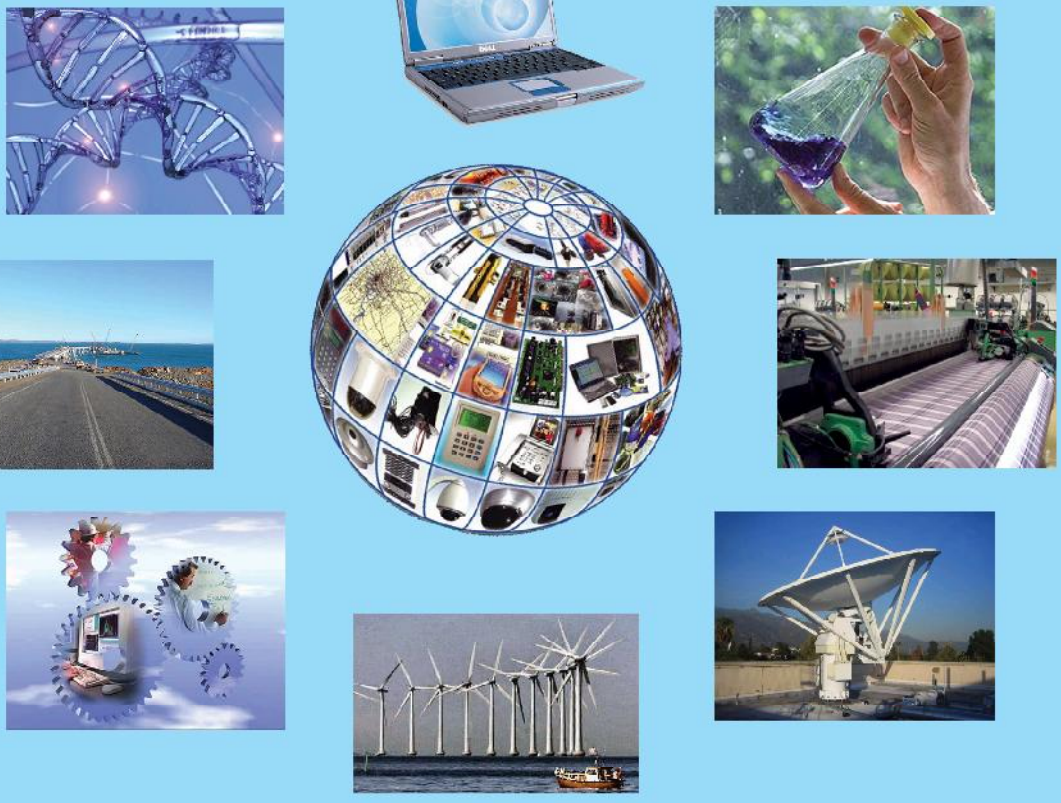
IJEST

ISSN: 0973-6255
Vol. 7 No. 1
January - June 2013

Indian Journal of Engineering, Science, and Technology

A Refereed Research Journal



Published by

BANNARI AMMAN INSTITUTE OF TECHNOLOGY

(Autonomous Institution Affiliated to Anna University of Technology, Coimbatore -
Approved by AICTE - Accredited by NBA and NAAC with "A" Grade)
Sathyamangalam - 638 401 Erode District Tamil Nadu India
Ph: 04295-226340 - 44 Fax: 04295-226666
www.bitsathy.ac.in E-mail: ijest@bitsathy.ac.in

BANNARI AMMAN INSTITUTE OF TECHNOLOGY
Stay Ahead

TÜV NORD ISO 9001

Indian Journal of Engineering, Science, and Technology

IJEST is a refereed research journal published half-yearly by Bannari Amman Institute of Technology. Responsibility for the contents rests upon the authors and not upon the IJEST. For copying or reprint permission, write to Copyright Department, IJEST, Bannari Amman Institute of Technology, Sathyamanagalam, Erode District - 638 401, Tamil Nadu, India.

Indian Journal of Engineering, Science, and Technology

ISSN: 0973-6255

Volume 7 Number 1 **January - June 2013**

CONTENTS

Optimization of Fuzzy Based PD Controller

K. Lakshmi[1] and P. Harikrishnan[2]

[1&2]Department of Control and Instrumentation Engineering, Anna University of Technology, Coimbatore - 641 047, Tamil Nadu
E-mail: lakshmikrishnan29@gmail.com, hariora@gmail.com

Abstract

A Proportional Integral derivative controller is a most commonly used feedback controller in industrial control system. Although a proportional-integral-derivative (PID) controller is popular and relatively simple in structure, it must also be pointed out that the unnecessary mathematical rigorosity, preciseness and accuracy involved with the design of the controllers have been a major drawback. This drawback can be highly eliminated by designing systems with PD controller to improve the system performance. In this paper genetic algorithm technique is used to design the Proportional Derivative Fuzzy logic controller. A comparative study of PD, fuzzy based PD and optimized fuzzy logic PD controller is analyzed. Simulation result shows that the optimized proportional derivative fuzzy logic controller improves the system performance in terms of rise time and settling time, besides reducing overshoot and steady state error. This approach is first simulated using MATLAB / SIMULINK using the techniques of PD- Fuzzy Logic controller. The output of the PD controller serves as an input to the DC motor. The desired speed of the motor is then achieved by tuning the fuzzy controller with the help of genetic algorithm technique.

Keywords: *Control, PD control, Fuzzy logic, Genetic algorithm, DC motor*

1. INTRODUCTION

The control of processes and systems in the industry is customarily done by experts through the conventional PID control techniques. This is as a result of its simplicity, low cost design and robust performance in a wide range of operating conditions. Although the PID controllers have gained widespread usage across technological industries, it has some drawback. This has made it difficult if not impossible for designers, engineers and technology experts to design intelligent complex systems, nonlinear systems higher order and time-delayed linear systems that can satisfactorily behave as expected while operating in the human-machine interface. However, various techniques and modifications to the conventional PID controllers have been employed in order to overcome these difficulties, this include the use of auto tuning PID controllers, adaptive PID controllers and also the implementation of compensation schemes to the PID controllers. Another alternative technique is the use of unconventional control techniques such as fuzzy logic, neural networks and genetic algorithm or a combination of two or more of these techniques. The combination of the genetic algorithm and the fuzzy logic control technique will be considered in this study. Genetic algorithms, which are adopted from the principle of biological evolution, are efficient search techniques that manipulate the codings representing a parameter set to reach a near optimal solution. Hence by strengthening fuzzy logic controllers with genetic algorithms the searching and attainment of optimal fuzzy logic rules and high-performance membership functions will be easier and faster. Although the benefits of harnessing the capabilities of genetic algorithms are huge, research efforts on optimizing fuzzy logic rules, membership functions and other parameters are challenging. In[8], genetic algorithm technique has been used to optimized fuzzy logic rules while in, a customized GA technique has been proposed to optimize the search for optimal fuzzy logic rules. Research efforts focused majorly on the optimal tuning of membership functions in[6] while in[8] both fuzzy logic rules, membership functions and other parameters were optimized using genetic algorithm. The combination of genetic algorithm and fuzzy logic controllers is normally shortened as GAFLC and this intelligent hybrid controller has found application in many scenarios like motor speed control [9, 10], temperature control [11] robotics [8] and in many other control systems. This study employs the fuzzy logic technique to design a Proportional-Derivative (PD) Controller and optimizes the inference rules, membership functions and scaling gains of this controller by using Genetic Algorithm (GA) and Particle swarm optimization technique. The resultant optimal fuzzy logic controller is used in the speed control unit which

comprises of a DC motor, control valve and other components. The performance of the GA Optimized Fuzzy Logic controller and PS Optimized Fuzzy Logic controller are compared with that of the conventional PD controller. The MATLAB/SIMULINK software forms part of the modeling and design tools employed in this paper.

2. FUNDAMENTALS ON FLBC

In recent years fuzzy logic has met a growing interest in many industrial applications. The attention toward this technique is due to its nonlinear features, independence from an accurate system modeling, and reduction of development and maintenance time [1]. The fuzzy controller tries to emulate a human operator: both of them operate in a knowledge-based way, and their knowledge relies on a set of linguistic if-then rules in one case, and on the human experience in the other [2]. In the field of electric drives fuzzy logic-based controllers (FLBCs) are more robust than conventional PID controllers when parameter detuning and load disturbance occur but it is not yet clear which design choices affect their robustness and stability.

Zadeh introduced fuzzy set theory in 1965 [3].With his original methodology it is possible to define the control laws of any process starting from a linguistic description of the control strategy to be adopted. Fuzzy logic-based controllers are characterized by simple relations (fuzzy rules) that describe the interaction between linguistic variables instead numerical ones. These rules are combined to form the decision table of the fuzzy controller and are structured as follows:

if x is A_i and y is B_i then z is C_i (1)

where A_i, B_i, and C_i are the linguistic values associated to the process state variables x, y and the controller output z . The if-part of the rule, called rule-antecedent, and then-part of the rule, called rule-consequent, describe inputs and output in term of fuzzy (linguistic) proposition.

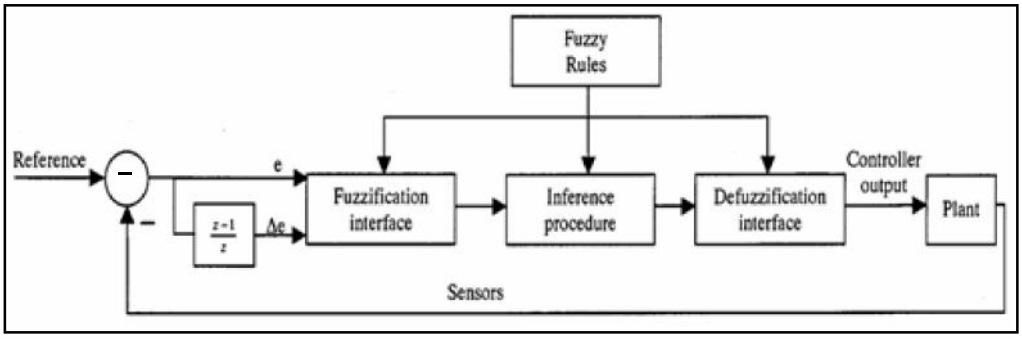


Fig. 2.1 Fuzzy Controller Block Diagram

The meaning of these linguistic values is given by the membership functions m_i, that associate a real number μ_i in the interval [0, 1], called grade of membership, to each input and output numerical-value (crisp-value).Input and output values have to be respectively normalized and denormalized, so that the crisp values handled by the FLBC are bounded in an interval U, called universe of discourse. A block diagram of a fuzzy controller is shown in Fig. 1. The set point value is compared with the controlled variable to derive the error (e) and the change of error (Де). These quantities are input signals for the FLBC, which requires three fundamental steps to generate the controller output:

1) fuzzification of actual input values;
2) fuzzy inference;
3) defuzzification of fuzzy output.

Fuzzification permits to convert the current input crisp-values into linguistic values, to make them compatible with the fuzzy expression in the rule-antecedent. Moreover it permits to find the degree of membership of the inputs to each corresponding membership function. Fuzzy inference associates a truth-value ϕ_i to each rule. This truth-value can be calculated with a T-norm operation (e.g., minimum, product, etc.) among the membership degrees of the antecedents corresponding to the current input values x and y . The output fuzzy set C_i has membership function $m'_{Ci}(z) = \min(m'_{Ci}(z)T_i)$, according to Mamdani implication. The obtained membership functions are then combined into an overall fuzzy set, given by $m_{out}(z) = \max(m'_{Ci}(z))$. The output of the inference procedure $m_{out}(z)$ must be transformed into a crisp value to become the controller output.

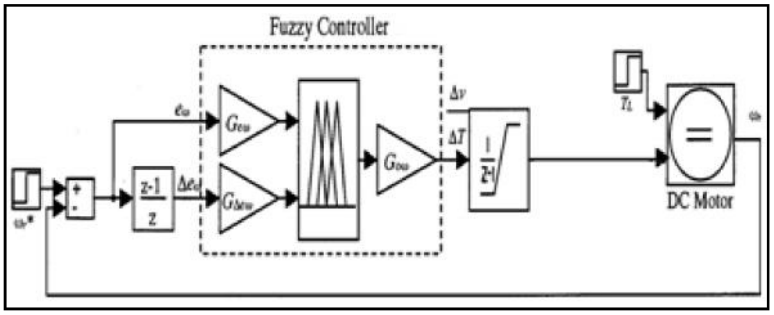


Fig. 2.2 DC drive block diagram

Membership functions allow us to graphically represent a fuzzy set. The *x* axis represents the universe of discourse, whereas the *y* axis represents the degrees of membership in the [0,1] interval.

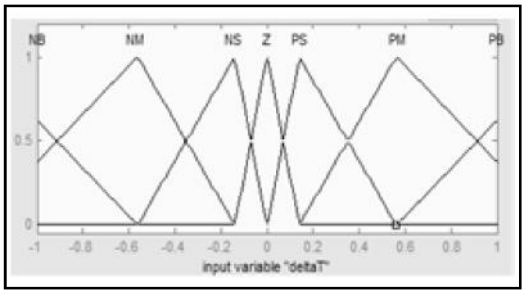


Fig. 2.3 Membership function for input variable deltaT (error)

Membership functions of error value, derivative error value and output variable of the fuzzy based proportional derivative controller are shown in figure 2.3, 2.4, 2.5.

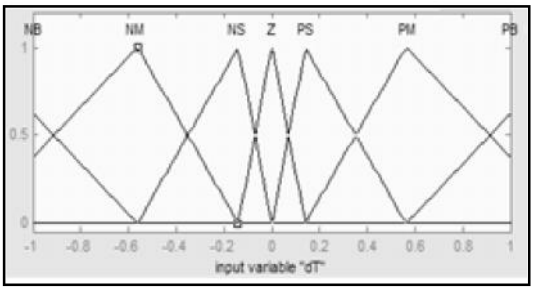


Fig. 2.4. Membership Function for Input Variable dt

The fuzzy variables themselves are adjectives that modify the variable (e.g. “large positive” error, “small positive” error, “zero” error, “small negative” error, and “large negative” error). As a minimum, one could simply have “positive”, “zero”, and “negative” variables for each of the parameters. Additional ranges such as “very large” and “very small” could also be added to extend the responsiveness to exceptional or very nonlinear conditions, but aren’t necessary in a basic system.

The fuzzy inference system consists of three linguistic variables (two inputs and one output) each having seven membership function sets. This results in 49-rule fuzzy inference system with inputs as the error and the rate of change in error. The output of the fuzzy logic inference system is the control action of the controller and the universe of discourse of all the variables are set within the range (-1, 1).

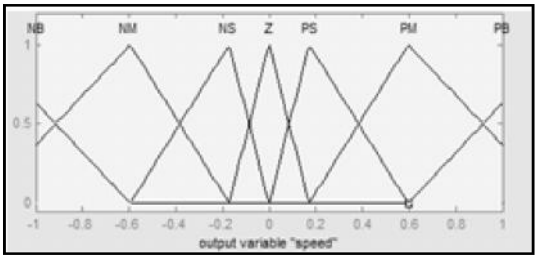


Fig. 2.5 Membership function for output variable speed

The GA-optimized fuzzy logic rules for the fuzzy Inference system is shown in Table 2.1.

Table 2.1 Optimized Fuzzy Rules

	Error						
De	NB	NM	NS	Z	PS	PM	PB
NB	NB	NB	NB	NM	NS	Z	PS
NM	NB	NB	NB	NM	Z	PS	PM
NS	NB	NB	NM	NS	Z	PM	PB
Z	NB	NM	NS	Z	PS	PM	PB
PS	NB	NM	Z	PS	PM	PB	PB
PM	NM	NS	Z	PM	PB	PB	PB
PB	NS	Z	PS	PM	PB	PB	PB

3. GENETIC ALGORITHM

3.1 Overview of Genetic Algorithms

The antecedent parameters, $\mathbf{c_{ij}}$ and y_{ij} and the consequent parameters, q_{ij} and r_{ij}, of the fuzzy inference system must be optimized for the good performance of a fuzzy controller. Most fuzzy system adaptation methods use the error back propagation algorithm based on the gradient descent method Since conventional training algorithms of a fuzzy controller do not work well when the plant has non minimum phase characteristics as mentioned before, a genetic algorithm is adopted as the training algorithm of a fuzzy controller.

