, 2006.

[4] G. Barber, and E. Tsibertzopoulos, "An Analysis of Using EPCglobal class-1 generation-2 RFID Technology for Wireless Asset Management," IEEE Military Communications Conference (MILCOM 2005), Vol. 1, pp. 245-251, October 2005.

[5] K. Ahsan, H. Shah, and Paul Kingston, "RFID Applications: An Introductory and Exploratory Study," IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 3, January 2010.

[6] Benjamin D. Braaten, and Robert P. Scheeler, "Design of Passive UHF RFID Tag Antennas Using Metamaterial-Based Structures and Techniques," Radio Frequency Identification Fundamentals and Applications, Design Methods and Solutions, Book edited by: Cristina Turcu, pp. 324, February 2010,ISBN 9789537619725.

[7] Daniel M. Dobkin, "The RF in RFID: Passive UHF RFID in Practice," Chapter 3: Radio Basics for UHF RFID, (2007), ISBN: 9780750682091.

[8] Smith, J. R. RFID-based techniques for human-activity detection. Communications of the ACM, 48(9), Sept. 2005.

[9] Songini, M. L. Wal-Mart details its RFID journey. ComputerWorld, Mar. 2006.

[10] Stanford, V. Pervasive computing goes the last hundred feet with RFID systems. IEEE Pervasive Computing, 2(2), Apr. 2003.

[11] Sweeney, L. k-anonymity: A model for protecting privacy. IJUFKS, 10(5):557–570, Oct. 2002.

[18] Want, R. The magic of RFID. ACM Queue, 2(7), Oct. 2004.

[12] Want, R. et. al. An overview of the PARCTAB ubiquitous computing experiment. IEEE Personal Communications, 2(6):28–33, Dec 1995.

[13] Want, R. et. al. Bridging physical and virtual worlds with electronic tags. In CHI, pages 370–377, 1999.

[14] E. Wu, Y. Diao, and S. Rizvi. High-performance complex event processing over streams. In Proc. of the 2006 SIGMOD Conf., June 2006

International Journal of Computer Applications Technology and Research
Volume 4– Issue 8, 618 - 622, 2015, ISSN: 2319–8656

# Enhanced Quality of Service Based Routing Protocol Using Hybrid Ant Colony Optimization and Particle Swarm Optimization

Neelam Kumari
Dept. Computer Science and Engineering
Beant College of Engineering & Technology

Arpinder Singh Sandhu
Dept. Computer Science and Engineering
Beant College of Engineering & Technology

## Abstract

The main problem of QoS routing is to setup a multicast hierarchy that may meet particular QoS constraint. In order to reduce the constraints of the earlier work a new improved technique is proposed in this work. In the proposed technique the issue of multi-cast tree is eliminated using clustering based technique. First of all multi-radio and multichannel based clustering is deployed and these cluster head are responsible for the multicasting. It will diminish the overall energy consumption of nodes and complexity of intelligent algorithms. The path will be evaluated based upon the ant colony optimization. Thus it has produced better results than other techniques.

**Keywords:** QoS, Multicast, Ant colony optimisation, clustering.

## 1.Introduction

A Mobile Ad-hoc Network is an accumulation of independent mobile nodes that can communicate together via Radio Lake. Your mobile nodes which has been in radio selection of each various other could right communicate, whereas others needs the aid of intermediate nodes to route his or her packets. The entire node carries a radio user interface to connect jointly. These networks usually are fully distributed, and perform at any place without the aid of any fixed infrastructure as gain access to points or base areas. Figure 1 shows a simple ad-hoc network having 3 nodes. Node 1 in addition to node 3 isn't within range of each and every other, however the node 2 enables you to forward packets between node 1and node a couple of. The node 2 will behave as a router in addition to these three nodes in concert form an ad-hoc system.

.

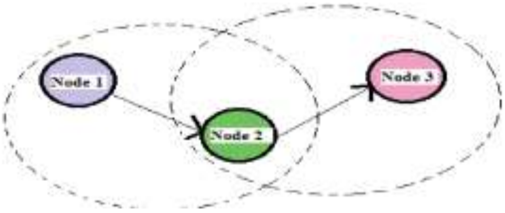


**Fig. 1 Example of mobile ad-hoc network**

## 1.1 MANETs characteristics

**1) Dispersed operation:** There isn't background network for that central control in the network operations; the control in the network is distributed one of the nodes. The nodes associated with a MANET should cooperate with each other and communicate among themselves and each and every node acts to be a relay as essential, to implement specific functions for example routing and security.

**2) Multi get routing**: When a node endeavours to send facts to other nodes that is out of its communication range, the packet the packet need to be forwarded via one or more intermediate node.

3) Autonomous fatal: In MANET, each mobile node is an independent node, which will function as both a host and a router.

**4) Vibrant topology:** Nodes are unengaged to move arbitrarily using different speeds; so, the network topology might change randomly and also at unpredictable time. The nodes in the MANET dynamically determine routing among themselves since they travel around, establishing their unique network.

**5) Light-weight terminals**: Within maximum cases, the actual nodes at MANET are generally mobile with much less CPU capability, minimal power storage and also small memory measurement.

**6) Shared Physical Medium:** The wireless communication medium is obtainable to any entity with all the appropriate equipment and also adequate resources. Keeping that in mind, access to the channel is not restricted.

### Advantages of MANET

Why people love an Ad-Hoc network add the following:

• They feature access to information and services no matter geographic position.

• Liberty from central circle administration. Self-configuring circle, nodes are also behave as routers. Less expensive compared to wired network.

• Scalable—accommodates the particular addition of additional nodes.

- Much better Flexiblibility.
- Robust on account of decentralize administration.
- The network might be set up at any place and time.

## 1.1.1 MANETs Challenges

**1)Limited bandwidth:** Wireless link keep have significantly reduced capacity than infrastructure networks. In addition, the realized throughput involving wireless communication after accounting for that effect of a number of accesses, fading, noise, and interference problems, etc., is often a reduced amount of than a radio’s highest transmission rate.

2) Energetic topology: Dynamic topology member's program may disturb the particular trust relationship amongst nodes. The trusts are often disturbed if a number of nodes are recognized as compromised.

3) Course-plotting Overhead: In Wi-Fi ad-hoc networks, nodes frequently change their area within network. And so, some stale routes are generated from the routing table that leads to unnecessary direction-finding overhead.

4) Undetectable terminal problem: The hidden terminal problem describes the collision of packets with a receiving node a result of the simultaneous transmission of the nodes that aren't within the direct transmission selection of the sender, but are within the transmission range on the receiver.

5) Bundle losses on account of transmission mistakes: Ad hoc Wi-Fi networks experiences a far more achievable bundle loss due to factors like while increased collisions a consequence of the presence of cannot be seen terminals, presence involving disturbance, unidirectional links, frequent way breaks due to mobility associated with nodes.

6) Mobility-induced approach changes: The system topology inside the ad hoc Wi-Fi network can be highly dynamic a consequence of the activity of nodes; for that reasons an on-going interval suffers typical path pauses. This situation often leads to frequent way alterations.

7) Battery demands: Devices used throughout these networks have restrictions for the power source so as to maintain portability, size and weight on the device. 8) Security threats: The wireless mobile random nature of MANETs provides new security challenges for the network design. For the reason that wireless medium is at risk of eavesdropping and random network functionality is made through node assistance, mobile ad hoc networks are intrinsically confronted with numerous security attacks.

## 1.1.1.1 MANETs Applications

1) Military battlefield: Ad-Hoc networking will allow the military to reap the benefits of commonplace network technology to help keep an information network relating to the soldiers, vehicles, and military information brain quarter.

Unknown terrain
Limit the Range of communication
Directional Antennas
Destroyed infrastructure

2) Collaborative perform: For some enterprise environments, the need regarding collaborative computing may very well be more important outside office environments when compared with inside and in which people do must have outside meetings to help cooperate and exchange information on a given challenge.

3) Local levels: Ad-Hoc networks can autonomously link instantaneously and temporary media network using notebook computers to spread and share information among participants for a e. g. meeting or classroom. Another appropriate community level application may very well be in home networks where devices can communicate straight to exchange information.

4) Personalized area network and bluetooth: A personal area network is a short range, localized network in which nodes are usually associated with a given person. Short-range MANET like Bluetooth can de-stress the inter verbal exchanges between various mobile devices as being a laptop, and also a mobile phone.

5) Commercial Market: Ad hoc may be used in emergency/rescue procedures for disaster relief efforts, e. gary the gadget guy. in fire, avalanche, or earthquake. Emergency rescue operations must be held where non-existing or even damaged communications national infrastructure and rapid deployment of any communication network is necessary.

## 1.1.1.1.1 SAFETY GOALS in MANET

All networking functions for instance routing and bundle forwarding, are performed by nodes themselves in a very self-organizing manner. Therefore, securing a cell ad -hoc network is quite challenging. The goals to evaluate if mobile ad-hoc circle is secure or even not are as follows:

1)Availability: Availability suggests the assets are generally accessible to sanctioned parties at suitable times. Availability applies both to data in order to services. It makes sure the survivability associated with network service despite denial of services attack.