Many optimization methods move from a single point in the decision space to the next using some transition rule to determine the next point. This point-to-point method is dangerous because it is a perfect prescription for locating false peaks in many peaked search spaces. By contrast, genetic algorithms work from a rich database of points, simultaneously climbing many peaks in parallel. Thus, the probability of finding a false peak is reduced over methods that go point-to-point. Therefore, genetic algorithms are less susceptible to getting stuck at local optima than conventional search methods.

In genetic algorithms, the term *chromosome* typically refers to a candidate solution to a problem, often encoded as a bit string. The genes are either single bits or short blocks of adjacent bits that encode a particular element of the candidate solution. An allele in a bit string is either zero or one. Each chromosome can be thought of as a point in the search space of candidate solutions. The genetic algorithms process populations of chromosomes, successively replacing one such population with another. The genetic algorithms require a fitness function that assigns a score to each chromosome in the current population.

i. Selection Operator: This operator selects individuals (chromosomes) in the population for reproduction. The goodness of each individual depends on its fitness. Fitness may be determined by an objective function. The fitter the chromosome, the more times it is likely to be selected to be reproduced.
ii. Crossover Operator: Two individuals are chosen from the population using the selection operator. This operator randomly chooses a crossover site along the bit strings and exchanges the subsequences before and after that crossover site between the two individuals to create two offspring. For example, the strings “000000” and “111 111” could be crossed over after the second locus in each to produce the two offspring “110000” and “001 1 11”.The two new offspring created from this mating are put into the next generation of the population. By recombining portions of good individuals, this process is likely to create even better individuals.
iii. Mutation Operator: With some low probability, a portion of the new individuals will have some of their bits flipped. Mutation can occur at each bit position in a string with some probability, usually very small. Its purpose is to maintain diversity within the population and inhibit premature convergence.

Table 3.1 Bit Distribution of the Chromosomes

Rule Base	MF for Proportional Input	MF for Derivative Input	MF for Output	Proportional Scaling Gain	Derivative Scaling Gain	Output Scaling Gain
9bits	7 bits	7 bits	7 bits	7 bits	7 bits	7 bits

The chromosome that has lower energy has higher Fitness. A genetic algorithm uses a cost function that evaluates the extent to which each individual is suitable for the given objectives, such as small undershoot and overshoot together with small overall error. It is required that a controller has desirable response to the step changes of set point and external disturbance. The fitness of an individual (chromosome) is calculated by means of the energy of the individual[4]. Therefore, the energy functions are defined by the following three equations:

$$E = +” \check{S} y_d(t) - y(t) \check{S} \, dt \quad \text{.....} \quad (1)$$

where $y\,d\,(\,t\,)$ and $y\,(\,t\,)$ are the step set point change and the actual output response, respectively, and T is a training time interval. E_1, E_2, and E_3 are overall sums of absolute errors, absolute value of undershoot, and absolute value of overshoot, respectively. The fitness function is given as follows:

$$F = \exp(\delta - E_t) \quad \text{.....} \quad (2)$$

$$E_t = \alpha E_1 + \beta E_2 + \gamma E_3 \quad \text{.....} \quad (3)$$

which is called a total-weighted error from now on; a, в, and г are the weighting coefficients.

To increase the efficiency of the conventional genetic algorithm, the proposed genetic algorithm has initial coarse tuning characteristics by initially representing each parameter in a chromosome by a small bit number. If the parameters in a chromosome are represented by a big bit number, the genetic algorithm can find the accurate optimal points in a limit of resolution but needs much more time to reach a convergence point. As it were, because the genetic algorithm is a time-consuming algorithm, it is unnecessary to represent by a big bit number from the beginning the parameters in chromosomes which are distant from the optimal points. However, it is necessary to represent it by a big bit number as many chromosomes (solution) gradually approach the optimal points. Therefore, when the simulation generation reaches one third of the maximum generation, the bit number is increased by one-third of its initial bit number. And then, when the simulation generation reaches two thirds of the maximum generation, the bit number is increased by two-thirds of its initial bit number. By this method, the genetic algorithm has initial coarse-tuning and final fine-tuning characteristics.

The crossover site is selected by two ways. The first is that the crossover site is selected randomly in a chromosome. The second is that the crossover site is selected between only parameters in a chromosome. This method slows a premature convergence without reaching optimal solutions and speeds up a final convergence The probability that one of the two ways is chosen depends on the number of the tuned parameters. In this work, the two ways go fifty-fifty. It is difficult to determine general characteristics through some simulation results since the genetic algorithm randomly searches optimal solutions. In some qualitative aspects, the first way increases diversity since the crossover site is randomly selected anywhere in a chromosome. Therefore, this way prevents a premature convergence without reaching optimal solutions.

4. SIMULATION RESULTS

4.1 Simulation Model

MATLAB M-files where utilized for the encoding, testing and decoding of each of the tuned FLC parameters. This includes the fuzzy logic rule base, the membership function definition of the linguistic variables and the scaling gains of the controller.

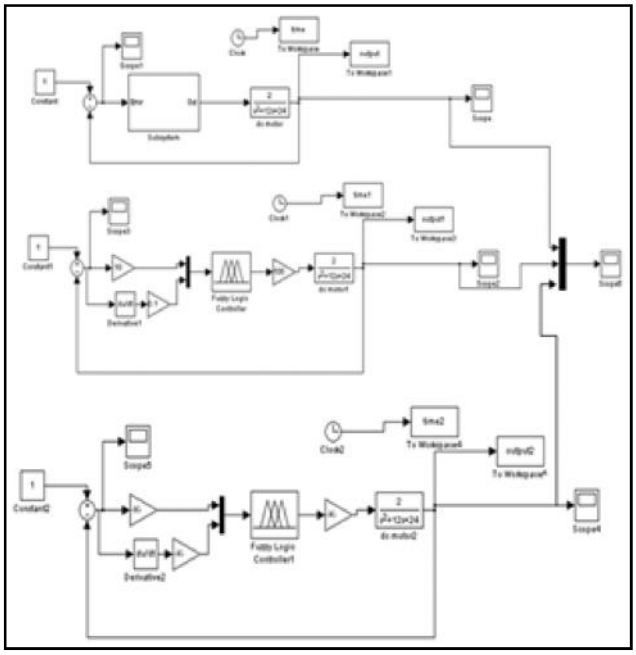


Fig. 4.1 Simulation model of different controllers

The output scaling gain of the controller was adjusted over a range of 0-250 while the proportional input scaling gain and the derivative input scaling gains were adjusted over a range of 0-50 and 0-1.5 respectively. The suitability of the ranges of the scaling gains was determined from the prior hand tuning of the controller. The output responses of Conventional Pd controller, Fuzzy based PD controller and optimized fuzzy PD controller are compared. The simulation model of those controllers is shown in Fig. 4.1.

4.2 MATLAB Simulation Results

The simulated results of the overall model are given in the responses as per the error occurring in different controllers respectively. They indicate the response of an individual controller. The simulated result of response of PD controller is shown in Figure 4.2.

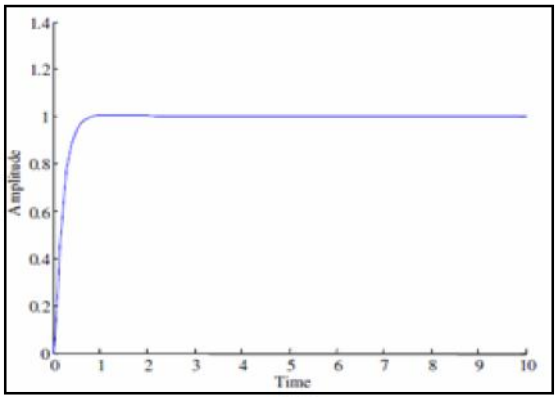


Fig. 4.2 Response of the conventional PD controller

The response in Figure 4.3 shows the controller output obtained from the fuzzy logic controller. From the response, we can infer that the system achieves the desired output. But the controller takes much time to reach the desired output in terms of settling time and rise time.

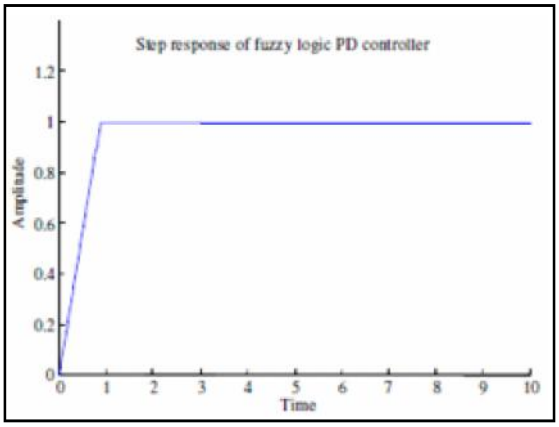


Fig. 4.3 Response of the fuzzy based PD controller system

The step response of Optimized fuzzy logic controller response is shown in figure 4.4.

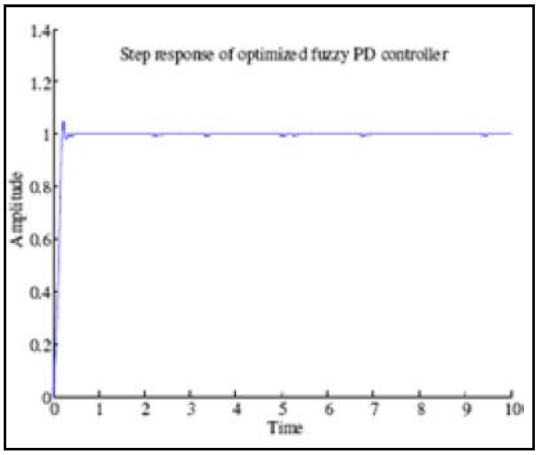


Fig. 4.4 Step response of optimized fuzzy logic controller

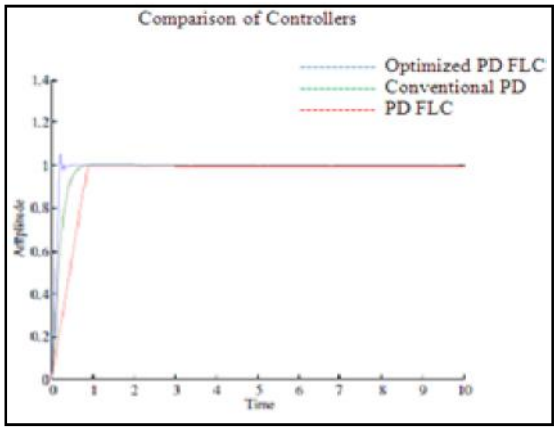


Fig. 4.4 Output response of different controllers

We can conclude that the Optimized fuzzy logic controller produces zero error and improves the system performance in terms of rise time and settling time. The reduced settling time of the controller gives the faster response compared to the fuzzy based PD controller.

The responses in Figure 4.4 show the comparison of conventional PD controller, Fuzzy based PD controller and the Optimized Fuzzy PD controller.

4.3 Comments on Results

The most desirable performance requires the controllers to have the smallest possible value for the rise time, overshoot and the settling time. It is also required for the final value should be as close as possible to the desired value which is unity.

From the table, it can be seen that the fuzzy logic controller can produce a desirable response performance with the use of only the proportional and Derivative Component (PD). This is contrary to the implementation of the conventional controller which requires the use of the proportional, derivative and integral component before a desirable or satisfactory response can be obtained. When compared to the conventional PD controller, the fuzzy logic PD controller shows a better performance in terms of overshoot while it exhibits a slightly lesser performance in terms of rise time and settling time. However, the results are changed when the fuzzy logic PD controller is further optimized.

The result of the GA-optimized fuzzy logic PD controller shows an outstanding performance in terms of achieving the desired value with very small values for the rise time and settling time. GA-optimized fuzzy logic PD controller produces a more desirable performance when compared to the conventional PD controller. At this point, it can be said that with optimal tuning measures, the fuzzy logic PD controller can perform better than the conventional PD controller.

Table 4.1 Performance Metrics for Conventional and Fuzzy Logic Controller

Parameters	Conventional PD Controller	Fuzzy Logic PD Controller	Optimized Fuzzy Logic PD Controller
Rise Time	0.3741	0.7094	0.1559
Overshoot	0.6748	0.0000	4.8434
Settling Time	0.6358	0.8735	0.8735
Final Value	1.0000	0.9958	0.9958

5. CONCLUSIONS

This study has succeeded in the design of an optimal Proportional Derivative fuzzy logic controller (PD-FLC) using genetic algorithm technique. The result was shown through simulation that the optimized fuzzy logic controller is performing better than a conventional PD controller when both controllers are subjected to the same operating conditions. The performance metrics taken into consideration are *the* overshoot, rise time, settling time and steady state error. MATLAB/SIMULINK model is used to fine-tune the controller parameters and simulate models. The simulation results show that the optimal fuzzy logic controller is functioning better than a conventional PD controller in terms of the rise and settling time. The output of the Optimized fuzzy logic controller can be used to drive the DC motor position control in industrial applications, robot manipulators, and home appliances.

REFERENCES

[1] D. Driankov, H. Hellendoorn and M. Reinfrank, "An Introduction to Fuzzy Control", Berlin, Germany: Springer Verlag, 1997.

[2] M. Kumar and D.P. Garg, "Intelligent Learning of Fuzzy Logic Controllers via Neural Network and Genetic Algorithm", Proceedings of the Japan-USA Symposium on Flexible Automation, July 19-21, Colorado, 2004, pp:1-8.

[3] M.W. Hwang and J.Y, Choi, "Hybrid Feedforward and Feedback Control of Wafer Temperature in RTP Using Genetic Algorithm and Fuzzy Logic", Proceedings of the IEEE International Conference on Intelligent Processing Systems, Oct. 28-31, IEEE Computer Society, Washington DC., USA., DOI: 10.1109/ICIPS.1997.672745, pp.93-97.

[4] C. C. Lee, "Fuzzy Logic in Control System: Fuzzy Logic Controller Parts I, II," IEEE Trans. Syst., Man, Cybern., Vol.20, Mar. 1990, pp.404-435.

[5] K.C.Ng and Y. Li, "Design of Sophisticated Fuzzy Logic Controllers Using Genetic Algorithms", Proceedings of the 3rd IEEE Conference on Fuzzy Systems, IEEE World Congress on Computational Intelligence, June 26-29, IEEE Computer Society, Washington DC., USA., pp: 1708-1712. DOI: 10.1109/FUZZY.1994.343598, 1994.

[6] M.Mohammadian and R.J. Stonier, "Tuning and Optimization of Membership Functions of Fuzzy Logic Controllers by Genetic Algorithms", Proceedings of the 3rd IEEE International Workshop on Robots and Human Commmunication, IEEE Computer Society, USA., DOI: 10.1109/ ROMAN.1994.365903, July 18-20, 1994, pp: 356-361.

[7] P.T.Chan, A.B.Rad and K.M. Tsang, "An Optimized Fuzzy Logic Controller", Proceedings of the 6th IEEE International Conference on Fuzzy Systems, July 1-5, IEEE Computer Society, USA., DOI: 10.1109/FUZZY.1997.622841, 1997, pp.975-980.

[8] C.Rekik, M. Djemel, N. Derbel and A. Alimi, "Design of Optimal Fuzzy Logic Controller with Genetic Algorithms", Proceedings of the IEEE International Symposium on Intelligent Control, Oct. 27-30, IEEE Computer Society, Washington DC. USA., DOI: 10.1109/ISIC.2002.1157745, 2002, pp. 98-103.

[9] J.Xiu, C. Xia and H. Fang, "GA-Based Adaptive Fuzzy Logic Controller for Switched Reluctance Motor Drive", Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21-23, IEEE Computer Society, Washington DC. USA., DOI: 10.1109/WCICA.2006.1713578, 2006, pp.8226-8230.

[10] W.S. Oh, Y.T. Kim, C.S. Kim, T.S. Kwon and H.J. Kim, "Speed Control of Induction Motor Using Genetic Algorithm Based Fuzzy Controller", Proceedings of the 25th Annual Conference of the IEEE Industrial Electronics Society, Nov. 29-Dec. 3, IEEE Computer Society, Washington DC., USA., DOI: 10.1109/IECON.1999.816464, 1999, pp.625-629.

[11] . . . . . Fuzzy Logic Controller", Masters Thesis, School of Electronic Engineering, Dublin City University. 14. Bucci, G., M. Faccio and C. Landi, 2000. New ADC with Piecewise Linear Characteristic: Case Study-Implementation of a Smart Humidity Sensor, 2003.

[12] W.R.Hwang and W.E. Thompson, "Design of Intelligent Fuzzy Logic Controllers Using Genetic Algorithms", Proceedings of the 3rd IEEE Conference on Fuzzy Systems, IEEE World Congress on Computational Intelligence, June 26-29, IEEE Computer Society, Washington DC., USA., DOI: 10.1109/FUZZY.1994.343566, 1994, pp: 1383-1388.