2) Privacy: Confidentiality ensures in which computer-related assets are generally accessed only by authorized parties. Protection of information which is exchanging through any MANET. It ought to be protected against any disclosure attack including eavesdropping- unauthorized examining of message.

3) Honesty: Integrity means that assets is usually modified only by authorized parties or even only in sanctioned way.. Integrity assures which a message being transported is never damaged.

4) Authentication: Authentication is defined as assurance men and women in communication are authenticated in lieu of impersonators. The recourses involving network needs to be accessed because of the authenticated nodes.

5) Acceptance: This property assigns distinct access rights to different types of users. For example any network management can be carried out by network officer only.

6) Strength to attacks: It must sustain the network functionalities whenever a portion of nodes is actually compromised or ruined.

7) Freshness: It ensures that malicious node doesn't resend previously seized packets.

## 1.1.1.1.1.1 ROUTING PROTOCOLS

Ad-Hoc circle routing protocols are commonly divided into about three main classes; Proactive, reactive and cross protocols as found in figure 2.

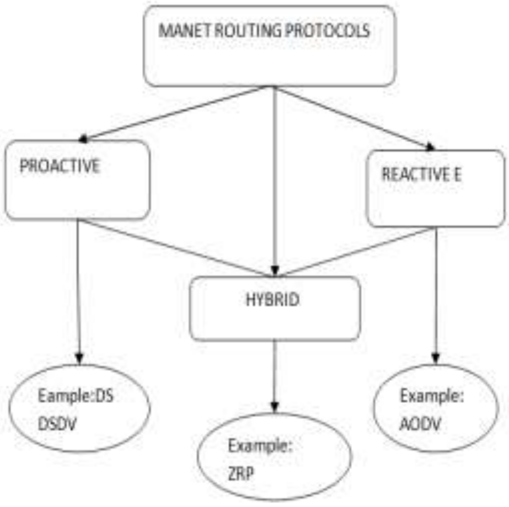


**Fig. 2 Classification of MANET routing protocol**

Proactive Protocols: Proactive, or table-driven routing protocols. In positive routing, each node needs to maintain a number tables to shop routing information, and any alterations in network topology have to be reflected by propagating updates through the entire network to be able to maintain a regular network view. Example of such schemes include the conventional routing techniques: Destination sequenced mileage vector (DSDV). They attempt to maintain consistent, up-to-date routing information in the whole network. It minimizes your delay in communication and invite nodes to rapidly determine which nodes are present or reachable within the network.

Example: Destination-sequenced distance vector (DSDV), Wireless routing protocol (WRP), Global state routing (GSR),OLSR (Optimized Link State Routing)

Reactive Protocols: Reactive routing is also called on-demand routing protocol since they just don't maintain routing data or routing activity on the network nodes when there is no communication. If a node would like to send a packet completely to another node then this protocol pursuit of the route within the on-demand manner and establishes the connection to be able to transmit and receive the packet. The route finding occurs by inundating the route request packets through the entire network. Examples of reactive routing protocols include the Ad-hoc On-demand Range Vector routing (AODV) and Dynamic Source routing (DSR).

Hybrid Protocols: They features a hybrid type that combines reactive and proactive routing standards. The Zone Course-plotting Protocol (ZRP) is often a hybrid routing process that divides your network into areas and specific zones. ZRP provides any hierarchical architecture where by each node needs to maintain additional topological data requiring extra memory space. Example: Zone routing protocol (ZRP),Distributed dynamic routing (DDR)

2. Related Work

Rajeev Agrawal (2001) [1] has adopted probabilistic modeling to model the effect due to multipath fading and shadowing. The BER for each link affected by the fading is estimated using the proposed model. Wireless Routing Protocol (WRP) maintains the BER associated with a particular link, a packet/ data is routed with optimum BER route from a set of discovered route by protocol. B.Malarkodi et al. (2009) [2] the impact of different mobility models on Multicast Routing Protocols. The results showed that the throughput of ADMR is higher than of ODMRP at high mobility. This is achieved at the cost of increase in delay and transmission over head. Under low mobility, ODMRP has higher throughput than AMDR.

V.A Gajbhiye and R. W.Jasutkar(2013) [3] showed that Swarm Intelligence based routing protocol has shown promising results in VANET. For this they compared and evaluated the performance of AODV, OLSR, and Swarm Intelligence based routing protocol in terms of throughput, latency and data packet delivery ratio for VANET. Simulation results have shown that SWARM Intelligence based routing protocol showed promising results in VANETs as compared to AODV and OLSR.

Nathaniel Gemelli et al. (2003) [4] Introduced Bluetooth wireless technology, examine current routing protocols and present the objectives and considerations for the design of a new Bluetooth routing protocol. The protocol design would consider the capabilities of the devices (nodes) within the range of the network. It was envisioned that capabilities Aware Routing (CAR) protocol would make routing decisions based on such. Factors as device power constraints E.Ahila Devi and K.Chitra(2014 )[5] Introduced a Privacy Protecting Secure and Energy Efficient Routing Protocol (PPSEER) was proposed. In this protocol, first the classifications of network node take place based on their energy level.

Hiba Hachichi et al. (2011) [6] created and maintained locally a hierarchy that was well suitable for routing packets in an Ad hoc network. The contribution of this work was mainly based on the construction of a virtual topology where cluster heads and gateways collaborate for searching the destination node.

Istikmal et al. (2013) [7] presented about investigation result of AODV, DSR and DSDV that applied an Ant-algorithm which were AODV-Ant, DSR-Ant, and DSDV-Ant. DSDV represents of proactive routing type protocol based on table driven, while AODV and DSR represents of reactive routing protocol type based on demand. Performance analysis included end to end delay, throughput, routing overhead and hop count for various scenario of node velocity, pause time and network traffic.

Sikkandar Ali and Vashik Ali et al.. (2012)[8] presented routing in wireless mobile ad-hoc networks using Destination Sequenced Distance Vector (DSDV) and Ad-hoc on demand Distance Vector (AODV) protocols. The performance of bandwidth, throughput and packet loss of DSDV and AODV has been modelled under various network configurations and mobility conditions..

Geethu Mohandas(2013) [9] The Mobile Ad hoc Networks (MANET) are networks with self-configuring capacity of mobile devices interconnected by wireless links. During the last few years, research in various aspects of MANETs has been prominent, prompted mainly by military, disaster relief, and law enforcement scenarios. An instinctive footstep was to take up such location-based operation to MANETS.

KomalPatel et al.(2006)[10] proposed a cross layer approach that uses the MAC layer link stability information to improve the routing efficiency. Signal strength of the link was captured from the MAC Layer and used at network layer to predict the future signal strength value using double exponential smoothing model. This information was used to categorize the link as stable or unstable.

Rashmi Rohankar et al. (2012) [11] analyzed the effect of random based mobility models on the performance of Proactive Routing Protocol (DSDV Destination Sequence Distance Vector) and Reactive Routing Protocol (AODV- on Demand Distance Vector, DSR- Dynamic Source Routing). Performance analysis was done with respect to end-to-end delay, throughput and Packet delivery ratio for varying node densities.

Yudhvir Singh(2010)[12] performed simulation based experiments were performed to analyzed the performance of On Demand Multicast Routing Protocol by evaluating Packet Delivery Ratio, End to End delay and average throughput. These results were compared with AODV and FSR routing protocols by varying number of nodes and mobility. The comparison showed that ODMRP for ad hoc networks performs better as compared to AODV and FSR.

3. Proposed Technique

Following are the various steps required to successfully simulate the proposed algorithm.

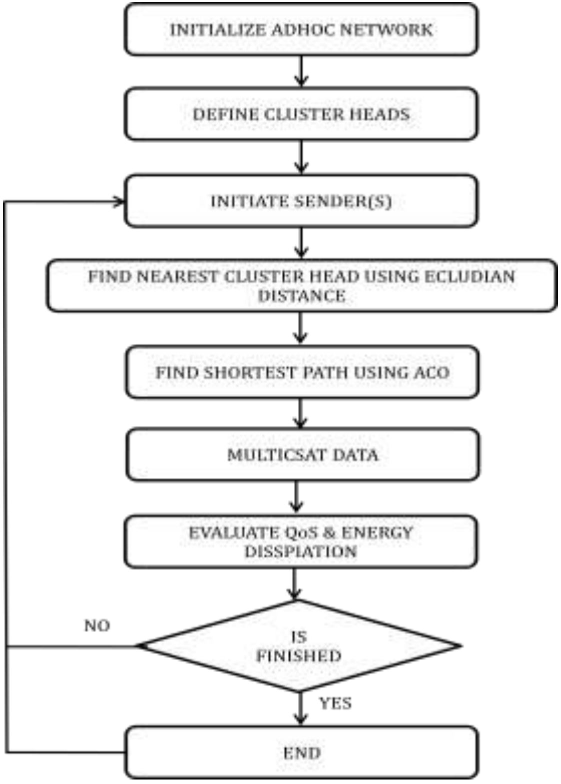

INITIALIZE ADHOC NETWORK
DEFINE CLUSTER HEADS
INITIATE SENDER(S)
FIND NEAREST CLUSTER HEAD USING ECLUDIAN DISTANCE
FIND SHORTEST PATH USING ACO
MULTICSAT DATA
EVALUATE QoS & ENERGY DISSPIATION
NO
IS FINISHED
YES
END

Figure 1: Flowchart of the proposed technique

**Step 1:** First of all initialize ad-hoc network with their respective characteristics like moving range, maximum dimensions, number of nodes etc.