[13] C.N. Ko, T.L. Lee, Y.Y. Fu and C.J. Wu, "Simultaneous Auto-Tuning of Membership Functions and Fuzzy Control Rules Using Genetic Algorithms", Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Oct. 8-11, IEEE Computer Society, USA., DOI: 10.1109/ICSMC.2006.384547, 2006, pp.1102-1107.

[14] Y.S. Lu and C.M. Cheng, "Design of a Non Overshooting PID Controller with an Integral Sliding Perturbation Observer for Motor Positioning Systems", JSME Int. J. Series C., 48: 103-110. http://www.jstage.jst.go.jp/article/jsmec/48/1/48_103/_article, 2005.

Wear and Emission Studies on Pungam Methyl Ester Blended With 2T Lubricating Oil

G. Senthil Kumar[1], K Balamurugan[2], P Karthi[3], R Karthi[4], S Karuppusamy[5]

[1,3,4&5]Department of Mechanical Engineering, Bannari Amman Institute of Technology, Sathyamangalam - 638 401, Erode District, Tamil Nadu

[2]Department of Mechanical Engineering, Institute of Road and Transport Technology, Erode - 638 316 Tamil Nadu

E-mail: senthil_3142@yahoo.co.in

Abstract

Much effort has been initiated on research and development of new types of lubricating oils to reduce wear, friction and corrosion in engine applications. Vegetable oils are based on soya bean, sunflower, castor, rapeseed, corn, pungam, canola and cotton seed. The vegetable lubricants are environmentally friendly alternative to mineral oils since they are biodegradable. The vegetable oils have many advantages like high viscosity index, low friction coefficient, high flash point, low volatile etc., over mineral oils. Pungam Oil Methyl Ester (POME) based bi-odiesel is a viable alternative to fossil fuels. In this project, POME and manufacturer's recommended oil mixed in definite proportions were tested as two stroke crankcase lubricants. ASTM four ball wear tests were conducted with POME to improve the property of the oil, which ensures less wear in the engine and the emission analysis for smoke was conducted on the definite proportion along with the manufacturer's recommended oil at various ratio to compare and analyze the effective environment friendly lubricant which could be used as an alternative. Tribological studies ensure less wear in engine and emission test proves less adverse effects on environment.

Key Words: *Vegetable lubricants, Pungam Oil Methyl Ester, ASTM four ball wear test, Tribological study, Emission analysis*.

1. INTRODUCTION

In India, non-edible oils like pungam oil and jatropha oil are available in abundance, which can be converted into methyl ester. The methyl esters of vegetable oils, known as biolubricant are becoming increasingly popular because of their low environmental impact and potential as a green alternative fuel for diesel engine and they would not require significant modification of existing engine hardware. Methyl ester of Pungam, Jatropha and Neem are derived through transesterification process. A well-known transesterification process made bio lubricant, Pungam seed oil was selected for bio lubricant production. Vegetable oils can be transesterified by heating them with a large excess of anhydrous methanol and an acidic or basic reagent as catalyst. Both the acid as well as alkaline esterification was subsequently performed to get the final product. In a transesterification reaction, a larger amount of methanol was used to shift the reaction equilibrium to the right side and produce more methyl esters as the proposed product.

Table 1 Properties of Pungam oil

PROPERTIES	VALUE
Water Content	0.05%
Specific Gravity	0.9366
Density	0.9358 gm/cc
Carbon Residue	0.80%
Ash Content	0.05%
Flash Point	212^0C
Fire Point	224^0C
Acid Value	16.8
Iodine Value	86.5
Boiling Point	330^0C
Cloud Point	20^0C
Pour Point	-40^0C
Calorific Value(Kcal/kg)	8742

2. LITERATURE REVIEW

R.K.Singh et al [1] has discussed on the fatty acid methyl ester which is derived from triglycerides by Transesterification. The acid value of pungam oil was more than 3, so it was converted to biolubricant by esterification followed by transesterification process. N.Prakash et al [2] has detailed about the Pungam oil was treated with a lower alcohol (methanol) in the presence of a base catalyst (KOH) to yield methyl esters of fatty acids (biolubricant) and glycerine. Transesterification of Pungam oil has been carried out using KOH catalyst. The optimum parameters for using KOH as catalyst were amount of catalyst 1.5 g, volume of methanol 45 ml, temperature 80°C and reaction time 60 min. A.K.Singh et al [3] Smoky emissions from two-stroke gasoline engines (2T) are a problem for the environment. Use of vegetable oil (oxygenate) is one solution. A biodegradable 2T-oil was developed from castor oil. Evaluation revealed that on one hand it reduced smoke by 50–70. T. Venkateswara Rao *et al* [4]. Experimental investigations have been carried out to examine properties, performance and emissions of different blends (B10, B20, and B40) of PME, JME and NME in comparison to diesel. T. Mohan Raj et al [5] Pungam seed oil is non-edible oil thus food versus fuel conflict will not arise if this is used for bio-lubricant production. A maximum of 75% bio lubricant was produced with 20% methanol in the presence of 0.5% sodium hydroxide. A. Veeresh Babu et al [6] Biodiesel was prepared from the non-edible oil of Pongamia pinnata by transesterification of the crude-oil with methanol in the presence of NAOH as catalyst.. A catalyst is usually used to improve the reaction rate and yield. NaOH was found to be a better catalyst than KOH in terms of yield. K.V. Thiruvengadaravi *et al* [7] Pre treatment of high free fatty acid containing Pungam pinnata oil using sulfuric acid catalyst has been optimized. Based on the experimental results, a methanol to oil ratio of 9:1, one percentage catalyst by weight, and a temperature of 60^0C, were selected as the optimum settings for the esterification process. P. V. Rao et al [8] The higher levels of NOx emission is attributed to the combined effect of bulk modulus, cetane number, oxygen, and unsaturated fatty acids. Decrease in premixed combustion and increase in diffused combustion is observed with preheated methyl ester. The reduction in peak value of premixed combustion is leads to the reduction of NOx levels. Sanjib Kumar Karmee et al [9] Methyl ester was prepared from the non-edible oil of Pungam pinnata by transesteriûcation of the crude oil with methanol in the presence of KOH as catalyst. A maximum conversion of 92% (oil to ester) was achieved using a 1:10 molar ratio of oil to methanol at 60^oC. Tetrahydrofuran (THF), when used as a co-solvent increased the conversion to 95%. Anand Kumar Pandey et al [10] Pungam oil was converted into methyl ester by the transesterification process. It involves making the triglycerides of Pungam oil to react with methyl alcohol in the presence of a catalyst (KOH/ NaOH) to produce glycerol and fatty acid ester.

3. EXPERIMENTAL SETUP

The experimental setup contains a 500 ml two-necked round-bottomed flask which was used as a reactor. The flask was placed in a water bath, whose temperature could be controlled within ±2 ^{0}C. One of the two side necks was equipped with a condenser and the other was used as a thermo well. A thermometer was placed in the thermo well containing little olive oil for temperature measurement inside the reactor. A magnetic stirrer was used with speed regulator for adjusting and controlling the stirrer speed. For production of Pungam Oil Methyl Ester (POME) esterification and transesterification process was carried out.

3.1 Esterification

A 250 ml of pungam oil was preheated up to 63^0C to remove the water content in the oil. 0.5 ml of sulphuric acid and 30 ml of methanol were added to the oil and stirred continuously maintaining a steady temperature of 65^0C up to 2hours. After the 2hour reaction the products were cooled and transferred into separating funnel. The free fatty acid was separated by separating funnel. This oil sample was used in transesterification to obtain methyl esters.

3.2 Tranesterification

In the same setup, the obtained esterified pungam oil from the esterification process was charged. The esterified oil was preheated up to 63^0, 1.425 gram of catalyst NaOH was made to dissolve into 30 ml of methanol along with the catalyst solution was added to the oil sample. The system was maintained airtight to prevent the loss of alcohol. The reaction mix was maintained at temperature just above the boiling point of the alcohol i.e. around 65°C for a period of 2hours. After the confirmation of completion of methyl ester formation, the heating was stopped and the products were cooled and transferred to a separating funnel. And lower glycerol layer was separated.

FLOW DIAGRAM FOR PREPARATION OF POME

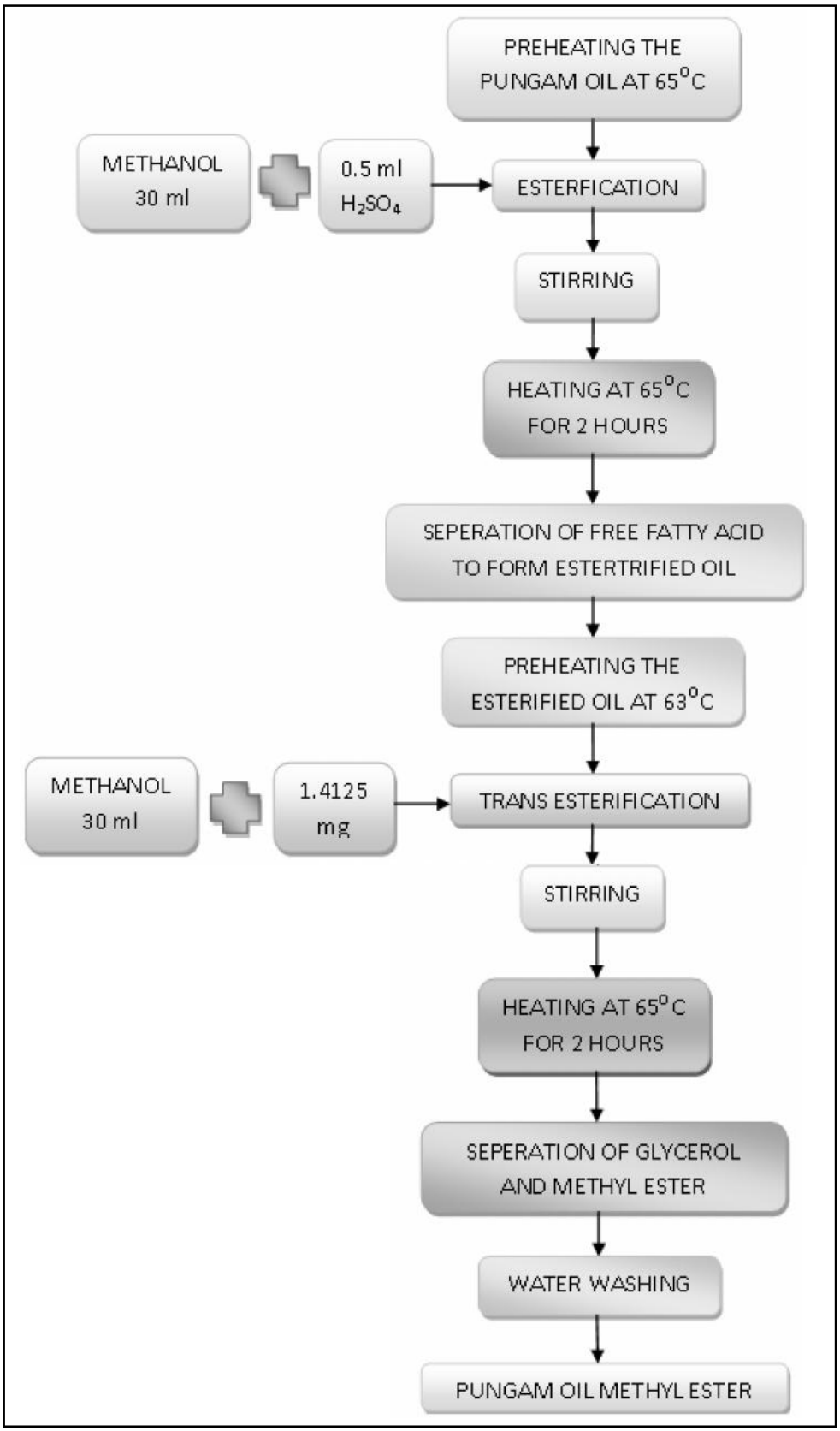


Fig.1 Flow diagram for preparation of POME

4. WATER WASHING

To remove un-reacted methoxide present in raw methyl ester, it is purified by the process of water washing with a preheated tap water at 50°c. The water wash has been done periodically at an interval of every 24 hours. The water wash has been carried out eight times to separate the glycerol from the methyl ester. The methyl ester produced from Pungam oil is known as pungam oil methyl ester (POME).

4.1 Drying

The obtained pungam oil methyl ester contains the water molecules due to water washing. In order to remove the water molecules the pungam oil methyl ester was placed in an oven and maintain a temperature about 65°c for a one hour.

5. TRIBOLOGICIAL STUDIES

Tribologicial studies are nothing but the analysis of wear and friction of lubricating oil. Here the wear and friction characteristics of lube oils are evaluated by four-ball wear and friction test apparatus.

A four-ball test apparatus for evolution of friction and wear characteristics of lubricants is disclosed. The test apparatus includes a motor and adjustable weight assembly that is supported by the upper test ball on the lower test balls positioned in a test cup. All the weight of the motor/weight assembly is supported by the test ball itself, and thrust bearings are not required for support of the test cup. The four-ball test apparatus eliminates numerous sources of error found in prior four-ball testers, and permits various instruments or fluid lines to be connected to the test cup without introducing errors into the test measurements.

5.1 Four-Ball Wear Test

One of the most important functions of any motor oil is wear protection. Because motorcycle engines operate under more severe operating conditions than automobiles, the ability of motorcycle oil to deliver adequate wear protection is especially important. The ASTM D-4172 Four-Ball Wear Test is the standard test used to determine a lubricant's ability to minimize wear in metal-to-metal contact situations. Three steel balls are secured and placed in a triangular pattern within a bath of the test lubricant. With load, speed and temperature kept constant, a fourth ball sits atop the other balls and is rotated and forced into them for one hour. Following the test, the lower three balls are inspected for wear scars at the point of contact. The diameters of the wear scars are measured and the results are reported as an average of the three scars. The lower the average wear scar diameter, the better the wear protection properties of the oil.

6. ASTM (D 4172-4194) STANDARD TEST METHOD FOR WEAR PREVENTIVE CHARACTERISTICS OF LUBRICATING FLUID (FOUR-BALL METHOD)

This method covers a procedure for making a preliminary evaluation of the anti-wear properties of fluid lubricants in sliding contact by means of the four-ball test machine. Evaluation of lubricating grease using the same machine is detailed in the test method D2266. The values stated in either inch-pound units or SI units are too regarded separately as standard. Within the test the inch-pound units are shown in brackets. The values stated in each system are not exactly equivalent; therefore each system must be used independently of the other. Combining values of the two systems may result in non-conformance with the specifications.

This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety and health practices and determine the applicability of regulatory limitations prior to use.

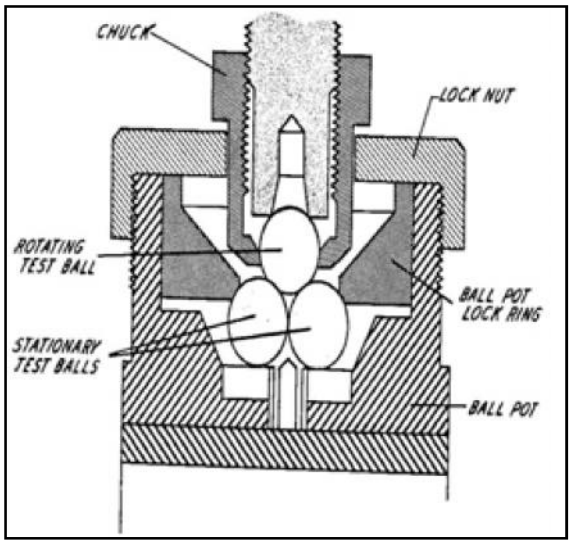


Fig.2 Conceptual setup

Fig.3 Actual setup

SERVO 2T oil

95% of 2T oil blended with 5% POME

Fig.4 Four Ball Wear Test Results

6.1 Optimistic Result Analysis

After conducting wear test in four-ball wear testing machine, it is evident that the ball using 2T oil as lubricants has excessive wear when compared to the ball using pungam methyl ester as lubricant. Moreover the wear seems to be much more reduced. When the pungam methyl ester has 1.5% additives are added to it.

This test therefore proves that the pungam methyl ester oil have better tribologicial properties than 2T oil for 2 stroke petrol engines while using the correct ratio of additives.

7. CHEMICAL TEST

Volatility is the major determinant of the tendency of hydrocarbon to produce potentially explosive vapours. The important tests are :

8. EMISSION TEST

Air pollution can be defined as an addition of any material which will have a deleterious effect on life upon the earth. Air pollution is increasing drastically with increase in number of vehicles, specially two wheelers, because they are the most economic way of transport in India. Theoretically, if the combustion in the combustion chamber is complete, the exhaust will contain only carbon dioxide (CO_2) and water vapour (H_2O). But when the fuel quantity supply is more, there is no sufficient oxygen available for complete combustion and part of the carbon converts into CO. Whereas, fuel quantity supplied is less, excess oxygen is made available to react with nitrogen which results in formation of NO. The different constituents which are exhausted from petrol engine and different factors which affect the formation are discussed below.

Table 2 Emission Test Result for POME Blended With 2T Oil

2T OIL %	PUNGAM OIL METHYL ESTER %	CO %	HC (PPM)	CO_2 %	NO_X %
100	0	4.30	8726	2.6	682
95	5	4.26	8725	4.35	429
90	10	4.35	8820	3.52	474
85	15	4.39	8905	3.90	408
80	20	4.41	8952	3.05	462
75	25	4.48	8987	2.69	387

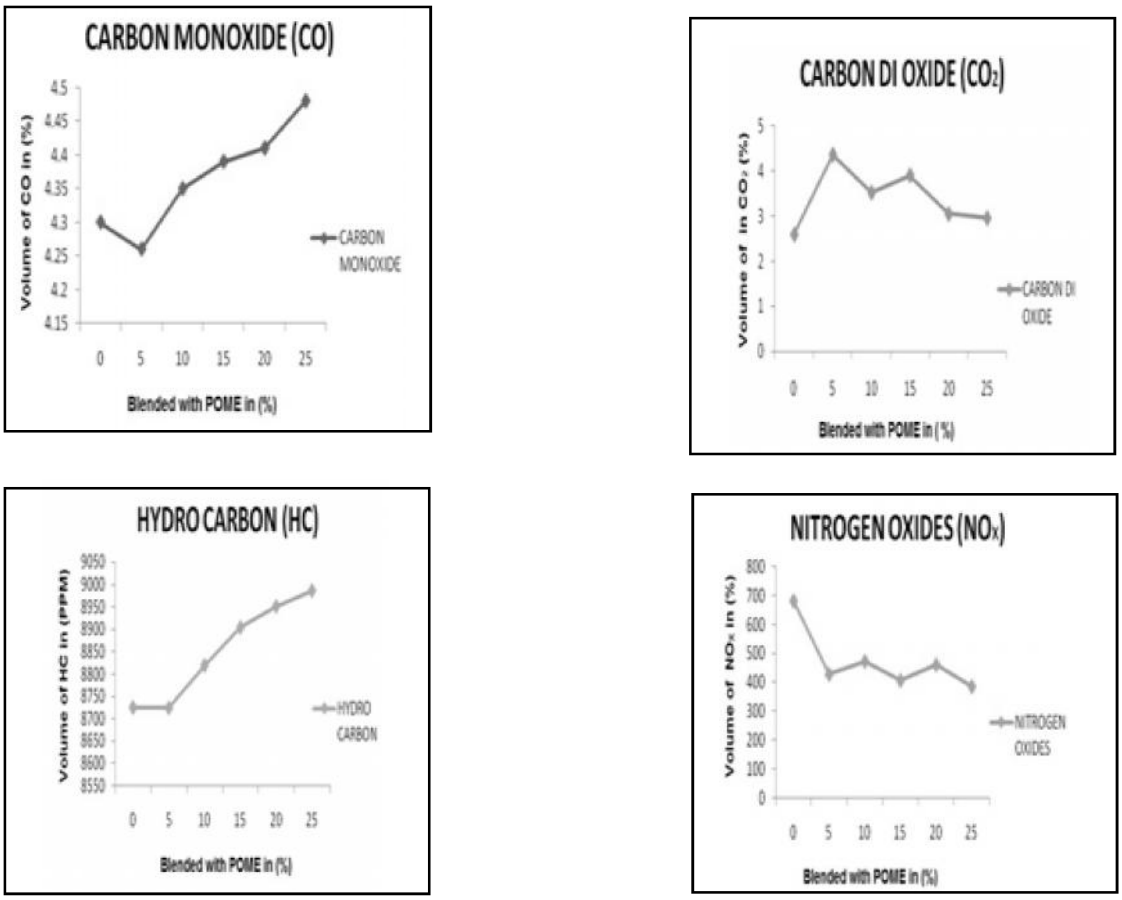


Fig. 5 Emission test result graph for POME blended with 2T oil

9. CONCULSION

The depletion of fossil fuel is one of the wildest challenges for the automobile sector. Numerous researches are already going on discovering the alternative fuels. This project concentrates on modifying the existing lubricant oil of two stroke engines by blending vegetable oil (pungam oil) with 2T oil which possesses high viscosity.

The 2T lubricating oil is blended with POME at various proportions and with additives. To examine the friction and wear behavior of the engine the blend is prepared in lab and the tribologicial study proves that the friction is reduced with comparing to the commercially available lubricants.

As per the Bharath Standard Norms the two stroke engine emissions are tested with the various blends. The emission analysis reveals that 95% 2T oil + 5% POME can be used in field with less particulates emission at the exhaust.

REFERENCES

[1] R.K.singh, A.Kiran kumar and S.Sethi , "Preparation of Karanj Oil Methyl Ester" National Institute of Technology- Rourkela, April-May 2006.

[2] N.Prakash, A.Arul jose, M.G.Devanesan and T.Viruthagiri, "Optimization of Pungam Oil Tranesterification" Indian Journal of Chemical Technology, Vol.13, 2006, pp.505-509.

[3] A.K.Singh, "Castor Oil Based Lubricant Reduces Smoke Emission in Two Stroke Engine" Industrial Crops and Product, Vol.33, 2011, pp. 287-295.

[4] T.Venkateswara Rao, G.Prabhakar Rao and K.Hema Chandra Reddy, "Experimental Investigation of Pungam, Jatropha and Neem Methyl Esters as Biodiesel on C.I. Engine" Jordan Journal of Mechanical and Industrial Engineering, Vol. 2, 2008, pp.117-122.

[5] T.MohanRaj**,** K. Murugu Mohan Kumar and Perumal Kumar, **"**Biodiesel from Pungam Seed Oil and Its Effects on Engine Performance with a Computerized Engine Test Rig" Pertanika Jouranal of Science & Technology, Vol.19, No.1, 2011, pp.117-127.

[6] A.Veeresh Babu, B.V.Appa Rao and P. Ravi Kumar, "Transesterification for the Preparation of Biodiesel from Crude-Oil of Pungam Pinnata" Thermal Science, Vol.13, No.3, 2009, pp.201-206.

[7] K.V.Thiruvengadaravi, J.Nandagopal, V. Sathya Selva Bala, S. Dinesh Kirupha, P. Vijayalakshmi and S. Sivanesan, "Kinetic study of the Esterification of Free Fatty Acids in Non-edible Pungam Pinnata Oil Using Acid Catalyst" Indian Journal of Science and Technology, Vol.2 No.12, Dec. 2009.

[8] P.V.Rao, "Effect of Properties of Pungam Methyl Ester on Combustion and NOx Emissions of a Diesel Engine" Journal of Petroleum Technology and Alternative Fuels, Vol. 2, No.5, 2011, pp. 63-75.

[9] A.Sanjib Kumar Karmee and Anju Chadha "Preparation of Biodiesel from Crude Oil of Pungam", Pinnata Bioresource Technology, Vol. 96, 2005.

[10] Anand Kumar Pandey and M R Nandgaonkar, "Experimental Investigation of the Effect of Esterified Pungam Oil Biodiesel on Performance, Emission and Engine Wear of a Military 160HP Turbocharged CIDI Engine", Proceedings of the World Congress on Engineering 2011, Vol.3, 2011.

Semantic Indexing of Text Documents Using Domain Knowledge

S. Logeswari[1] and S. Narmadha[2]
[1&2]Department of Computer Science and Engineering, Bannari Amman Institute of Technology,
Sathyamangalam - 638 401, Erode District, Tamil Nadu
E-mail: slogesh76@gmail.com

Abstract

Fast retrieval of the appropriate information from the database has always been a significant issue. In order to retrieve the relevant information efficiently from database we propose a new method called semantic indexing. Eye diseases in MeSH ontology is taken as the domain reference. A topic-based dynamic weighting scheme is used to index the text in the document. The topic weight is calculated based on the frequency of the term, child nodes and synonyms that are presented in the ontology. Topic frequency is used to determine the importance and existence of the topics. Experimental shows that proposed semantic indexing outperform the term based method.

Keywords: *MeSH ontology, Topic weight, Topic frequency, Semantic indexing, Tf-idf*

1. INTRODUCTION

A cluster is a anthology of data objects that are related to one another. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Clustering is also called as data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. As a data mining function, cluster analysis can be used as a stand-alone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. The purpose of text mining is to process unstructured information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining algorithms.

Document clustering has many imperative applications in the area of data mining and information retrieval. Many of the existing document clustering techniques uses the "bag of words" model to represent the content of a document. However, this representation is only effective for grouping related documents when these documents share a large proportion of lexically equivalent terms. In other words, instances of synonymy between related documents are ignored, which can reduce the effectiveness of applications using a standard full-text document representation. A novel semantic based mining model is proposed. The goal of this approach is to mine text through the analysis of higher level characteristics (called concepts), and minimizing the vocabulary problems. Instead of applying text mining techniques on terms or keywords labeling or extracted from texts, the discovery process works over concepts extracted from texts. After that, mining techniques are applied over the concepts discovered.

2. RELATED WORKS

Vector Space Model (VSM) is used to represent text as a vectors of identifiers in the term-based representation. The first is term based model in this The most popular Term Frequency_Inverse Document Frequency (TF_IDF) is often considered as the default weighting scheme. However, this scheme is pure statistical and does not incorporate any information about semantic or category that belongs to a term. In the phrase based model, the phrases are converted into atomic units. The phrases are represented either as a suffix tree or as a document index graph. The suffix tree is a data structure that presents the suffix in a way that allow for particularly fast implementation of many important string operations. The document index graph index the web document based on the phrases rather than single terms. The limitations of the existing term based and phrase based models make the clustering process as a challenging task. In VSM representation, the order in which the terms appear in the document is lost. Synonymy and polysemy problems are ignored by the conventional methods. A concept based mining method based on core words of a class [2] is proposed. The text categorization methods mainly focus on keywords, which cannot deal with synonyms and polysemy scenario properly. In this method, the core words are identified and extracted. How-net maps are used to map keyword

space to concept space based on these core words, and finally complete the text categorization process in the concept space. Both Naive Bayes and k-Nearest Neighbour text categorization methods are used to evaluate the performance of this method. The results indicate that the new core words oriented concept mapping can effectively improve text categorization precision compared with other keywords oriented text categorization methods.

In this paper, we propose a semantic based indexing scheme which uses the domain knowledge as reference for extracting the meaning of the terms. In this proposed scheme, the descriptor terminologies that are representing the topics in the particular ontology are getting stored along with their identifiers and synonyms. A topic-based weighting scheme is used to index the text in the document. Documents are indexed based on the terms that are presented in the ontology.

3. TOPIC BASED INDEXING

Indexing of document based on related or semantically related keyword. Topic based weighting scheme is proposed to index the text. It involves with identifying topic candidates, determine their importance, and detect similar and synonymous topics. The indexing algorithm use topic frequency to determine their importance and existence of topics.

Unlike traditional TF-IDF weighting scheme, a topic based weighting scheme computes the importance of the underlying text by converting the documents into a bag of concepts. Medical Subject Headings (MeSH), published by the National Library of Medicine mainly consists of the controlled vocabulary and a MeSH Tree. The controlled vocabulary contains several different types of terms, such as Descriptor, Qualifiers, Publication Types, Geographic, and Entry terms. MeSH descriptors are organized in a MeSH Tree, which can be seen as a MeSH Concept Hierarchy.

3.1 Index Table

In the proposed method, an index table consists of all descriptor terms in the ontology for Neoplasm in MeSH ontology.

Index_table = (ID, descriptor terms, entry terms, value)

Here ID represents the term number; descriptor terms are main concepts or main headings. Entry terms are the synonyms or the related terms to descriptors. Value is represented as *n* in a tree and it denotes parent, augmented by the index of *n* among its siblings, adding dot to separate them. For example, in Figure1.

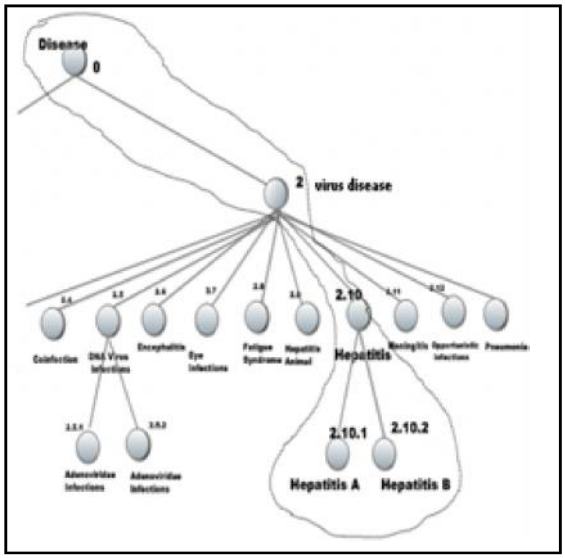


Fig.1 Descriptor terms in the form of tree

Based on the keyword, the whole path is retrieved from the root to the leaf and their weights are assigned dynamically based on parent and child relationship. For example the path retrieved for the keyword Hepatitis is also shown in the figure1.

3.2 Topic Based Weighting Scheme

A topic-based weighting scheme is used to index the text in the document. Documents are indexed based on the terms that are presented in the ontology. Tokenization and stop word removal are done as the pre-processing steps in the clustering process. A keyword and an abstract are given as input to the proposed method. The document is getting pre-processed and the words are getting stored in a text file. The given keyword is searched in the MeSH ontology for its existence. If it exists, the corresponding path and its synonyms will also get captured. All the terms in the ontology are compared with the given abstract. Topic frequency is used to determine the importance and existence of the topics. The weight of the abstract is calculated by using the frequency of the term and their weight. The total weight S for the topic is calculated is shown in equation (1).

$$S(topic) = \frac{\sum_{i=1}^{n} S_{word}(w_i)}{N} \quad (1)$$

The weight of the individual words can be calculated as follows:

$$S_{word}(w_1) = nR^{i}_{w_1} \text{ x } SR^{i}_{w_1} + nR^{i}_{w_1} \text{ x } SR^{k}_{w_1} \quad (2)$$

Where N denotes the number of unique words in the abstract,

k = represents the relation of the word either identity or synonymy,

$nR^{k}_{w_1}$ =represents the number of occurrence of the relations and

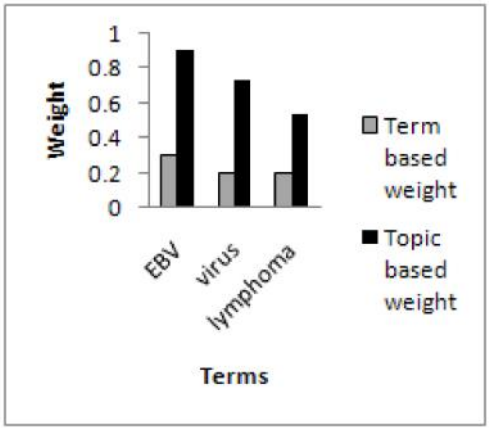
$SR^{k}_{w_1}$ =denotes the weight assigned to that relation.

The highest score represents the importance of the word in the abstract.

4. EXPERIMENTAL RESULTS

For the experimental purpose, twenty five documents are collected from MEDLINE and preprocessed.The terms are compared with the MeSH ontology and for its existence. If it exists, the keyword is assigned with a static weight as one and its children are assigned with lesser values level by level. The weight of an individual document is calculated based on their frequency for term based indexing. For the topic based indexing, weight of an individual document is calculated based on the frequency of the terms, their synonyms and the static weights using the equation (1). The higher topic weight shows that the document is more related to that topic.