**Step 2:** Define cluster heads having multi-radio and multi-channel facility.

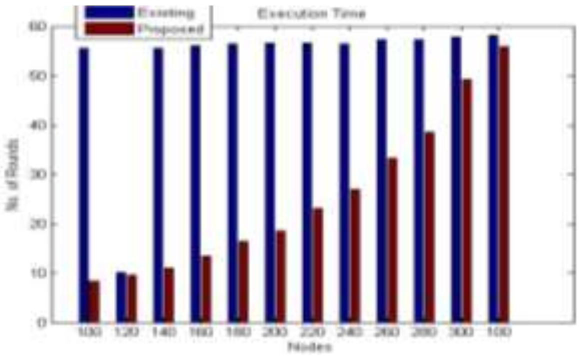
**Step 3:** Sender(s) will be initiated to multicast its data to defined nodes.

**Step 4:** Sender will hand over its data to nearest cluster head using Euclidian distance.

**Step 5:** Cluster head will multicast data to available cluster heads depends upon the ACO based shortest path.

**Step 6:** Evaluate energy dissipation as well as other QoS features, and move to step 3

## 4. Result Analysis



**fFig1: Execution Time**

The Figure represent the information about the Execution time of Existing and Proposed Technique X-axis signify the value of number of Nodes and Y-axis correspond to the value of Number of Rounds. Moreover, in which two type of color are used i.e Red and Blue. Red color symbolize the Proposed technique and Blue color symbolize the Existing Technique.

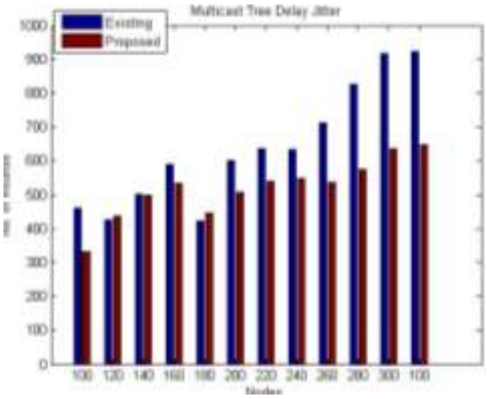


**Fig2: Multicast Tree Delay Jitter**

The Diagram represent the information about the Multicast Tree Delay Jitter of Existing and Proposed Technique X-axis signify the value of number of Nodes and Y-axis correspond to the value of Number of Rounds. Moreover, in which two type of color are used i.e Red and Blue. Red color symbolize the Proposed technique and Blue color symbolize the Existing Technique.

## 5. Conclusion

The main problem of QoS routing is to setup a multicast hierarchy that may meet particular QoS constraint. Nevertheless, the situation of making a multicast tree below several constraints is available to be NP Complete. Therefore, the issue is often settled by heuristics or smart optimization. Lately, some meta-heuristic algorithms including the ant colony algorithm, genetic algorithm and compound swarm optimization have been employed by the analysts to eliminate the multi-constrained QoS routing problem. In order to reduce the constraints of the earlier work a new improved technique is proposed in this work. In the proposed technique the issue of multi-cast tree is eliminated using clustering based technique. First of all multi-radio and multichannel based clustering is deployed and these cluster head are responsible for the multicasting. It will diminish the overall energy consumption of nodes and complexity of intelligent algorithms. The path will be evaluated based upon the ant colony optimization. Thus it has produced better results than other techniques.

This work has not considered the effect of node failures on the network. Therefore in near future we will evaluate the node failures while data communication is in progress.

## 6. REFERENCES

[1] Agrawal, Rajeev. "Performance of routing strategy (bit error based) in fading environments for mobile adhoc networks." IEEE International Conference on *Personal Wireless Communications*, pp. 550-554, 2005.

[2] Malarkodi, B.P. Gopal and B.Venkataramani. "Performance Evaluation of Adhoc Networks with Different Multicast Routing Protocols and Mobility Models." *IEEE International Conference on* advanced in recent technologies in communication and computing, pp. 81-84, 2009.

[3] Gajbhiye, V. A and R. W. Jasutkar. "Biologicaly inspired routing protocol for vehicular ad hoc network."6th IEEE *International Conference on Advanced Infocomm Technology*, pp. 202-206, 2013.

[4] Gemelli, Nathaniel, Peter LaMonica,Paul Prtzke, John Spina."Capabilities aware routing for dynamic ad hoc networks." IEEE *International Conference on Integration of knowledge Intensive Multi-Agent Systems*, pp. 585-590, IEEE, 2014. 2003.

[5] Devi, E.Ahila, and K.Chitra. "Security based energy efficient routing protocol for Ad hoc network." *IEEE International Conference on Control, Instrumentation, Communication and Computational Technologies*, pp. 1522-1526, 2014.

[6] Hachichi, Hiba, Samia Chelloug and Fatima Athmouni. "A virtual topology for routing in adhoc networks." *IEEE International* Conference on Electronics, Communication & Phtonics, 2011.

[7] Ali, Sikkandar AliVashik, W.R.Salem Jeyaseelan and Shanmugasundaram Hariharan. "Enhanced Route Discovery in Mobile Adhoc Networks." *IEEE Third International Conference on.* Computing Communication and Networking Technologies, pp. 1-5,2012.

[8] Istikmal,Leanna, V. Y and Basuki Rahmat. "Comparison of proactive and reactive routing protocol in mobile ad hoc network based on "Ant-algorithm." *IEEE International Conference on Computer, Control, Informatics and its Applications*, 2013.

[9] Mohandas, Geethu, Salaja Silas and Shini Sam. "Survey on routing protocols on mobile ad hoc networks." *IEEE International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing*, pp. 514-517,2013.

[10] Patel, Komal, S. Srivastava and R. B. Lenin. "MAC Layer aware Stable Link Routing (MACSLR) in Mobile Ad hoc Networks." IEEE *International Symposium on Ad hoc and Ubiquitous Computing*, pp. 298-301, 2006.

[11] Rohankar, Rashmi, Rinkoo Bhatia and Deepak Kumar Sharma."Performance analysis of various routing protocols (proactive and reactive) for random mobility models of Ad hoc networks."Ist IEEE *International Conference on Recent Advance in Information Technology*, 2012.

*[12]* Yudhvir Singh, Yogesh Chaba, Monica Jain and Praba Rani. "Performance Evaluation of On-Demand Multicasting Routing Protocols in Mobile Adhoc Networks." IEEE, *International Conference on Recent Trend in Information, Telecommunication & Computing*, pp.298-301, 2010.

# Internal Architecture of Junction Based Router

Tulikapriya Sinha
Symbiosis Institute of Technology,
Symbiosis International University,
Pune, India

Shraddha Patil
Symbiosis Institute of Technology,
Symbiosis International University,
Pune, India

Smita Khole
Symbiosis Institute of Technology,
Symbiosis International University,
Pune, India

**Abstract**: The router is an important component in NoC as it provides routes for the communication between different cores. A router consists of registers, switches, arbitration and control logic that collectively implement the routing and flow control function required to buffer and forward flits to their destination. This router will be implemented on FPGA using Spartan-3 kit. This paper describes the internal blocks of a junction based router and there operation.

**Keywords**: Router, NoC, FPGA, Verilog, arbiter.

## 1. INTRODUCTION

### 1.1Network on Chip (NoC)

Network on chip or network on a chip is a communication subsystem on an integrated circuit (commonly called a "chip"), typically between IP cores in a System on a Chip (SoC). Network-on-Chip (NoC) architectures provide a good way of realizing efficient interconnections and largely alleviate the limitations of bus-based solutions. NoCs can span synchronous and asynchronous clock domains or use un-clocked asynchronous logic. NoC technology applies networking theory and methods to on-conventional bus and crossbar interconnections. NoC improves the scalability of SoCs (System on Chips), and the power efficiency of complex SoCs compared to other designs.[1]

Traditionally, ICs have been designed with dedicated point-to-point connections, with one wire dedicated to each signal. For large designs, in particular, this has several limitations from a physical design viewpoint. The wires occupy most of the area on the chip, interconnects dominate both performance and dynamic power dissipation, as signal propagation in wires across the chip requires multiple clock cycles. NoC links can reduce the complexity of designing wires for predictable speed, power, noise, reliability, etc., thanks to their regular, well controlled structure. From a system design viewpoint, with the advent of multi-core processor systems, a network is a natural architectural choice. A NoC can provide separation between computation and communication, support modularity and IP reuse via standard interfaces, handle synchronization issues, serve as a platform for system test, and, hence, increase engineering productivity. The whole router design can be implemented on a FPGA (Field programmable gate arrays).

There are three main components of NoC.

- Resource
- Resource Network Interface (RNI)
- Router

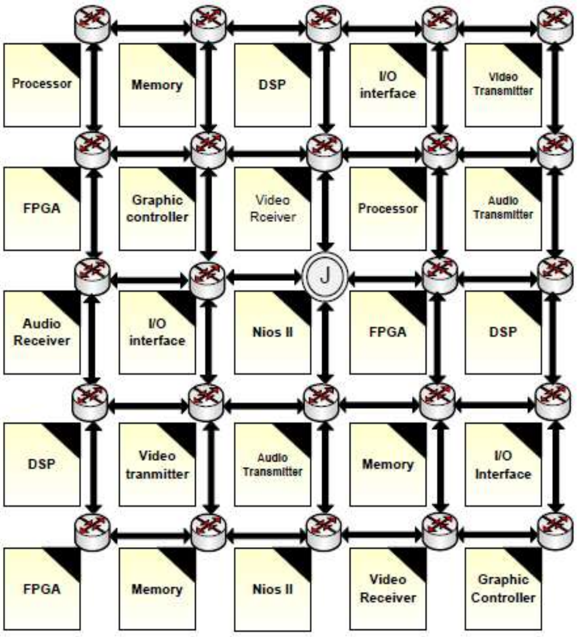


Figure 1 Example of Network on Chip platform[1]

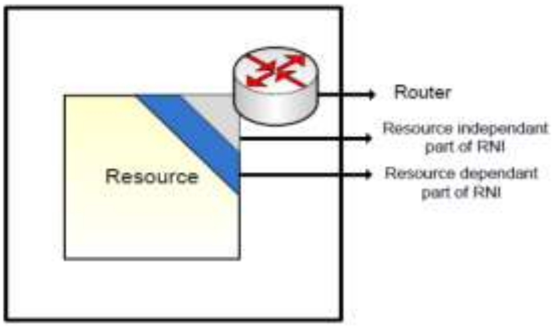

Router
Resource independant part of RNI
Resource dependant part of RNI
Resource

Figure 2: NoC Components [1]

A Core is connected to the router through RNI. RNI is the interface between router and the core. Core sends packet to the router through RNI, and from router it is send to the other routers or core depending on the destination.

Benefits of NoC Architecture:

- Independent implementation and optimization of layers.
- Simplified customization per application.
- Supports multiple topologies and options for different parts of the network.
- Simplified feature development, interface interoperability, and scalability.

## 1.2 Design for Junction based routing

Routing in NoC can be classified in many ways. Router design depends on the routing protocol and routing algorithm used. [1]

Two kinds of routing algorithms are source routing and distributed routing.

In source routing, path to be followed by the flit is locked from source to destination. The complete route information is available in head flit. Here, router takes decision by looking at the head flit. Path tables are stored inside the resource network interface (RNI). These path tables contain the complete path information for a specific destination in the network. Path information is calculated by applying routing algorithms.

In distributed routing, routing decisions are taken inside every router on the path and hence is a complex design due to extra hardware. There is no information about the path inside the header packet.

Compared to distributed routing, source routing has speed advantage because the routing information is stored in the packet itself. But source routing leads to overhead to store complete path information in the header of each packet.

An algorithm called Junction based source routing was developed to overcome these flaws. Junction Based Source Routing limits the required path information to be stored in every packet to a small number of bits which correspond to only a few hops as shown in figure 3. There are temporary destinations called junctions to cover the large distance such that sub-paths are always smaller than or equal to a maximum hop count. If a packet needs to go through a junction, the source just appends path information from source to the junction. On reaching the junction, the packet picks up path information to reach the destination from this junction.

The design of a NoC router depends on various aspects of NoC architecture and the performance requirement. The Junction Router contains path information in tables to reach any destination. The table can implement either in the router itself or in the resource network interface (RNI) or in the resource (core). There are three distinct cases for a packet to reach a junction:

i. If the destination of the packet is the resource connected to the Junction itself, then it should be routed to the resource through RNI.

ii. If the destination is not very far and the packet header has enough information to reach the destination, then the router forwards the packet just by looking at the relevant field in the header.

iii. If the destination is far, then the junction will be the intermediate destination. This will be clear from the relevant field in the header. In this case, Junction modifies the path information in the packet header for onward journey to the destination, if possible otherwise to another junction as intermediate destination. [1]

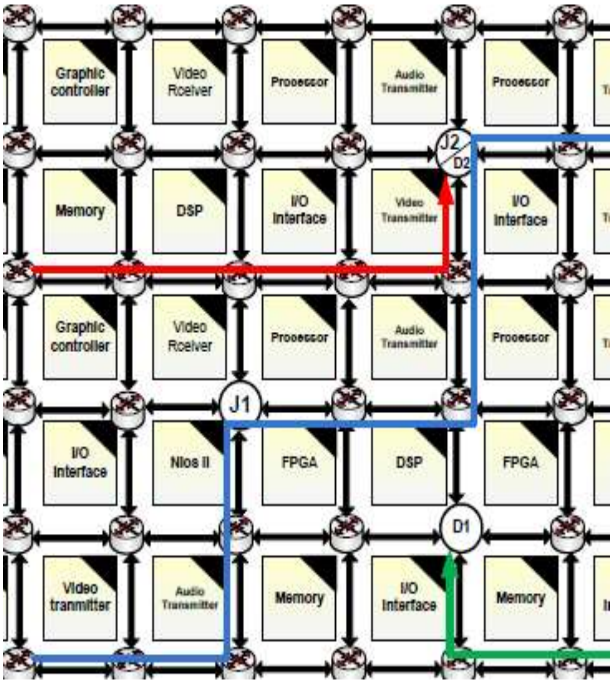


Fig.3 Junction Based Network on Chip based system [1]

The router architecture is a typical Network on Chip router. Design of a router depends on the routing algorithm used. Here junction based resource router has been described. This router architecture supports both kinds of routers which mean that the router can be used as a normal router as well as a junction router. Main components in the router architecture are as follow:

i. Arbiter and Control
ii. Crossbar
iii. Input Handler [1].

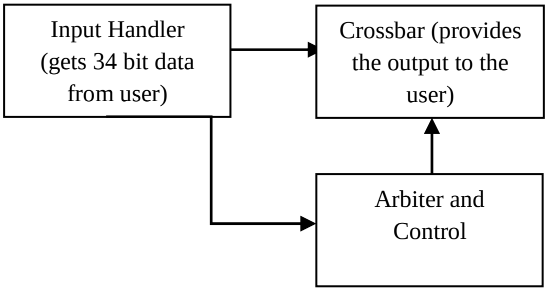

Input Handler (gets 34 bit data from user)
Crossbar (provides the output to the user)
Arbiter and Control

Figure 4: Main components of Router

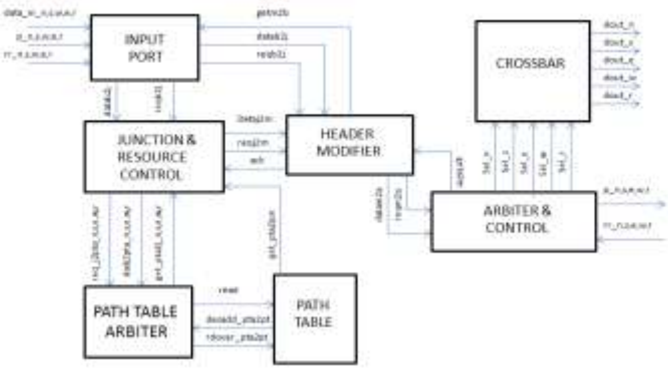


Figure 5: Block Diagram depicting all blocks of Junction based Router

## 1.3 Flit Format:

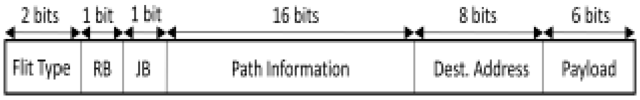


Figure 6: Flit Format.[1]

RB- Resource Bit.
JB- Junction Bit.
Dest. Address- Destination Address
The above diagram represents the Flit Format or the distribution of bits of the 34 bit data.
Flit type:
00- Head Flit.
01- Body Flit.
10- End Flit.
11- Full Flit.

### *Head Flit:*

When the Flit type is “00”, its Head Flit. This flit is initially sent to lock a particular path for transmission of data. This contains the destination address, direction bits, etc.

### *Body Flit:*

When the Flit type is “01”, its Body Flit. This flit follows the path reserved by the Head flit. It contains the main data which needs to be transmitted.

### *End Flit:*

When the Flit type is”10”, its End flit. When this flit is sent, it means that the data has been sent and no more data needs to be transmitted. This indicates end of transmission of data.

### *Full Flit:*

When the Flit type is “11”, its full flit. When this flit is sent, the path which was locked for transmission of data is unlocked for next data to be transmitted.

## 2. DESCRIPTION OF COMPONENTS PRESENT IN ROUTER:

## 2.1 Input Handler:

This block receives the data and stores it temporarily in a stack present. By using a stack of width 34bits and depth 4bits. The outputs are sent to different parts of Crossbar and Arbiter .This block consists of 3 subparts:

i. Input Port
ii. Header Modifier
iii. Junction and Resource Control.

### *2.1.1. Input Port*

- The main function of this block is to store the incoming flits and send them to header modifier block.
- For the storing the data temporarily, stack is used, which operates on ‘First In First Out’ (FIFO) concept.
- There are many temporary registers used to control the stack
- As the data is sent into the router, it becomes the data for this block, signal is generated and the data is written into the stack.
- Stack consists of a depth of 4, which means it can save maximum of 4 data at a time. When the stack is full, other data go into the waiting state for stack to empty at least one row.