Table 1 Comparison of Term Based and Topic Based Weights



5. CONCLUSION

Semantic based indexing reduces the dimensionality of the data which is efficient even for very large databases and provides an understandable description of the discovered clusters by their frequent term sets. But in the existing system, individual terms were considered for indexing and also only two relations, identity and synonymy were used for calculating the weight of the topics. Based on the topics, the background concepts can be identified. One of the future directions is to use a concept based similarity measure for improving the clustering process.

REFERENCES

[1] B.Tahayn, R.K.Ayyasamy, S.Alhashmi and S.Eu-Gene, "A Novel Weighting Scheme for Efficient Document Indexing and Classification", journal of IEEE International Conference on Information Technology. Vol. 2, 2010, pp. 783-788.

[2] F.Shehata and Mohamed S.Kamel, "An Efficient Concept Based Mining Model for Enhancing Text Clustering", journal of IEEE Transactions on Knowledge and Data Engineering, Vol. 22, 2010, pp.1360-1371.

[3] Y.Takeru, Y.Hidekazu and O.Sigeru, "Information Filtering Using Index Word Selection Based on the Topics", World Academy of Science, Engineering and Technology, Vol.50, 2009, pp.296-302.

[4] B.Simona, S.Nefti and Y.Rezgui, "A Concept Based Indexing Approach for Document Clustering", Journal of IEEE International Conference on Semantic Computing, 2008, pp.26-33.

[5] A.Amine, Z.Elberrichi, L.Bellatreche, M.Simonet and M.Malki, "Concept-Based Clustering of Textual Documents Using SOM," journal of IEEE Transactions on Text Document, 2008, pp.156-163.

[6] S.Zhifang and L.Yao, "Inducting Concept Hierarchies From Text Based on FCA", Fourth International Conference on Innovative Computing, Information and Control, 2009, pp. 1080-1083.

[7] S.Shehata, F.Karray and M.Kamel, "Enhancing Text Clustering Using Concept-based Mining Model", Proceedings of the Sixth International Conference on Data Mining (ICDM'06),journal of IEEE Transactions on Text Document, 2008, pp.1043-1048.

[8] M. LAN, C.L. Tan, H.B. Low, "Proposing a New Term Weighting Scheme for Text Categorization",

Proceedings of the National Conference on Artificial Intelligence, 2006, pp.285-289.

[9] M. Lan, S.Y. Sung, H.B. Low, and C.L.Tan, "A Comparative Study on Term Weighting Schemes for Text Categorization", Proceedings of the International Joint Conference on Neural Networks (IJCNN-05), 2005, pp.546-551.

[10] J.Weston, S.MukheIjee, O.Chapelle, M.Pontil, T.Poggio and V.Vapnik, "Feature Selection for SVMS", In Proceedings of the Advances in Neural Information Processing Systems,Vol. 13, 2000, pp. 398-410.

Multi-query Optimization of SPARQL Using Clustering Technique

R.Gomathi[1], C.Sathya[2] and D.Sharmila[3]

[1&2]Department of Computer Science and Engineering, [3]Department of Electronics and InstrumentationEngineering,
Bannari Amman Institute of Technology, Sathyamangalam - 638 401, Erode District, Tamil Nadu
E-mail: gomsbk@gmail.com

Abstract

A W3C standard for processing RDF data is a SPARQL query language, a technique that is used to encode data in meaningful manner. We investigate the foundations of SPARQL query optimization by grouping into individual clusters using common substructures in the multiple SPARQL queries, propose a comprehensive set of query rewriting rules for the clustered group and finally Query execution provide the final result of optimized query. The proposed technique is efficient and scalable for multiple SPARQL query.

Keywords: *Common substructures, Multiple SPARQL, MQO, RDF*

1. INTRODUCTION

RDF is the data format of interlinked data. RDF is a directed, labeled graph data format for representing information in the Web. RDF is an essence of triple format namely subject, predicate and object.

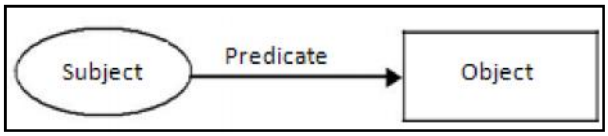


Fig 1.1 Triple format representation

SPARQL a query language and a protocol for retrieving RDF data which has been formulated and designed by the W3C RDF Data Access Working Group. SPARQL is a query language for pattern matching for RDF graphs. SPARQL syntax is similar to SQL, but SPARQL is more powerful, which enables queries spanning multiple disparate (local or remote) data sources containing heterogeneous semi structured data.

The SPARQL query language is related to the following specifications:

i. The SPARQL Protocol for RDF [SPROT] specification defines the remote protocol for issuing SPARQL queries and receiving the results.
ii. The SPARQL Query Results XML Format [RESULTS] specification defines an XML document format for representing the results of SPARQL SELECT and ASK queries.

SPARQL takes the description of what the application wants, in the form of a query, and returns that information, in the form of a set of bindings or an RDF graph.

Query optimization is the most critical phase in query processing. Multi Query optimization is a technique in which multiple query plans for satisfying a query are examined and a good query plan is identified. Complexity arises in MQO which leads to NP-Hard. There may be many plans to find the best strategy. Cost based query optimizers evaluate the resource of various query plans and use the basis for plan selection using algorithms. The search space can become quite large depending on the complexity of the SPARQL query.

Complex queries are becoming common, due to the advent of technological tools that help examine information from large data stores. These complex queries share a lot of common sub-expressions since i) extensive views for different query that share a common value ii) There are nested queries that are correlated where outer query and inner query variables are not common but form a common sub-expression. Keeping the above challenges we design a framework for MQO with the following contributions:

i. Summarize the similar pattern in the SPARQL query.
ii. Summarized patterns will be clustered based on the common substructures.
iii. Clustered queries will be rewritten and finally query execution is performed.
iv. Experiments prove that the model is very efficient and scalable.

2. RELATED WORK

Complex queries are becoming common in decision support systems. These complex queries have a lot of common sub-expressions [1], either within a single query, or multiple queries. Multiquery optimization exploits common sub-expressions to reduce evaluation cost. Three cost-based heuristic algorithms: Volcano-SH and Volcano-RU, which are based on simple modifications to the Volcano search strategy, and a greedy heuristic is used'. A performance study of comparing the algorithms, using workloads consisting of queries from the TPC-D benchmark. The study shows that algorithms provide significant benefits over traditional optimization, at a very acceptable overhead in optimization time.

The problem of *Basic Graph Pattern* (BGP) optimization for SPARQL queries and *main memory* graph implementations of RDF data is formalized. The characteristics of heuristics for selectivity based static BGP optimization are studied. Customized summary statistics for RDF data enable the selectivity estimation [2] of *joined* triple patterns and the development of efficient heuristics. Using the Lehigh University Benchmark (LUBM), the performance of the heuristics for the queries provided by the LUBM is discussed.

Efficient management of RDF data is an important factor in realizing the Semantic Web vision. Drawbacks are becoming increasingly pressing as Semantic Web technology is applied to real-world applications. Current data management solutions for RDF data does not scale properly, and explore the fundamental scalability limitations [3] of these approaches are examined. Improving performance for RDF databases using property tables is analysed. Vertically partitioning approach is used to study the RDF data. Further, column-oriented DBMS is used which shows an increase in performance magnitude, with query processing time is reduced.

The salient points of RDF-3X are: 1) a generic solution for storing and indexing RDF triples 2) a powerful yet simple query pro- cessor that leverages fast merge joins to the largest possible extent, and 3) Choosing optimal join orders is executed using query optimizer through which a cost model based on statistical synopses for entire join paths is identified. The performance of RDF-3X, [4] in comparison to the previously best systems, has been measured on several datasets with more than 50 million RDF triples and benchmark queries that include pattern matching and long join paths in the underlying data graphs.

BitMat introduces –(i) a compressed bit-matrix structure for storing huge RDF graphs, and (ii) a novel, light-weight SPARQL join query processing method that employs an initial pruning technique along with variable-binding-matching algorithm on BitMats [5] to produce the final results. Query processing method does not build intermediate join tables and works directly on the compressed data. Results show that the competing methods are most effective with highly selective queries. On the other hand, BitMat delivers 2-3 orders of magnitude better performance on complex queries over massive data.

Loosely-structured Exploratory queries requires only minimal user knowledge of the source network. Exploratory query evaluation usually [6] involves the evaluation of many distributed queries. The optimization problem for exploratory queries is overcome by proposing several multi-query optimization algorithms that compute a global evaluation plan which minimizes the total communication cost, a major bottleneck in distributed queries. The algorithms proposed are necessarily heuristics, as computing an optimal global evaluation plan is shown to be np-hard. Finally, an implementation of our algorithms and its illustrations shows their potential not only for the optimization of exploratory queries, but also for the multiquery optimization of large set queries is presented.

A set of novel query processing techniques, for large number of XML stream queries involving value joins over multiple XML streams [7] and documents referred to as *Massively Multi-Query Join Processing* techniques is proposed. These techniques enable the sharing of representations of inputs to multiple joins, and the sharing of join computation. These techniques are also applicable to relational event processing systems and publish/subscribe systems that support join queries. Experimental results to demonstrate the effectiveness of our techniques is presented. Thousands of XML messages with hundreds of thousands of join queries on real RSS feed streams is processed. Techniques gain more than two orders of magnitude speedup compared to the naive approach of evaluating such join queries.

Queries with common sequences of disk accesses can make maximal [8] use of a buffer pool. We developed a middleware to promote the necessary conditions in concurrent query streams, and achieved a speedup of

2.99 in executing a workload derived from the TCP-H benchmark.

Clustering is the unsupervised classification of patterns into groups (clusters). The clustering problem [9] has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. Clustering is a difficult problem in common. An overview of pattern clustering methods is presented. A taxonomy of clustering techniques, identifying cross-cutting themes and recent advances is also proposed. It also describes some important applications of clustering algorithms such as segmentation of the image, recognition of object, and information retrieval of information.

3. PROPOSED ARCHITECTURE

MultiQuery Optimization mainly involves Query Processing, Query Rewriting and Execution as shown in the Figure 3.1.

Query Processing: Query Processing converts the SPARQL query into query graph pattern which is equivalent to the query. This query graph pattern presents the query execution in sequence and optimization of the query takes place.

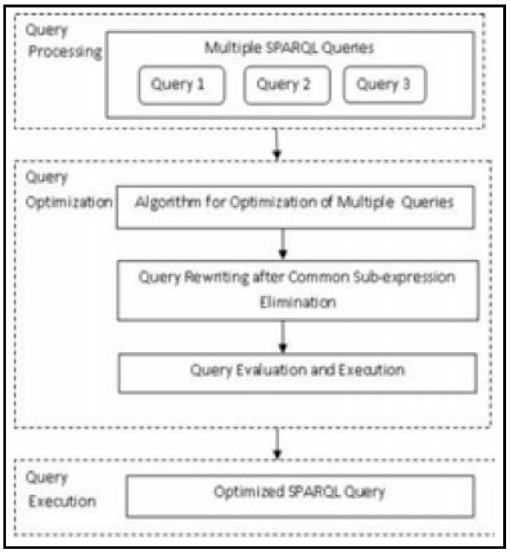


Fig.1 Steps in multiquery optimization

Query Optimization and Execution: Selecting the best strategy for query processing is Query Optimization. Algorithm finds the best query processing strategy. The steps involved in the algorithm are as follows: Consider the data in the table (Fig.2) and the given SPARQL query.The query corresponds to triples Faculty and Course which has an equivalent value GraduateCourse0 as object. First OPTIONAL field in the query returns object of predicate Institute, if predicate exists. Second OPTIONAL field in the query returns object of predicate Student, if its predicate exists correspondingly. When the query is evaluated over the Input data D, it results in Q_{OPT} as shown in the Figure 3.

Subject	Predicate	Object
FullProfessor0	Faculty	FullProfessor
FullProfessor0	Institute	Organisation
FullProfessor0	Course	GraduateCourse0
FullProfessor0	Student	GraduateStudent
FullProfessor1	Faculty	PostDoctrate
FullProfessor1	Course	GraduateCourse0
FullProfessor1	Student	GraduateStudent
FullProfessor2	Faculty	FullProfessor
FullProfessor2	Institute	Organisation
FullProfessor2	Course	GraduateCourse1

Fig.2 Input data D

Sample SPARQL Query:

SELECT *?FacultyType ?Institute ?Student* WHERE {*?x* FacultyType *?Faculty?x* Course *GraduateCourse0*, OPTIONAL{ *?x* Institute *?Organisation* } OPTIONAL { *?x* Student *?GraduateStudent* }}

Step 1: The input query is partitioned into clusters using K-means clustering.
Step 2: Clusters are formed based on the common sub-expression in the queries that are provided as Input.
Step 3: Formed clusters are rewritten into either of Sample 1 and Sample 2 query pattern.
Step 4: The rewritten query is distributed to the input query and the result is Optimized SPARQL Query.

4. ILLUSTRATIONS

A pattern matching query recommended by W3C is SPARQL. There are two types of query variations we focus on:
Sample1: Q := SELECT OP WHERE TP.
Sample2: Q_{OPT} := SELECT OP WHERE TP (OPTIONAL TP_{OPT})$^{+}$
where OP is the Output Result and TP is the set of Triple Pattern. Let D be the data graph, and TP searches the triple pattern in D. The difference between the two queries is the OPTIONAL clause.
SELECT *?FacultyType ?Institute ?Student* WHERE {*?x* FacultyType *?Faculty?x* Course *GraduateCourse0*, OPTIONAL{ *?x* Institute

?Organisation } OPTIONAL { *?x* Student *?GraduateStudent* }}

FacultyType	Institute	Student
FullProfessor	Organisation	
FullProfessor		GraduateStudent
PostDoctrate		GraduateStudent

Fig.3 Output $Q_{OPT}(D)$

Graphically the query graph pattern in the figure 4.3 will consists of four tuples : V- Vertices, E-Edges, Constants and Variables. Vertices represents the subject and object of the triple pattern, gray vertices represent the constants, white vertices represent variables. Predicates are represented in Edges, Dashed edges represent predicates with OPTIONAL graph pattern Q_{OPT} and solid edges represent required graph pattern Q.

The main problem of MultiQuery Optimization is to set query of Sample 1, compute a new set of Q_{OPT} of Sample 1 or Sample 2 queries. There are two requirements for rewritinga queries: It may produce the same result for both Q and Q_{OPT} or the the query evaluation time should be low.

Illustration of Multi Query Optimization is shown in the figures. Figure 4(a) to 4(d) shows the graph patterns of 4 queries and 4(e) shows the graph pattern that rewrites all the 4 queries into a single query. ?xY?p and ?qY?p is the common sub-expression in the figures (a) to (d). These common sub-expression will be rewritten in the figure 4(e) using OPTIONAL clause.