### *2.1.2 Header Modifier*

- When the block receives request signal from the input port, data is sent from the port.
- When the data is received, flits are checked first. If the flits are Head flit or Full flit, then the RB (Resource Bit) is checked.
- If Resource Bit=1, then wait for request signal from junction to the Header modifier. When the signal is received, along with this the data from Junction and Resource control which is of 34bits is also received.
- First 4 bits of Header Modifier is sent to the Arbiter, these become the inputs of Arbiter block, which is discussed later.
- The 4 bits of Header Modifier are divided in following manner:
    - ➢ 2 bits for Flit type, first two bits.
    - ➢ 2 bits for direction i.e. from where the data is coming, next two bits.
- The entire 34bit data is sent to the Crossbar as its input from this block.
- If RB=0, then concatenation operation of data is performed. Here the path information is modified, first two bits are sent behind the last 2 bits of path information. This modified data is then sent to the crossbar as 34bits data and 4 bit data is sent to the Arbiter and control block.
- If the Flit type is body flit or end flit, data in the modifier is directly sent to the crossbar and the 4 bit data is sent to the Arbiter and control block.

### *2.1.3 Junction and Resource Control*

The main function of this block is to receive flits from input port, send destination address to Path Table Arbiter Component and receive new path information , To send flit with new path information to Header Modifier.

- When the block receives request signal from the input port, data is sent from the port.
- As soon as the data is received , flit type is checked. If the flit type is Head or Full type , then it checks for Resource Bit condition.
- If resource bit is 1, then it loads the data in destination address which goes from junction to path table arbiter else it goes into waiting state till it receives the address with resource bit as 1 .
- Further, if junction and resource control block is getting signal from path table arbiter then this block gets new path information from path table.
- Junction and resource control concatenates the new path information such that the first two bits are sent behind the last 2 bits of path information.
- If the flits are Body or End type, then it goes into waiting state till it receives Head or Full as flit type.

## 2.2 Arbiter and control

### *2.2.1 Path Table Arbiter*

- The main function of this block is to handle all the new path information requests and forward them to the path table, where the new path information is generated.

- Firstly, there is a request signal from Junction and Resource Control block. Run an FSM to set priority if data is coming from more than one direction:
  - North -------- First
  - South ------- Second
  - West -------- Third
  - East --------- Fourth
  - Resource -- Last

  Resource is given lowest priority to avoid entrance of a new flit when present data is being processed.
- As the request is accepted to be serviced, the destination address is forwarded to the path table to obtain new path information. Simultaneously, a read signal is generated and forwarded to the path table.
- The block remains in waiting state until a signal is obtained from the path table.
- As soon as this signal is obtained, signal is generated and sent to Junction Resource Control block.
- After this, the block enters the initial state of scanning requests.

### *2.2.2 Path Table*

- Main function of this block is to forward the new path information to the Junction and Resource Controller.
- For this purpose, there is a new path information created in the initial steps, input of destination address from the Path Table Arbiter, payload from the Input Port and an FSM is run to decide the direction bits i.e. 001 for north, 010 for south, 011 for west and 100 for east.
- The above three are concatenated and stored in a temporary register of 17 bits length.
- This data is then written into a stack. Its operation is similar to that of stack used in Input Port.
- At positive transition of clock, the path info is concatenated and written in the fifo,
- As soon as a read signal is received, the data is sent to the Junction and Resource Control block. After this, a signal is sent to the Path Table Arbiter indicating the completion of the task.

### *2.2.3 Arbiter using Round RobinAlgorithm:*

Arbiter & Control is considered as the Brain of the router because of its following functions described below:

- ➢ handles request for output direction.
- ➢ takes decision for output direction by decoding the Head flit.
- ➢ locking and unlocking the path.
- ➢ checks the space in the buffer of next router.
- ➢ To send signals for selecting the output direction.
- ➢ To send signal to the next router or resource to save data in buffer. [1]

Arbiter and Control makes use of Round robin Algorithm to assign priority to directions/ routing of the data. There are 5 directions which needs to be considered while making this router, North, South, West, East and Resource.

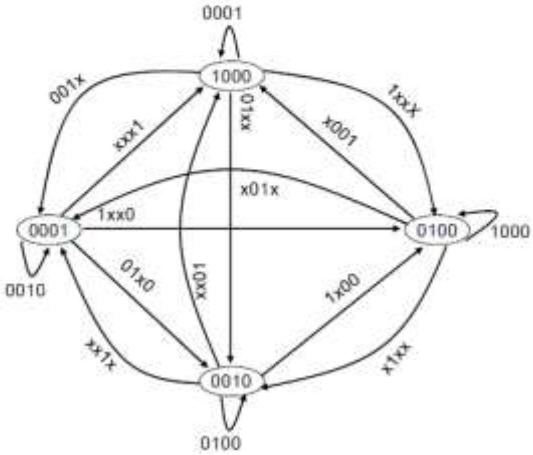


Figure7: FSM for using Round Robin Algorithm[2]

North- 1000
West- 0100
South- 0010
East- 0001

While using this algorithm assume a 3*3 mesh, resource will be the 5th router or the centre router, as discussed diagram above. Hence resource direction has not been taken into considerations. By using round Robin algorithm each state checks for all signals from all directions and then passes the signal according to their priority.

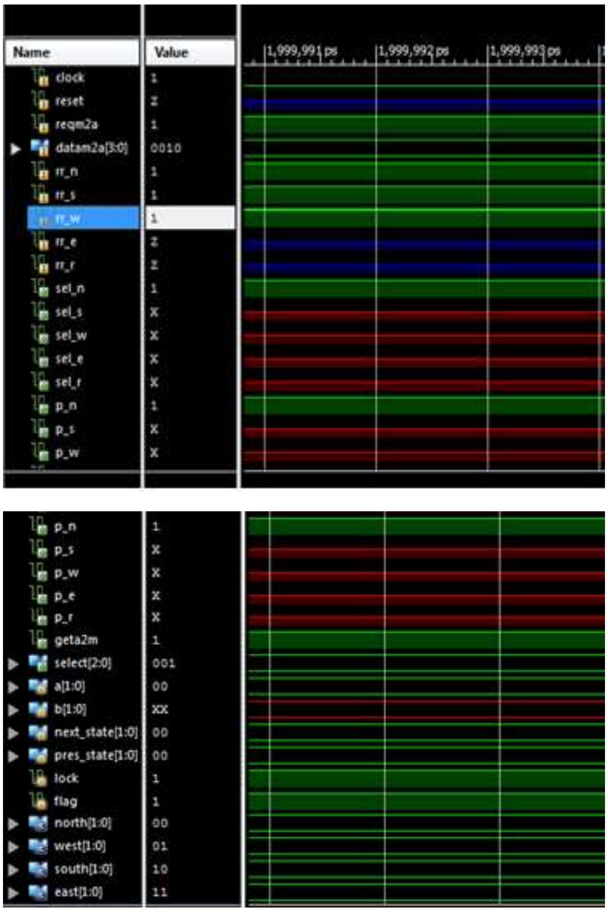
Highest priority is given to north, then west, south and east. The direction into consideration is anti-clockwise.

Arbiter works in following way:

- When it receives the signal from Header Modifier which is a 4 bit data, the first 2 bits are used for checking Flit type.
  - ➢ If the Flit type is head, it means that arbiter needs to assign a direction for the flow of data i.e. lock the path for data.
  - ➢ If the flit type is body or end, data follows the path assigned/ the path which is locked.
  - ➢ If the flit type is full, it means that the data transmission is over and the path should be unlocked.
- It checks from which direction Arbiter has received the signal and accordingly assigns the next state. Directions are assigned and used to avoid congestion of Data.
- When the direction the data is in the present state then the path is locked. Locking of path takes place only for flit type head.
- Along with this select is also assigned some value of 3bits. These values are given to crossbar for deciding the direction through which the output is expected.

  For example if the Present State is ‘North’ then select= 001 and select line for north=1, others are assigned as 0.

- This cycle continues till the data does not reach destination.



| Name | Value |
|---|---|
| clock | 1 |
| reset | z |
| reqm2a | 1 |
| datam2a[3:0] | 0010 |
| rr_n | 1 |
| rr_s | 1 |
| rr_w | 1 |
| rr_e | z |
| rr_r | z |
| sel_n | 1 |
| sel_s | x |
| sel_w | x |
| sel_e | x |
| sel_r | x |
| p_n | 1 |
| p_s | x |
| p_w | x |

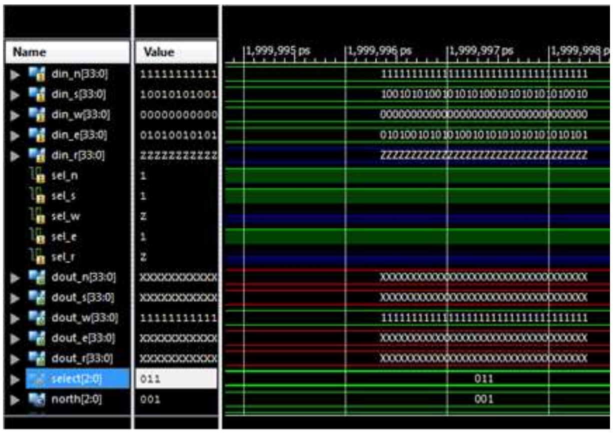
|  |  |
|---|---|
| p_n | 1 |
| p_s | x |
| p_w | x |
| p_e | x |
| p_r | x |
| geta2m | 1 |
| select[2:0] | 001 |
| a[1:0] | 00 |
| b[1:0] | xx |
| next_state[1:0] | 00 |
| pres_state[1:0] | 00 |
| lock | 1 |
| flag | 1 |
| north[1:0] | 00 |
| west[1:0] | 01 |
| south[1:0] | 10 |
| east[1:0] | 11 |

Figure 8: Simulation Result of Arbiter. Flit type is Head, North is given higher priority hence select=001 and, signal for north goes high(p_n) and select line for north (sel_n) goes high.