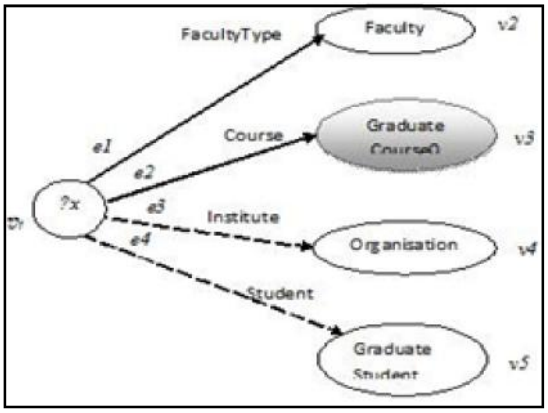


Fig.4 A query graph

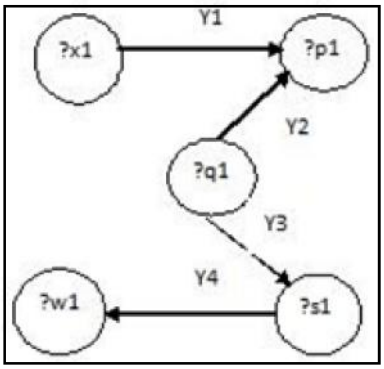


Fig. 4(a) Input query Q1

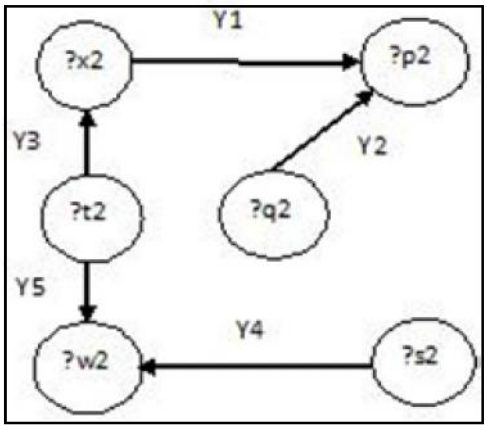


Fig. 4(b) Input Query Q2

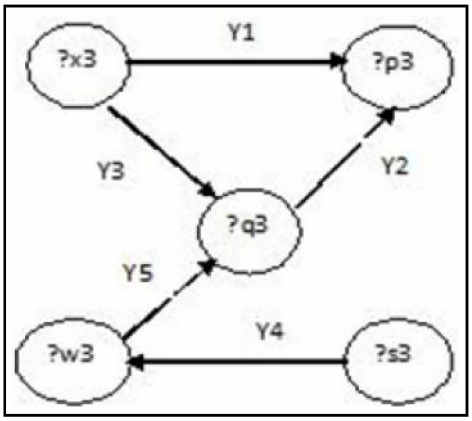


Fig. 4(c) Input Query Q3

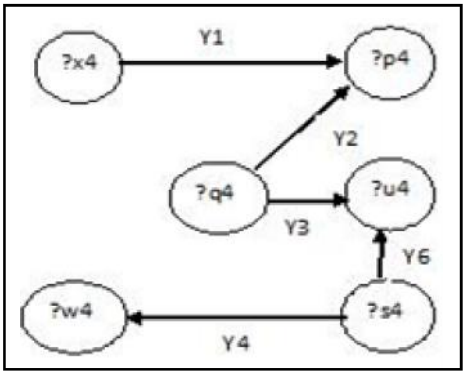


Fig. 4(d) Input Query Q4

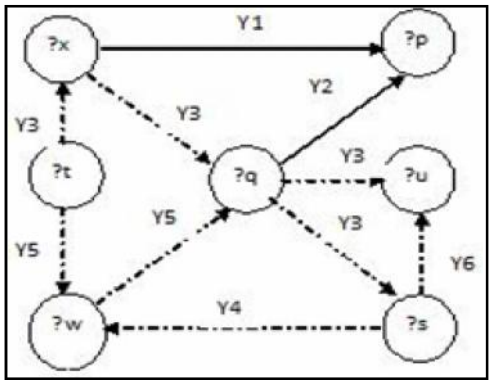


Fig.4(e) Example query for Q_{OPT}

4.1 Dataset

Lehigh University Bench Mark (LUBM) dataset is used for evaluation. This benchmark dataset describes universities with students and departments with limited predicates. LUBM dataset limits the complexity of SPARQL queries.

5. EXPERIMENTAL RESULTS

LUBM data set is used as input which has triple patterns (subject, predicate and object).This RDF data is passed to the preprocessing step where the triple pattern is been split based on the predicate in the PS module. The output from the PS is passed to the POS module where the split is based on the object in the triple pattern. The final result is stored in RDF store.

Multiple SPARQL Queries are given as input. Common sub-expression is clustered using k-means. The clustered query is rewritten using OPTIONAL element and finally optimized query is generated. The output for the Optimized SPARQL query is retrieved from the RDF store. The Execution time of optimized query is measured. red in Milliseconds as shown in the Figure 5 below.

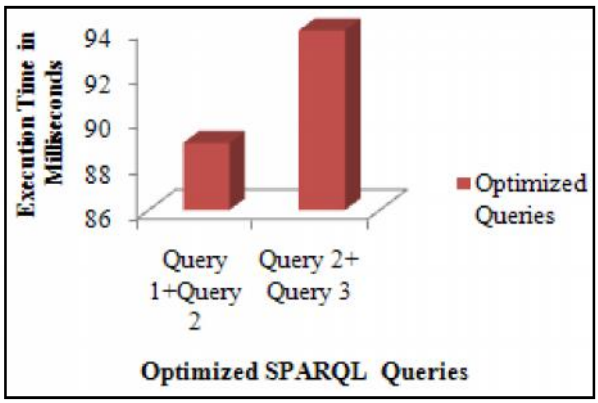


Fig.5 Execution time of multiple SPARQL queries

The Execution time of Single Query is compared with the Optimized Query. The results shows that Optimized SPARQL Query execution is better than Single SPARQL Query execution. The Execution Time for Single SPARQL Query and Optimized SPARQL Query is shown in the Figure 6 below.

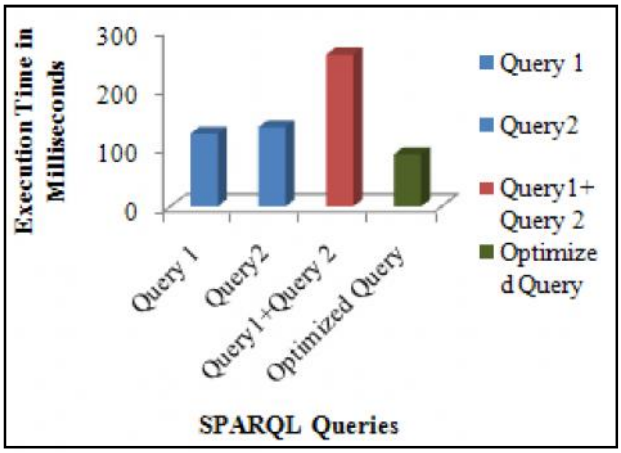


Fig.6 Comparision of execution time of single SPARQL query and optimized SPARQL query

6. CONCLUSION

The problem of MultiQuery Optimization for SPARQL has been studied. Proposed Architecture involves an algorithm to identify the common sub-expression and partitioned the input set of queries into clusters. The clusters are rewritten to evaluate and finally optimized query is obtained. The algorithm provides an efficient, effective and scalable optimization technique. Future work can be extended to general graph databases.

REFERENCES

[1] P. Roy, S. Seshadri, S. Sudarshan and S. Bhobe, “Efficient and Extensible Algorithms for Multi Query Optimization”, In SIGMOD, 2000.

[2] M. Stocker, A. Seaborne and A. Bernstein, “SPARQL Basic Graph Pattern Optimization Using Selectivity Estimation”, In WWW, 2008.

[3] D. J. Abadi, A. Marcus, S. R. Madden and K. Hollenbach, “Scalable Semantic Web Data Management Using Vertical Partitioning”, In VLDB, 2007.

[4] T. Neumann and G. Weikum, “RDF-3X: a RISC-Style Engine for RDF”, In PVLDB, 2008.

[5] M. Atre, V. Chaoji, M. J. Zaki and J. A. Hendler, “Matrix Bit Loaded: A Scalable Lightweight Join Query Processor for RDF Data”, In WWW, 2010.

[6] Kementsietsidis, F. Neven, D.V.de Craen and S. Vansummeren, "Scalable Multi-query Optimization for Exploratory Queries Over Federated Scientific Databases", PVLDB, 2008.

[7] M. Hong, A. J. Demers, J. Gehrke, C. Koch, M. Riedewald and W. M.White, "Massively Multi-query Join Processing in Publish/subscribe Systems", In SIGMOD, 2007.

[8] K. O'Gorman, D. Agrawal and A.E. Abbadi, "Multiple Query Optimization by Cache-aware Middleware Using Query Teamwork", In ICDE, 2002.

[9] K. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review", ACM Comput. Surv., 1999.

[10] M. Schmidt, M. Meier and G. Lausen, " Foundations of SPARQL Query Optimization", In ICDT, 2010.

Low Power Ternary Shift Register Using CNTFETS

V. Sridevi[1] and T. Jayanthy[2]
[1]Sathyabama University, Jeppiaar Nagar, Rajiv Gandhi Road, Chennai - 600 119, Tamil Nadu.
[2]Panimalar Institute of Technology, Varadharajapuram, Poonamallee, Chennai - 600 123, Tamil Nadu.
E-mail: asridevi_2005@yahoo.com

Abstract

In the last few decades, interest in multivalued logic has grown rapidly due to its potential advantages over binary logic for designing energy efficient digital systems. In this paper, a ternary D flip flop with preset and clear inputs is designed using Carbon Nanotube Field Effect Transistor based ternary logic gates. The chiralities of the carbon nanotubes (CNT) used for constructing CNTFET based ternary logic circuits are (19, 0), (13, 0) and (10, 0) of diameters 1.487nm, 0.783nm and 1.018nm with threshold voltages of 0.293V, 0.428V and 0.557V respectively. The designed ternary D flip flop is used as basic building gate for constructing serial in and serial out (SISO) shift registers with improved design and energy efficiency. Finally simulation results using Hspice simulator are reported to show that the proposed CNTFET ternary logic circuits consume significantly less power with considerable reductions in power delay product as compared to conventional binary logic circuits.

Keywords: *Chiralities, CNTFET, D flip flop, Multivalued logic, Power delay product, SISO shift register, Ternary*

1. INTRODUCTION

Multi-valued logic replaces the classical Boolean characterization of variables with either finitely or infinitely many values such as ternary logic [1] or fuzzy logic [2], since it reduces the complexity of interconnects and chip area [3]. By employing ternary logic, serial and serial-parallel arithmetic operations can be carried out faster. In many cases, MVL logic has been combined with binary logic to enhance the performance of CMOS technologies [4]. Three kinds of MVL circuits are current-mode, voltage-mode and mixed-mode or hybrid mode. Several current-mode MVL circuits have been fabricated which shows better performances compared to binary circuits [5]-[8]. But the power consumption of current mode circuits is high due to their inherent nature of constant current flow during the operation. Voltage mode circuits consume a large current only during the logic level switching, thus offers less power consumption. CNTFET replaces conventional devices for low power and high performance design, due to ballistic transport and low off current properties, [9 – 13]. As the threshold voltage of the CNTFET is determined by the diameter of the CNT, a multi-threshold design can be achieved by employing CNTs with different diameters. Recently, efforts have been done for designing combinational circuits using multi-threshold CNTEFs [14].

This paper is organized as follows: Section 2 deals with the fundamental ideas of CNTFET. Simulation results of combinational ternary logic circuits such as ternary D flip flop, ternary SISO shift register etc have been discussed in section 3. Finally in section 4, the research paper has been concluded with the work undertaken in this research work and the scope for improvement of the circuit level transistor models.

2. CARBON NANOTUBE FIELD EFFECT TRANSISTOR

Carbon nanotube field effect transistors consists of semiconducting Carbon nanotubes, which is acting as conducting channel, bridging the source and drain contacts. The diameter of the CNT can be calculated as, $D_{CNT} = \frac{\sqrt{3}a_o}{®}\sqrt{n_1^2 + n_1 n_2 + n_2^2}$ where a_o = 0.144 nm is the inter-atomic distance and the threshold voltage of the intrinsic CNT is given by, $V_{th} = \frac{\sqrt{3}}{3}\frac{aV_®}{eD_{CNT}}$ where a=2.49ï is the carbon to carbon atomic distance, V_π = 3.033eV is the carbon π-π bond energy in the tight binding model and e is the unit electron charge. Thus, the threshold voltage of the CNT is inversely proportional to the diameter of the CNT in turn the chiral vector. In this

paper, the chiralities of the CNTs used for modeling of CNTFETs are (19,0), (13,0) and (10,0) with diameters 1.487 nm, 0.783 nm and 1.018 nm respectively.

3. CIRCUIT LEVEL IMPLEMENTATION AND ASSESSMENT OF TERNARY LOGIC

A compact SPICE model including non-idealities [20 - 22] is used for simulations which has been designed for unipolar, MOSFET-like CNTFET based circuits, also considers Schottky Barrier Effects, Parasitics, including CNT, Source/Drain, and Gate resistances and capacitances, and CNT Charge Screening Effects. HSPICE simulator has been used to simulate the proposed ternary-logic based combinational circuits. The threshold voltages of CNTFETs used in the circuits are shown in Table 1.

Table 1 Threshold Voltages of CNTs

Chirality (n_1,n_2)	Diameter (nm)	Threshold Voltage (V)
(19,0)	1.487	0.293
(10,0)	0.783	0.557
(13,0)	1.018	0.428

3.1 Ternary NAND

The basic function of ternary NAND gate is defined by $Y_{NAND} = \overline{Min\{X_1, X_2\}}$ and the truth table is shown in Table 2.

The circuit realization of ternary NAND gates requires 5 n-CNTFETs and 5 p-CNTFETs as shown in Figure 1. The chirality of the CNT used for transistors T_1, T_2, T_5, T_6 is (19, 0), for T_3, T_4 is (13, 0) and for T_7, T_8, T_9, T_{10} is (10, 0). The diameter of CNT used for T_1, T_2, T_5, T_6 is 1.487, for T_3, T_4 is 1.018 and for T_7, T_8, T_9, T_{10} is 0.783. The threshold voltage for T_1, T_2, T_5, T_6 is 0.293V, for T_3, T_4 is 0.428V and for T_7, T_8, T_9, T_{10} is 0.557V. From the transient response of ternary NAND shown in Figure 2, the proper functioning of the circuit for various input variables can be observed.

Table 2 Truth Table of Ternary NAND Gate

X_1	X_2	Y_{NAND}
0	0	2
0	1	2
0	2	2
1	0	2
1	1	1
1	2	1
2	0	2
2	1	1
2	2	0

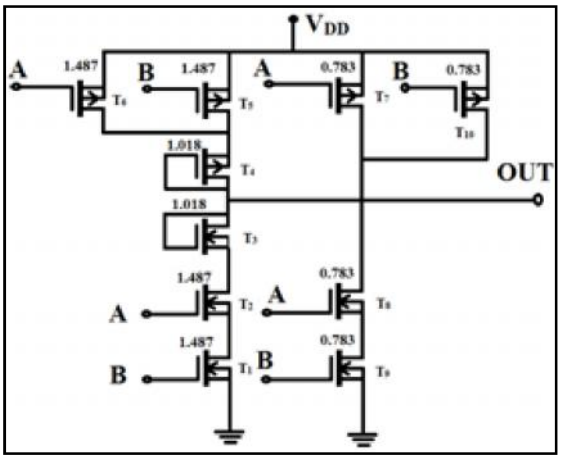


Fig.1 Structure of Ternary NAND gate

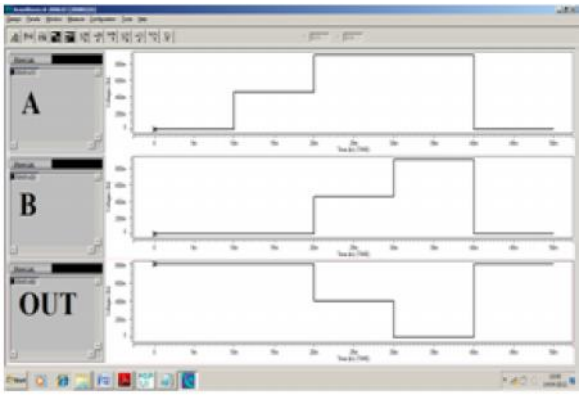


Fig.2 Behaviour of Ternary NAND gate

3.2 Ternary D Flip Flop with Binary Clock

A similar structure of flip flops used in binary is applied for constructing ternary D flip flop by replacing all the binary logic gates with ternary logic gates. The structure of ternary D flip flop is shown in Figure 3 and truth table in Table 3. The designed D flip flops can be used in many applications such as Shift registers, frequency dividers, counters etc. In this paper, it is used to design shift registers.

Table 3 Truth Table of Ternary D Flip Flop with Binary Clock

CLK	Data	Q	Q_{prev}
2	0	0	X
2	1	1	X
2	2	2	X
0	X	Q_{prev}	-

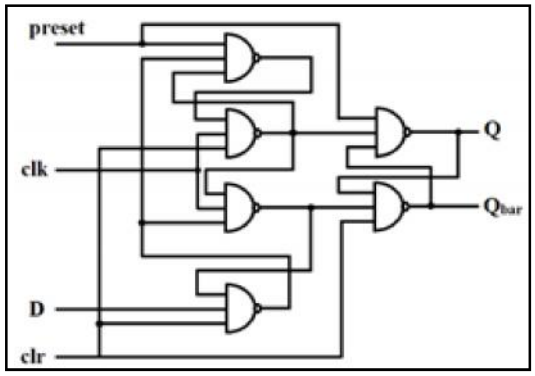


Fig.3 Structure of ternary D flip flop

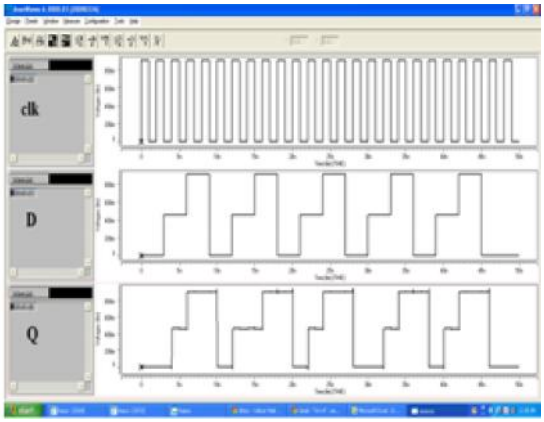


Fig.4 Behavior of ternary D flip flop with binary clock

The data supplied to the D flip flop is ternary input data whereas the clock input is binary. When the clock input is high, the D flip flop read the input value and transmits the input value to the output. The designed ternary flip flop is able to transmit all three logic levels i.e. 0, 1 and 2. When the clock input is low, the D flip flop retains its previous state. The performance of the flip-flop is verified using Hspice simulator. Figure 4 shows the behaviour of flip flop.

3.3 Ternary D flip flop with ternary clock

For the clock input of logic 0 and logic 2, the output behaviour of the ternary D flip flop with ternary clock input is similar. For the intermediate value (logic 1), the value of Q output depends on data input as shown in Table 4 and Figure 5.

Table 4 Truth table of Ternary D Flip Flop with Ternary Clock

CLK	Data	Q	Q_{prev}
2	0	0	X
2	1	1	X
2	2	2	X
1	0	0	0
		1	1,2
1	1	1	X
1	2	1	0,1
		2	2
0	X	Q_{prev}	-

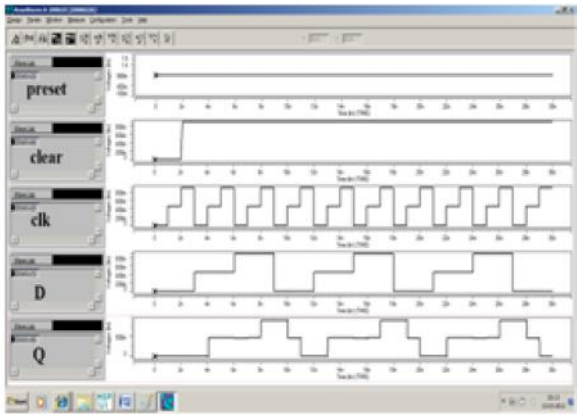


Fig.5 Behavior of ternary D flip flop with ternary clock

3.4 Ternary SISO Shift Register

In digital circuits, a shift register is a cascade of flip flops, sharing the same clock, which has the output of anyone but the last flip flop connected to the “data” input of the next one in the chain, resulting in a circuit that shifts by one position the one dimensional “bit array” stored in it, shifting in the data present at its input and shifting out the last bit in the array, when enabled to do so by a transition of the clock input.

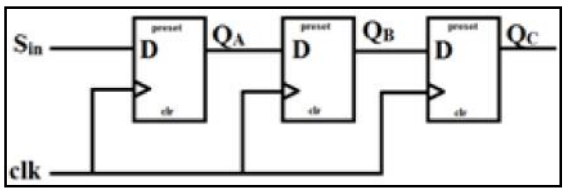


Fig.6 Structure of ternary SISO shift register

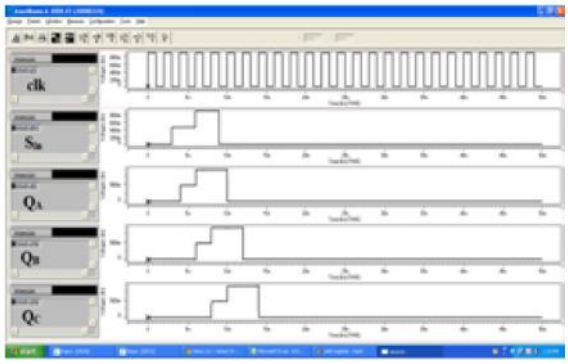


Fig.7 Behavior of SISO shift register

Figure 6 shows the basic three bit ternary serial-in serial-out shift register implemented using ternary D flip flops. The circuit functions as follows. A reset applied to the 'clr' input of all the flip flops resets their Q outputs to 0s. The waveforms shown in Figure 7 include the clock pulse train, the waveform representing the data to be loaded onto the shift register and the Q outputs of different flip flops. The flip flops shown respond to the LOW to HIGH transition of the clock pulses as indicated by their logic symbols.

During the first clock transition, the Q_A output goes from logic '0' to logic '1'. The outputs of the other two flip flops remain in the logic '0' state as their D inputs was in the logic '0' state at the time of clock transition. During the second clock transition, the Q_A output goes from logic '1' to logic '2' and the Q_B output goes from logic '0' to logic '1', again in accordance with the logic status of the D inputs at the time of relevant clock transition. During the third clock transition, the Q_A output goes from logic '2' to logic'0 ', the Q_B output goes from logic '1' to logic '2' and the Q_C output goes from logic '0' to logic '1'. Thus, we have seen that a logic '2' that was present at the data input prior to the occurrence of the second clock transition has reached the Q_B output at the end of fourth clock transitions. This bit will reach the Q_C output at the end of sixth clock transitions. In general, in a three bit ternary shift register of the type shown in Figure 6, a data bit present at the data input terminal at the time of the nth clock transition reaches the Q_C output at the end of the $(n+6)^{th}$ clock transition.

3.5 Comparison of Binary-logic and Ternary-logic

Table 5 shows the comparison results of binary and ternary logic circuits. The average delay of the binary D flip flop is $4.9854 * 10^{-11}$ S and average power consumed is $6.172 * 10^{-6}$ W. Therefore the power delay product is $3.0769 * 10^{-16}$ J. The average delay of the binary SISO shift register is $5.134 * 10^{-11}$ S and average power consumed is $7.045 * 10^{-6}$ W. Therefore the power delay product is $3.6169 * 10^{-16}$ J. The average delay of the ternary D flip flop is $3.0378 * 10^{-11}$ S and average power consumed is $4.8492 * 10^{-6}$ W. Therefore the power delay product is $1.473 * 10^{-16}$ J. The average delay of the ternary SISO shift register is $3.9773 * 10^{-11}$ S and average power consumed is $5.2869 * 10^{-6}$ W. Therefore the power delay product is $2.1027 * 10^{-16}$ J. Figure 8(a-c) and Figure 9(a-c) shows that the proposed ternary logic design is faster and consumes less power compared to binary logic family.

Table 5 Comparison of Binary and Ternary Logic Circuits

Circuit	Binary Logic			Ternary Logic		
	Delay (10^{-11}s)	Avg Power (10^{-6} W)	PDP (10^{-16} J)	Delay (10^{-11}s)	Avg Power (10^{-6} W)	PDP (10^{-16} J)
D Flip Flop	4.985	6.172	3.076	3.037	4.849	1.473
SISO Shift Register	5.134	7.045	3.616	3.977	5.286	2.102

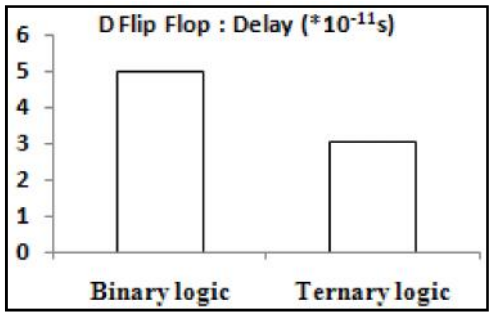


Fig. 8(a) Comparison of binary and ternary D flip flop-delay

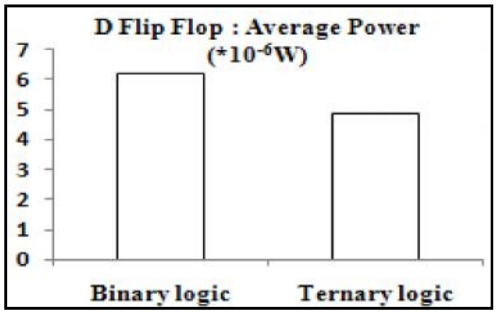


Fig. 8(b) Comparison of binary and ternary D flip flop-average power

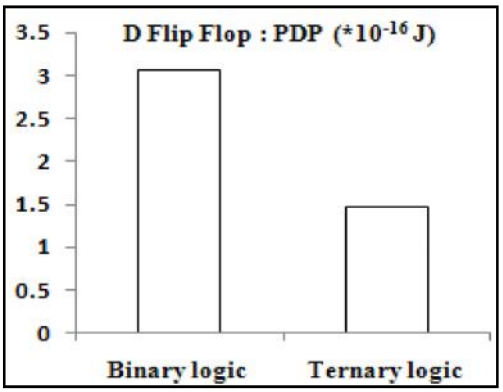


Fig. 8(c) Comparison of binary and ternary D flip flop-PDP

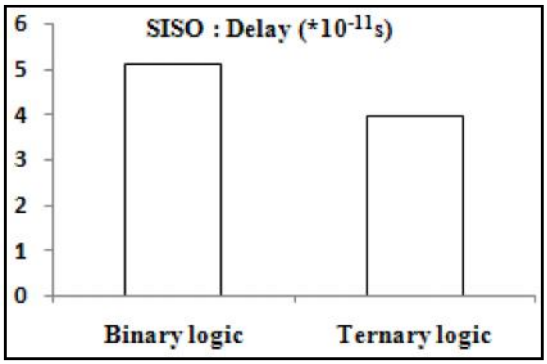


Fig. 9(a) Comparison of binary and ternary shift register-delay

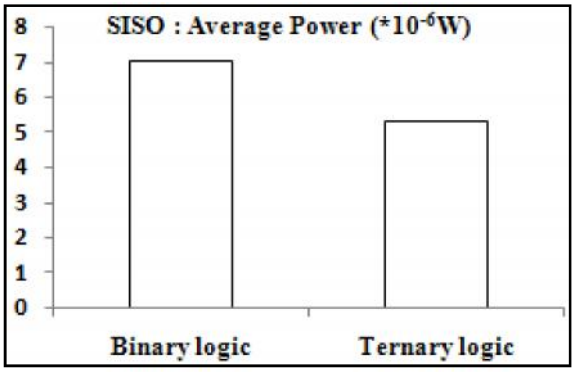


Fig. 9(b) Comparison of binary and ternary shift register – average power

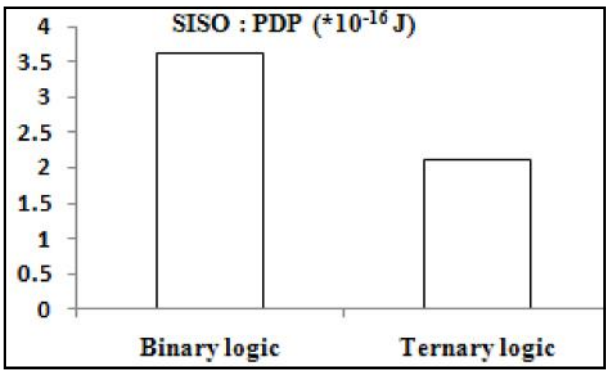


Fig. 9(c) Comparison of binary and ternary shift register - PDP

4. CONCLUSION

The prospects of applying ternary logic in computation have been discussed in this paper. Multi-threshold CNTFETs have been used for realizing ternary circuits such as clocked D flip flop, Shift registers etc. All simulations have been performed in HSPICE and the simulated results validated the correct operation of the realized circuits. Comparison made between binary and ternary logic showed that the circuits designed using ternary logic is predicted to be faster than classical binary circuits and works at even lower power.

REFERENCES

[1] M. Mukaidono, "Regular Ternary Logic Functions - Ternary Logic Functions Suitable for Treating Ambiguity", IEEE Trans. Computers, Vol. C-35, No. 2, 1986, pp.179- 183.

[2] T. Araki, H. Tatsumi, M. Mukaidono, F. Yamamoto, "Minimization of Incompletely Specified Regular Ternary Logic Functions and its Application to Fuzzy Switching Functions", in Proc. IEEE International Symposium on Multiple-Valued Logic, 1998, pp.289-296.

[3] P.C. Balla, A. Antoniou, "Low Power Dissipation MOS Ternary Logic Family", IEEE J. Solid-State Circuits, Vol.19, No.5, 1984, pp.739-749.

[4] D. A. Rich, "A Survey of Multivalued Memories", IEEE Trans. Computers, Vol. 35, No. 2, 1986, pp.99-106.

[5] Hanyu, M. Kameyama, "A 200 MHz Pipelined Multiplier Using 1.5V-supply Multiple Valued MOS Current-mode Circuits with Dual-rail Source-coupled Logic", IEEE Journal of Solid-State Circuits, Vol.30, No.11, 1995, pp.1239-1245.

[6] B. Radanovic, M. Syrzycki, "Current-mode CMOS Adders Using Multiple-Valued Logic", Canadian Conference on Electrical and Computer Engineering, pp.190-193, 1996.

[7] J. Shen et al., "Neuron-MOS Current Mirror Circuit and Its Application to Multi-Valued Logic", IEICE Trans. Inf. & Syst., E82-D,5 1999, pp.940-948.

[8] D. H. Y. Teng, R. J. Bolton, "A Self-restored Current-mode CMOS Multiple-valued Logic Design Architecture", 1999 IEEE pacific Rim Conf. on Communications, Computers and Signal Processing (PASRIM'99), 1999, pp. 436-439.

[9] J. Appenzeller, "Carbon Nanotubes for High-Performance Electronics-Progress and Prospect," Proc. IEEE, Vol. 96, No. 2, 2008, pp.201-211.

[10] A. Rahman, J. Guo, S. Datta, M.S. Lundstrom, "Theory of Ballistic Nanotransistors," IEEE Trans. Electron Device, Vol. 50, No.10, 2003, pp.1853-1864.

[11] A. Akturk, G. Pennington, N. Goldsman and A. Wickenden, "Electron Transport and Velocity Oscillations in a Carbon Nanotube", IEEE Trans. Nanotechnol, Vol.6, No. 4, 2007, pp.469-474.

[12] H. Hashempour, F. Lombardi, "Device Model for Ballistic CNFETs Using the First Conducting Band," IEEE Des. Test. Comput., Vol. 25, No. 2, 2008, pp.178-186.

[13] Y. Lin, J. Appenzeller, J. Knoch, P. Avouris, "High-performance Carbon Nanotube Field-Effect Transistor with Tunable Polarities," IEEE Trans. Nanotechnol, Vol.4, No.5, 2005, pp. 481-489.

[14] Peiman Keshavarzian and Keivan Navi, "Universal Ternary Logic Circuit Design Through Carbon Nanotube Technology", International Journal of Nanotechnology, Vol. 6, No. 10, 2009, pp.942-953.

A Novel Approach for Online Identity Management System Using AADHAAR Unique Identification Number

T. Sivakumar[1], A. Ummu Salma[2] and T. Anusha[3]
[1&2]Department of Information Technology, [3]Department of Computer Science and Engineering
PSG College of Technology, Coimbatore-641 004, Tamil Nadu
E-mail: sk@ity.psgtech.ac.in, abd_salma@yahoo.com, anu@cse.psgtech.ac.in

Abstract

Evolution of Unique National identity number plays a significant role across nation-wide. It provides the government with accurate data on residents, enable direct benefit programs, and allow government departments to coordinate investments and share information.Unique identification project is as an initiative that would provide identification for each resident across the country and would be used primarily as the basis for efficient delivery of welfare services. It would also act as a tool for effective monitoring of various programs and schemes of the Government. The role is to issue a Unique Identification number (UID) that could be verified and authenticated online, in a cost-effective manner, which is robust enough to eliminate duplicate and fake identities. This requires the individual's data to be collected, transmitted and stored in a central repository. Thus, the issues involved in this system could be security while transmission and storage, fast retrieval of data, storage of huge amount of data and biometric data to be enrolled. The security mechanisms that can be provided are confidentiality, integrity and authentication. Hence, this paper proposes architecture for providing solutions to the technological challenges, such as security and speed. Instead of using the traditional client-server architecture, a proxy-server mechanism is used for fast retrieval of data and the security mechanism concentrated here is confidentiality.

Keywords*: Authentication, Data security, Speed efficiencyUnique Identification Number*

1. INTRODUCTION

Aadhaar is a 12-digit unique number which the Unique Identification Authority of India (UIDAI) is issuing for all residents of India. This project is as an initiative that would provide identification for each resident across the country and would be used primarily as the basis for efficient delivery of welfare services. It plays a significant role across nation-wide. The role is to issue a Unique Identification number (UID) that can be verified and authenticated in an online, cost-effective manner, which is robust enough to eliminate duplicate and fake identities.