## 2.3 Crossbar

The main function of the crossbar is to send the data from one path to the assigned path.

- Select lines are assigned for choosing the direction or path of data flow.
- There are 5 select lines for 5 directions which are controlled by Arbiter.
- When the select line is known then data is collected from the Header Modifier and sent out through the same line.
- For example, if select=001, which means North state and in this state if west select line is high, then data is collected from modifier and sent as data out from west direction.



| Name | Value |
|---|---|
| din_n[33:0] | 11111111111 |
| din_s[33:0] | 10010101001 |
| din_w[33:0] | 00000000000 |
| din_e[33:0] | 01010010101 |
| din_r[33:0] | zzzzzzzzzz |
| sel_n | 1 |
| sel_s | 1 |
| sel_w | z |
| sel_e | 1 |
| sel_r | z |
| dout_n[33:0] | xxxxxxxxxxx |
| dout_s[33:0] | xxxxxxxxxxx |
| dout_w[33:0] | 11111111111 |
| dout_e[33:0] | xxxxxxxxxxx |
| dout_r[33:0] | xxxxxxxxxxx |
| select[2:0] | 011 |
| north[2:0] | 001 |

Figure 9 : Simulation report of Crossbar

## 2.4 Interfacing of the blocks

Verilog coding is usually done in segments and later a number of smaller segments is integrated together to give a common output. The process of interfacing involves the following mentioned steps:

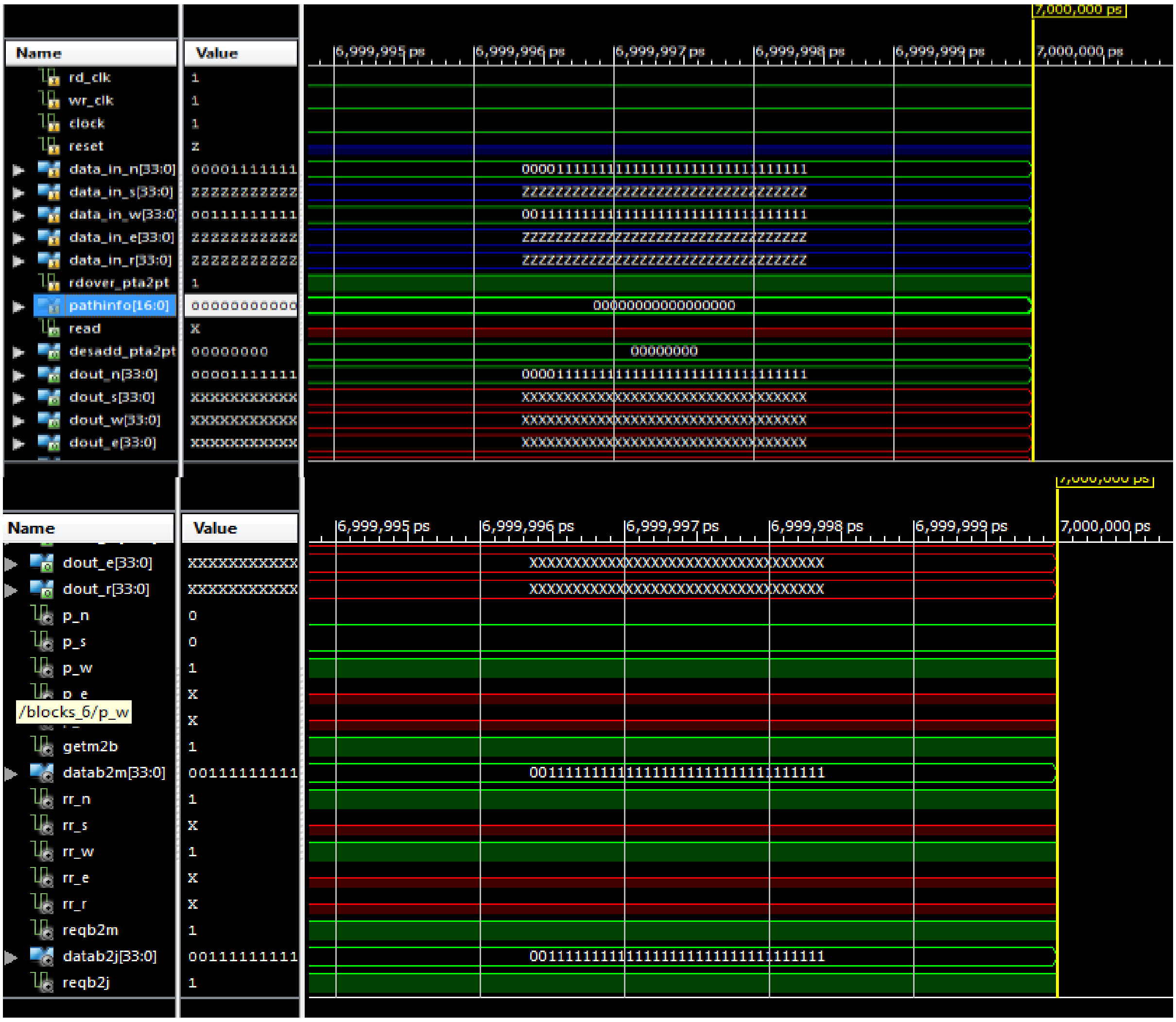
- Writing the codes individually.
- Creating a top module and calling the above programs as instances in the top module.
- Instantiating a program is calling the code is a defined format, that is, module ‘name’ ‘instance’(ports).
- After all the codes are instantiated and added in the top module group and a common simulation is checked.

## 2.5 Conclusions

This work presented the implementation of router which is divided into small blocks input port, header modifier , junction and resource control , path table, path table arbiter, crossbar and arbiter and control. The blocks were implemented through Verilog codes using ISE XILINX version 14.6 software. The simulation facilitates clear understanding of the functionality of each block in the router for a network on chip.

Six blocks have been interfaced satisfactorily which includes input port, header modifier, junction and resource control, path table arbiter, arbiter and crossbar.

Interfacing of the 7th block, Path Table is in the process.



| Name | Value |
|---|---|
| rd_clk | 1 |
| wr_clk | 1 |
| clock | 1 |
| reset | z |
| data_in_n[33:0] | 00001111111 |
| data_in_s[33:0] | zzzzzzzzzzz |
| data_in_w[33:0] | 00111111111 |
| data_in_e[33:0] | zzzzzzzzzzz |
| data_in_r[33:0] | zzzzzzzzzzz |
| rdover_pta2pt | 1 |
| pathinfo[16:0] | 00000000000 |
| read | x |
| desadd_pta2pt | 00000000 |
| dout_n[33:0] | 00001111111 |
| dout_s[33:0] | xxxxxxxxxxx |
| dout_w[33:0] | xxxxxxxxxxx |
| dout_e[33:0] | xxxxxxxxxxx |

| Name | Value |
|---|---|
| dout_e[33:0] | xxxxxxxxxxx |
| dout_r[33:0] | xxxxxxxxxxx |
| p_n | 0 |
| p_s | 0 |
| p_w | 1 |
| p_e | x |
|  | x |
| getm2b | 1 |
| datab2m[33:0] | 00111111111 |
| rr_n | 1 |
| rr_s | x |
| rr_w | 1 |
| rr_e | x |
| rr_r | x |
| reqb2m | 1 |
| datab2j[33:0] | 00111111111 |
| reqb2j | 1 |

Figure 10: Simulation Report of 6 blocks( Input port, Header Modifier, Junction and Resource Control, Path Table Arbiter, Arbiter and Crossbar) have been interfaced together. The inputs are given to data_in pin in any direction required and the output is obtained from data_out in the same direction as the input given.

## 3. ACKNOWLEDGMENTS

We would like to thank our mentor Ms. Priti Shahne for her continuous support and giving us the opportunity to work on this project.

## 4. REFERENCES

1. Muhammad Awais Aslam,” Router Architecture for Junction Based source routing: design and FPGA Prototyping”, Tekniska Hogskolan, 2011.

2. Jer-Min Jou and Yun-Lung Lee,” An Optimal Round-Robin Arbiter Design for NoC”, Journal of Information Science and Engineering,2010.

3. Tobias Bjerregaard and Shankar Mahadevan,” A Survey of Research and Practices of Network-on-Chip”, ACM computing surveys, 2006.

4. Ville Rantala, Teijo Lehtonen and Juha Plosila, “Network on Chip Routing Algorithms”, TUCS technical Report, 2006.

5. Stephen L. Chamberlin,” Design and implementation of Router using Xilinx FPGA”, Massachusetts Institute of Technology.

6. Sunil, Shaik Khadar Sharif, 3praveen Vanaparthy, “Fpga Implementation of Five Port Router Network”, International Journal of Engineering Development and Research.

7. Adrijean Adriahantenaina, Hervé Charlery, Alain Greiner, Laurent Mortiez, Cesar Albenes Zeferino, “SPIN: a Scalable, Packet Switched, On-chip Micro-network”.

8. Samir Palnitkar, “Verilog- HDL, a guide to Digital and Synthesis”.

9. J. Duato, S. Yalamanchili, and L. Ni, Interconnection Networks: An Engineering Approach. IEEE Computer Society Press, 1997.

International Journal of Computer Applications Technology and Research
Volume 4– Issue 8, 629 - 632, 2015, ISSN: 2319–8656

# Spam Detection in Social Networks Using Correlation Based Feature Subset Selection

Sanjeev Dhawan
Department of Computer Science & Engineering,
University Institute of Engineering and Technology,
Kurukshetra University, Kurukshetra-136119,
Haryana, India

Meena Devi
Department of Computer Science and Engineering
University Institute of Engineering and Technology,
Kurukshetra University, Kurukshetra-136119,
Haryana, India

**Abstract:** Bayesian classifier works efficiently on some fields, and badly on some. The performance of Bayesian Classifier suffers in fields that involve correlated features. Feature selection is beneficial in reducing dimensionality, removing irrelevant data, incrementing learning accuracy, and improving result comprehensibility. But, the recent increase of dimensionality of data place a hard challenge to many existing feature selection methods with respect to efficiency and effectiveness. In this paper, Bayesian Classifier with Correlation Based Feature Selection is introduced which can key out relevant features as well as redundancy among relevant features without pair wise correlation analysis. The efficiency and effectiveness of our method is presented through broad.

**Keywords**: Bayesian Classifier, Feature Subset Selection, Naïve Bayesian Classifier, Correlation Based FSS, Spam, Non-Spam

## 1. INTRODUCTION

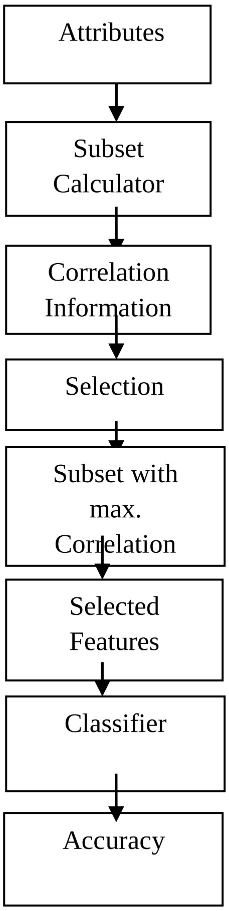
It is impossible to tell exactly who was the first one to come upon a simple idea that if you send out an advertisement to a number of people, then at least one person will react to it no matter what is the proposal. E-mail provides a very good way to send these millions of advertisements at no cost for the sender, and this unfortunate fact is nowadays extensively exploited by several organizations. As a result, the e-mailboxes of millions of people get cluttered with all this so-called unsolicited bulk e-mail also known as “spam” or “junk mail”. Being incredibly cheap to send, spam causes a lot of problems to the Internet community: large amounts of spam-traffic between servers cause delays in delivery of solicited email, people with dial-up Internet access have to spend bandwidth downloading junk mail. Sorting out the unwanted messages takes time and introduces a risk of deleting normal mail by mistake. Finally, there is quite an amount of pornographic spam that should not be uncovered to children. A number of ways of fighting spam have been proposed. There are “social” methods like legal measures (one example is an anti-spam law introduced in the US) and plain personal participation (never respond to spam, never publish your e-mail address on WebPages, never forward chain-letters. . .). There are 60 “technological” ways like blocking spammer’s IP-address (blacklist), e-mail filtering etc.. Unluckily, till now there is no perfect method to get rid of spam exists, so the amount of spam mail keeps increasing. For example, about 50% of the messages coming to my personal mailbox are unsolicited mail. For blocking spam at the moment Automatic e-mail filtering appears to be the most effective method and a tough competition between spammers and spam-filtering methods is going on: the better the anti-spam methods get, so do the tricks of the spammers. Several years ago most of the spam could be reliably handle by blocking e-mails coming from certain addresses or filtering out messages with certain subject lines. To overcome these spammers began to specify random sender addresses and to append random characters to the end of the message subject. Spam filtering rules adjusted to consider separate words in messages could deal with that, but then junk mail with specially spelled words (e.g. B-U-Y N-O-W) or simply with misspelled words (e.g. BUUY NOOW) was born. To fool the more advanced filters that relies on word frequencies spammers append a large amount of “usual words” to the end of a message. Besides, there are spams that contain no text at all (typical are HTML messages with a single image that is downloaded from the Internet when the message is opened), and there are even self-decrypting spams (e.g. an encrypted HTML message containing JavaScript code that decrypts its contents when opened). So, as you see, it’s a never-ending battle. There are two basic approaches to mail filtering knowledge engineering (KE) and machine learning (ML). In the former case, a set of rules is created according to which messages are categorized as spam or legitimate mail. A typical rule of this kind could look like “if the Subject of a message contains the text BUY NOW, then the message is spam”. A set of such rules should be created either by the user of the filter, or by some other authority (e.g. the software company that provides a particular rule-based spam-filtering tool).The major drawback of this method is that the set of rules must be constantly updated, and maintaining it is not convenient for most users. The rules could, of course, be updated in a centralized manner by the maintainer of the spam filtering tool, and there is even a peer-2-peer knowledgebase solution, but when the rules are publicly available, the spammer has the ability to adjust the text of his message so that it would pass through the filter. Therefore it is better when spam filtering is customized on a per-user basis. The machine learning approach does not require specifying any rules explicitly. Instead, a set of pre-classified documents (training samples) is needed. A specific algorithm is then used to “learn” the classification rules from this data. The subject of machine learning has been widely studied and there are lots of algorithms suitable for this task. This article considers some of the most popular machine learning algorithms and their application to the problem of spam filtering. More-or-less self-contained descriptions of the algorithms are presented and a simple comparison of the performance of my implementations of the algorithms is given. Finally, some ideas of improving the algorithms are shown.

## 2. CHALLENGES IN SPAM DETECTION

One of the barriers to legislation against spam is the fact that not everyone uses exactly the same definition. It doesn’t help that laws may be made at different levels even within the same country, let alone laws in different countries. With so many different and sometimes conflicting laws, prosecution can be very difficult. Another barrier both to legislation and practical filtering is that email is not designed in such a way that the sender can always be traced easily. There is no authentication of the sender built in to the protocol used by email, leaving it possible for people to forge sender information. This makes it hard to trace back and prosecute the sender, or to avoid receiving messages from a known spammer in the future. There are several proposals to adapt this protocol like Microsoft's "Caller ID for email". Spam changes with time as new product are introduced and seasons change. For example, Christmas-themed spam is not usually sent in June. But beyond that, there are targeted changes happening in spam. Perhaps the largest problem of spam filtering is that spammers have intelligent beings working to ensure that "direct email marketing" (the marketing term for spam) is seen by as many potential customers as possible. Many anti-spam tools are freely available online, which means that spammers have access to them too, and can learn how to get through them. This makes spam detection a co-evolutionary process, much like virus detection: both sides change to gain an advantage, however temporarily. Although it does change, spam is not completely volatile. Terry Sullivan found that while spam does undergo periods of rapid changes, it also has a core set of features which are stable for long periods of time. Spam changes from person to person. This is partly due to targeting on the part of the address harvesters, who try to guess the interests of the recipients so that the response rate will be higher. But more importantly, legitimate mail also varies from person to person. In theory it should be possible to discover spam without much attention to the legitimate mail. However, the great success of classifiers which use both, such as Graham's Bayesian classifier and the CRM114 discriminator [Yer04], implies that use of data from both legitimate and spam email is very beneficial. One final thing to note in the difficulty of spam classification is that all mistakes in classification are not equal. False negatives, messages that have accidentally been tagged as non-spam, are usually seen by the user. They may be annoying, but are usually easy to deal with. However, false positives, messages that have been accidentally tagged as spam, tend to be more problematic. When a single legitimate message is in a pile of spam, it is much easier to miss seeing it. (A typical user will not read all spam, but instead scans subject and from lines quickly to see if anything legitimate stands out.) While there is relatively little impact if a person receives a single spam, missing a real message which might be important is much more dangerous. One research firm suggests that companies lose $3 billion dealing with false positives.

## 3. PROPOSED WORK

In previous work various spam detection algorithm have been proposed ranging from text based to feature based using classifiers such as naïve bayes, SVM, ANN, kNN and decision tree etc. However Naïve Bayesian Method is utilized by 99% of the company. The reason for this is their classification efficiency. But these probabilistic methods take in consideration all the feature of the spam making the overall accuracy ranging from 65 to 74 %. So we require a more efficient method to improve spam detection and false alarm reduction. The feature subset algorithm tries to formulate the vector space of the features by filtering of subset selecting the most prominent feature of spam and removing unwanted features. The filtering allows the reduction in search space and noise. After filtering using FSS we have applied attribute selection based naïve Bayesian probabilistic classifier and achieved 17-20% more accuracy.



## 4. FEATURE SUBSET SELECTION

Feature subset selection is used for identifying and removing as much irrelevant and redundant information as possible and thus it reduces the dimensionality of the data and may allow learning algorithms to run faster and more effectively. In some cases, accuracy on future classification can be improved; in others, the result is a more compact, well interpreted representation of the aimed concept.

## 5. CORRELATION BASED FSS

CFS algorithm relies on a heuristic for assessing the cost or merit of a subset of features. This heuristic takes into account the usefulness of individual features for forecasting the class label along with the level of intercorrelation among them. The hypotheses on which the heuristic is based is:

Sound feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.

Features are relevant if their values vary systematically with category membership. A feature is useful if it is correlated with or forecaster of the class; otherwise it is irrelevant. Empirical grounds from the feature selection literature show that, along with irrelevant features, redundant information

should be wiped out as well. A feature is said to be redundant if one or more of the other features are highly correlated with it. The above definitions for relevance and redundancy lead to the idea that best features for a given classification are those that are highly correlated with one of the classes and have an insignificant correlation with the rest of the features in the set.
If the correlation between each of the components in a test and the outside variable is known, and the inter-correlation between each pair of components is given,then the correlation between a composite consisting of the summed components and the outside variable can be predicted from

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k - (k-1)\overline{r_{ii}}}} \qquad (5.1)$$

Where

$r_{zc}$ = correlation between the summed components and the outside variable.

$k$ = number of components (features).

$\overline{r_{zi}}$ = average of the correlations between the components and the outside variable.

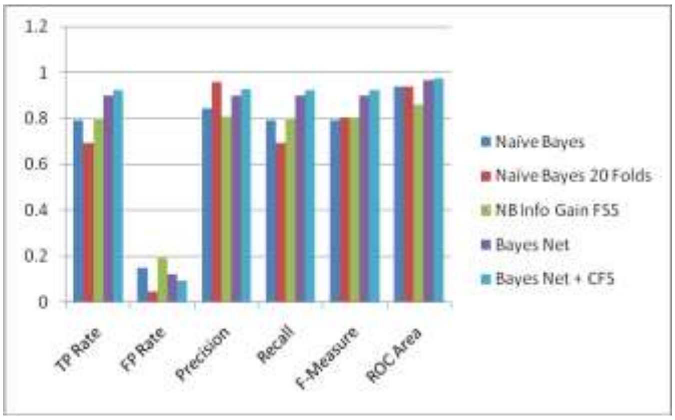
$\overline{r_{ii}}$ = average inter-correlation between components.

Equation 5.1 represents the Pearson’s correlation coefficient, where all the variables have been standardized. The numerator can be thought of as giving an indication of how predictive of the class a group of features are; the denominator of how much redundancy there is among them. Thus, equation 5.1 shows that the correlation between a composite and an outside variable is a function of the number of component variables in the composite and the magnitude of the inter-correlations among them, together with the magnitude of the correlations between the components and the outside variable. Some conclusions can be extracted from (5.1):

- The higher the correlations between the components and the outside variable, the higher the correlation between the composite and the outside variable.
- As the number of components in the composite increases, the correlation between the composite and the outside variable increases.
- The lower the inter-correlation among the components, the higher the correlation between the composite and the outside variable.

## 6. CLASSIFICATION RESULTS

| Classifier | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Correct |
|---|---|---|---|---|---|---|---|
| Naïve Bayes | 0.793 | 0.152 | 0.842 | 0.793 | 0.794 | 0.937 | 79.3871 |
| Naïve Bayes 20 Folds | 0.692 | 0.046 | 0.959 | 0.692 | 0.804 | 0.937 | 79.5262 |
| NB Info Gain FSS | 0.8 | 0.196 | 0.808 | 0.8 | 0.802 | 0.861 | 80.0478 |
| Bayes Net | 0.9 | 0.123 | 0.9 | 0.9 | 0.899 | 0.965 | 89.9587 |
| Bayes Net + CFS | 0.924 | 0.096 | 0.925 | 0.924 | 0.924 | 0.974 | 92.4147 |



## 7. CONCLUSION AND FUTURE SCOPE

Feature subset selection (FSS) plays a vital act in the fields of data excavating and contraption learning. A good FSS algorithm can efficiently remove irrelevant and redundant features and seize into report feature interaction. This also clears the understanding of the data and additionally enhances the presentation of a learner by enhancing the generalization capacity and the interpretability of the discovering mode. An alternative way employing a classifier on a corpus of e-mail memos from countless users and a collective dataset.
In this work we have worked on improving SPAM detection based on feature subset selection of Spam data set. The Feature Subset selection methods such as Info Gain Attribute selection and Correlation based Attribute Selection can be perceived as the main enhancement to Naïve Bayesian/ probabilistic methods. We have analyzed the Probabilistic SPAM Filters and attained more than 92% of success in filtering SPAM.
However many open issues still remain open such as, the system deals only with content as it has been translated to plain text or HTML. Since some spam is sent where most of the message is in an image, it would be worth looking at ways in which images and other attachments could be examined by the system. These could include algorithms which extract text from the attachment, or more complex analysis of the information contained within the attachment. We can also work on a technique to recognize web junk e-mail according to finding these boosting pages in place of web spam page itself. We will begin from a small set of spam seed pages to get a hold of boosting pages. Then web junk e-mail pages are supposed to be identified making use of boosting pages. We can also work on a better larger dataset; the system should be tested over a longer period than the one-year one available in the public domain.

## 8. REFERENCES

[1] Hayati, Vidyasagar Potdar and Pedram, “Evaluation of spam detection and prevention frameworks for email and image spam: a state of art,” In Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services, ACM, pp. 520-527, 2008.

[2] Becchetti, Luca, Carlos Castillo, Debora Donato, Ricardo Baeza-Yates and Stefano Leonardi, “Link analysis for web spam detection,” ACM Transactions on the Web (TWEB), vol. 2, no. 1, 2008.

[3] Ioannis Kanaris, Konstantinos Kanaris, Ioannis Houvardas, And Efstathios Stamatatos, “Words Vs. Character N-Grams For Anti-Spam Filtering,” International Journal on Artificial Intelligence Tools, pp. 1–20, 2006.

[4] Joshua Attenberg, Kilian Weinberger, Anirban Dasgupta, Alex Smola, and Martin Zinkevich, “Collaborative Email-Spam Filtering with the Hashing Trick,” CEAS, 2009.

[5] Tu Ouyang, Soumya Ray, Michael Rabinovich and Mark Allman,” Can network characteristics detect spam effectively in a stand-alone enterprise?,” In Passive and Active Measurement, (Springer Berlin Heidelberg, 2011), pp. 92-101, 2011.

[6] Rushdi Shams and Robert E. Mercer,”Classifying Spam Emails using Text and Readability Features,” IEEE 13th International Conference on Data Mining (ICDM), pp. 657-666, 2013.

[7] Lei Yu, Huan Liu,” Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution” Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[8] Liumei Zhang, Jianfeng Ma, and Yichuan Wang, “Content Based Spam Text Classification: An Empirical Comparison between English and Chinese,” 5th International Conference on Intelligent Networking and Collaborative Systems (INCoS), IEEE, pp. 69-76, 2013.

[9] Igor Santos, Carlos Laorden, Borja Sanz, and Pablo Garcia Bringas, “JURD: Joiner of Un-Readable Documents to reverse tokenization attacks to content-based spam filters”, Consumer Communications and Networking Conference (CCNC), IEEE, pp. 259-264, 2013.

[10] De Wang, Danesh Irani, and Calton Pu, “ A study on evolution of email spam over fifteen years,” IEEE 2013 9th International Conference on In Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), pp. 1-10, 2013.

[11] Bujang, Yanti Rosmunie, and Husnayati Hussin, “Should we be concerned with spam emails? A look at its impacts and implications,” 2013 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M), IEEE, pp. 1-6 2013.

[12] Manek, Asha S., D. K. Shamini, Veena H. Bhat, P. Deepa Shenoy, M. Chandra Mohan, K. R. Venugopal, and L. M. Patnaik, “ReP-ETD: A Repetitive Preprocessing technique for Embedded Text Detection from images in spam emails,” 2014 IEEE International Advance Computing Conference (IACC), pp. 568-573, 2014.

[13] Bosma, Maarten, Edgar Meij, and Wouter Weerkamp, “A framework for unsupervised spam detection in social networking sites, Advances in Information Retrieval,” Springer Berlin Heidelberg, pp. 364-375, 2012.

[14] Dave, Vacha, Saikat Guha, and Yin Zhang, “Measuring and fingerprinting click-spam in ad networks,” In Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication, ACM, pp. 175-186, 2012.

[15] Karthika Renuka and Visalakshi, “Latent Semantic Indexing Based SVM Model for Email Spam Classification,” Journal of Scientific & Industrial Research, vol. 73, pp. 437-442,July 2014.