The number will be stored in a centralized database known as CIDR (Central ID Repository) and linked to the basic demographics and biometric information – photograph, ten fingerprints and iris – of each individual. Once a person is on the database, the person will be able to establish his identity in online easily. It will become the single source of identity verification online. Residents would be spared the hassle of repeatedly providing supporting identity documents each time they wish to access services such as obtaining a bank account, passport, driving license and so on. By providing a clear proof of identity, Aadhaar will also facilitate entry for poor and underprivileged residents into the formal banking system and the opportunity to avail services provided by the government and the private sector. It also gives migrants mobility of identity.

The benefits of Aadhaar card are:

- Great potential for not-so-privileged, poor and the marginalized people, mostly living in the rural areas.
- Clear proof of identity.
- Facilitate entry for poor and underprivileged residents into the formal banking system.
- Opportunity to avail services provided by the government and the private sector.
- Giving migrants mobility of identity.
- Financial inclusion with deeper penetration of banks, insurance and easy distribution of benefits of government schemes.

Each Aadhaar number will be unique to an individual and will remain valid for life. It is a random number generated, devoid of any classification based on caste, creed, religion and geography [9]. A sample UID card is shown in Figure 1.

Fig.1 Sample AADHAAR card

2.TECHNOLOGICAL CHALLENGES OF UIDAI

The biometric data of the individuals used are fingerprint, iris and face. Since all three data are used and stored for every individual, the size of the database (Total population of India x size of biometric data combined) becomes large. Also, the resident's information is stored in the database. Hence it must be stored in a secure manner so that any unauthorized or illegal person cannot view the data. The data must also be transmitted across the Internet in a safe manner without any intervention and modification to provide online identity and authentication services.

The major technical challenges faced by this project are [6].

2.1 Volume

Creating and managing a database of 1.2 billion people spread over a huge area will involve immense work. Around five megabytes of data will be required to store the compressed fingerprint images (all ten fingers) of each individual, meaning the size of the entire database will be atleast six petabytes (6,000 terabytes, or 6,000,000 gigabytes), making it among the world's largest databases.

Table 1Storage Requirements of Biometric Data

Type of Information	Storage Size Per Subject	Storage Size for Entire Population
Fingerprint Image	7.5 MB	8000-10000 TB
Face Image	4 KB	5 TB
Iris Image	150 KB	170-200 TB

2.2 Speed

Each new entry has to be validated against existing entries to remove the possibility of duplication. Over next few years, this would mean comparing each new application against, say, one billion entries in the database at a reasonable speed. Also, UIDAI proposes online authentication through cell-phones and using basic technology. While authentication is a simpler process, the proposed time of three to four seconds for the same makes it challenging.

2.3 Security

It is important to maintain privacy of the individual's data, avoid identity theft and provide a good authentication mechanism.

2.3.1 Privacy

As people become more connected online and pieces of their identities are more readily accessible, the potential to steal the data and access it illegally grows exponentially. Dealing with sensitive information, it has to include security features that will ward off the hackers.

2.3.2 Identity Theft

The lack of a single, user-friendly online identity access platform increases the risk of identity theft and identity fraud, and makes it difficult for people to find out they've been hit until it's too late. Hence, this must be avoided [1].

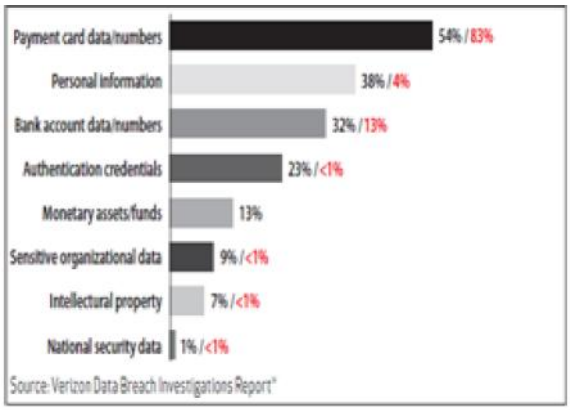


Fig.2 Compromised data types by percent of breaches and percent of records [1]

2.3.3 Authentication

As effective as these security methods can be, significant challenges still remain, leaving organizations

stuck between a rock and a hard place. Stronger security increases the cost exponentially and is not always worth the tradeoff. Other forms of security are difficult to use and expensive to support, preventing their mass adoption. Probably the single biggest challenge to organizations, both public and private, is the high cost to implement and manage authentication [1].

Fig.3 NIST security levels 1-4 [1]

2.4 Biometrics

Along with fingerprints, iris scan and face of every individual will also be collected and stored. This in turn increases the size of the database.

In this paper, a novel methodology is proposed to provide suitable solution for the challenges like speed and security for providing and efficient and secure online identity service proceedings book. Similarly, for journal references, include name of the journal, volume no, issue no, year of publication and page numbers. For book publications, along with authors name and book title, include year of publication, publisher name and place of publication. The template (and examples) for typing references is given in [5].

3. SPEED AS A CHALLENGE AND ITS SOLUTION

The client/servermodel of computing is a distributed application structure that partitions tasks or workloads between the provider of a resource or service, called server, and service requesters, called clients. Often clients and server communicate over a computer network on separate hardware, but both client and server may reside in the same system.A server machine is a host that is running one or more server programs which share their resources with clients. A client does not share any of its resources, but requests a server's content or service function. Clients therefore initiate communication sessions with the servers which await incoming requests. Figure 4 shows the basic architecture of client-server architecture.

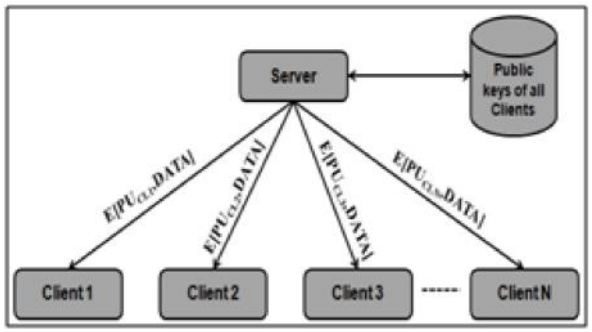


Fig.4 Traditional client-server architecture

There are several drawbacks to this architecture, such as,

i. One server and many clients' leads to traffic overload.
ii. Slow response.
iii. Packets get dropped.
iv. Buffer overflow.
v. Server should maintain a symmetric key for all clients.
vi. Single node failure.
vii. The server has to maintain the keys of all the clients.
viii. It also has to perform parallel encryption when multiple clients send request.

An alternative approach to overcome these problems is to use client-proxy-server architecture. This proposed architecture has a main server and many proxy servers. These proxy servers in turn have their appropriate sub proxies or clients. Figure 5 shows the suggested client-proxy-server architecture for the proposed Online Identity Management System (OIMS) using AADHAAR ID.

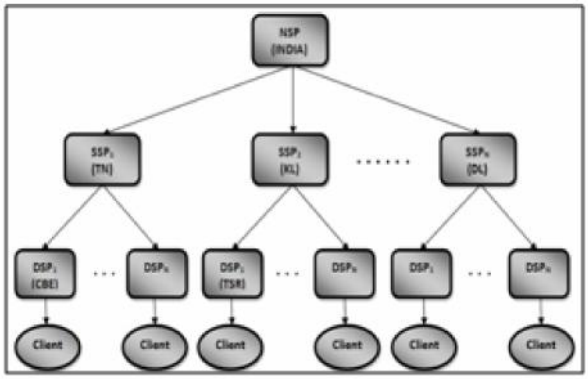


Fig.5 Client-proxy-server architecture

The main server is called NSP (National Service Provider), and its immediate proxy servers are called SSP (State Service Provider) and the sub proxy servers are said to be DSP (District Service Provider).

From Figure 5, the clients can be any government/ private organizations such as Passport office, RTO, Banks, Employment office, etc. The number of SSPs equals number of states in India and number of DSPs equal to number of districts in the corresponding state.

As this project is dealing with a large amount of data, the volume is considered to be one of the main issues. The architecture proposed maintains database at proxy level (district, state, national) thus handling the large amount of data in an efficient manner.The speed of transmitting the data over the network will obviously be higher than compared to client-server model.There occurs no duplication of identity since the fingerprint of a person is confidentially unique and the UID is perfectly unique.

A proxy server has a variety of potential purposes, including:

i. To keep machines behind it anonymous, mainly for security.
ii. To speed up access to resources (using caching).
iii. To prevent downloading the same content multiple times and save bandwidth.
iv. To log / audit usage.
v. To scan transmitted content for malware before delivery.
vi. To scan outbound content, e.g., for data loss prevention.
vii. Access enhancement/restriction.

3.1 Typical Working Model of the Proposed Architecture

When the client requests for the data, it is first checked in the proxy. If the proxy has the data corresponding to the UID sent by the client, then the proxy sends the details to the client. If the data is not present in the proxy, then the request is forwarded to the server. Based on the UID, the copy of the matching data is sent back to the proxy. Finally, the proxy responds to the client with the requested information. The proxy keeps a copy of the data (Figure 6).

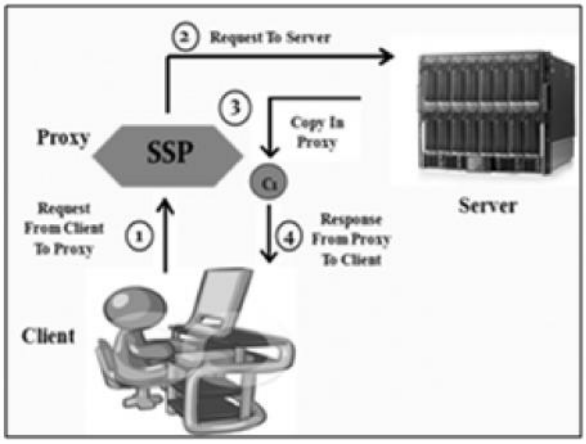


Fig.6 Working model of client-proxy-server architecture

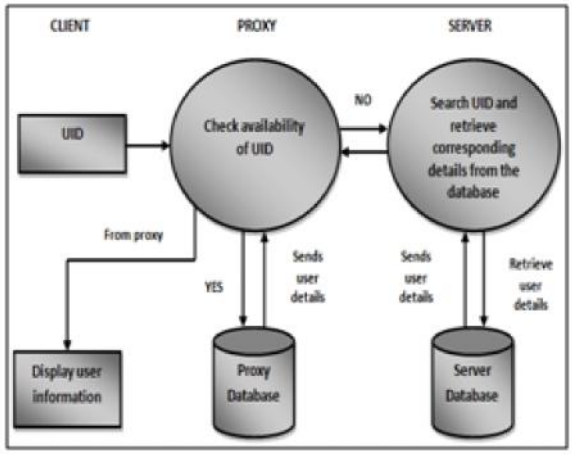


Fig. 7 Dataflow diagram for retrieving data

The government is collecting all data from every individual. Initially, all the details are stored in the NSP (CIDR). Similarly, when a data is been requested, it moves to the SSP and then DSP depending on from where the request has been sent. With this architecture, the details of the residents of that area/district within the district database itself is maintained, i.e., If a client moves to another district/state and wishes to view his information, then request will be sent to that district/state proxy he/she currently belongs using the unique id and the details will be made available in the local proxy. Finally, the details will be displayed. Figure 7 shows the data flow diagram for retrieving data and Figure 8 shows the UML diagram for the same. The three main issues namely volume of the data to handle, speed of transmitting the data over the network and the security to provide has been met by this architecture in a better way.

3.2 Major Modules

i. **Clients-** They represent the residents or individuals who enroll their data and get their Unique ID. They can also be any government/private organizations like Banks, Passport office, RTO, etc.
ii. **DSP (District Level Proxies)-** District Service Provider; the number of DSPs is equal to the number of districts in each state.
iii. **SSP (State Level Proxies)-** State Service Provider; the number of SSPs is equal to the number of states.
iv. **NSP/Server-**Main Database / CIDR, where the entire individual's information is stored.

Fig.8 UML diagram for retrieving data

3.3 Major Security Issues

i. Data confidentiality during transmission – Encryption
ii. Data origin authentication – Digital Signature
iii. Data integrity service – Digital Signature
iv. User Authentication
v. Cryptographic key generation
vi. Key distribution and management

In this paper, we propose a simple and cost-effective approach to provide security of the user identities while transmission.

4. SOLUTION FOR SECURITY ISSUES

This chapter discuss about the fundamentals of security services and mechanisms. The concepts of Multi-key RSA and algorithm are also discussed. Table 2 shows the various security services, mechanisms and algorithms which provide the concern services.

Table 2 Security Services and Mechanisms

Security Services	Protection Mechanisms	Algorithm
Authentication	Login and password	MD5 , SHA
Confidentiality	Data Encryption Symmetric(Secret key) Asymmetric(Public key)	DES, AES RSA
Integrity	Message Digest(Hashing)	Hashing

4.1 Security using Multi-key RSA

In the above mentioned architecture (Figure 5), it is clearly seen that data or unique id is being transmitted when data is not present in the corresponding proxy database. Hence security services, such as confidentiality, integrity, authentication and digital signature is needed for efficient and secure transmission of the data over the public network.

The primary advantage of public-key cryptography is increased security and convenience: private keys never need to be transmitted or revealed to anyone. Another major advantage of public-key systems is that they can provide digital signatures that cannot be repudiated. Authentication via secret-key systems requires the sharing of some secret and sometimes requires trust of a third party as well. As a result, a sender can repudiate a previously authenticated message by claiming the shared secret was somehow compromised by one of the parties sharing the secret.

4.1.1. Milti-key RSA Algorithm

Step 1: Generate two large prime numbers (p and q).
Step 2: Generate a sequence of encryption keys $\{e_i\}_{0d"id"N}$such that,$1<e_i<Ø$ andgcd(e_i ,Ø) = 1, for 0 d" i d" n
Step 3: Generate a corresponding sequence of decryption keys $\{d_i\}$ such that,$1< d_i<Ø$ and $(e_0 x\ e_i)\ x\ d_i =1$ (modØ)
Step 4: The server will send 'n' (pxq) and the encryption keys 'e_i' over a secure channel to the proxy servers.
Step 5: It also encrypt the original data as follows: $D_o = (D_{-1})^{\wedge}e_o \bmod n$
Step 6: D_o can be sent over an insecure channel to the proxy.

4.1.2. Extended Euclidean Algorithm

// Tofind the decryption key d_i
Step 1: If gcd(m, b) = 1, then b has a multiplicative inverse modulo m
Step 2: (A1, A2, A3) ! (1, 0, m); (B1, B2, B3) ! (0, 1, b)
Step 3: if B3 = 0, then return A3 = gcd(m, b); no inverse
Step 4: if B3 = 1, return B3 = gcd(m, b);B2 = b^{-1} mod m
Step 5: Q = |a3/b3|
Step 6: (T1, T2, T3) ! (A1-QB1, A2-QB2, A3-QB3)
Step 7: (A1, A2, A3) ! (B1, B2, B3)
Step 8: (B1, B2, B3) ! (T1, T2, T3)
Step 9: goto step 2

4.2 Data Confidentiality using Multi-key RSA

Figure 9 shows the mechanism for providing data confidentiality during transmission.

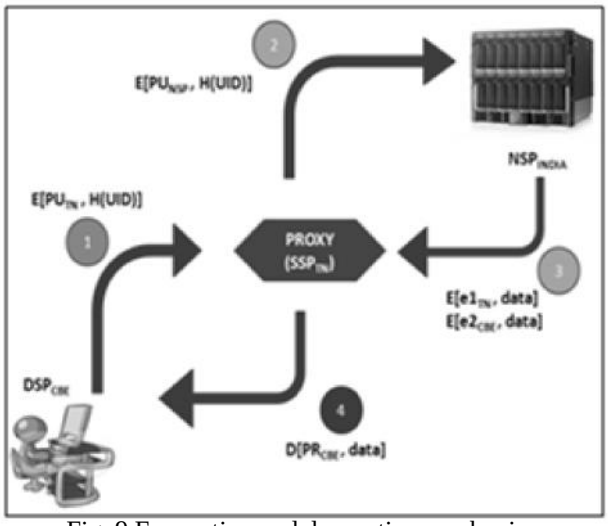


Fig. 9 Encryption and decryption mechanism

1. E[PU_{TN}, H(UID)]

The advantage of using H(UID) is that this one way version of UID is compared at proxy. It is safe against proxy intruders.

Step 1: The request is sent from the client (DSP_{CBE}) to the proxy (SSP_{TN}) for retrieving the details of the clients sending the UID. Here, while sending the request from the client, the H(UID) is encrypted with the public key of SSP_{TN} and sent to TN (SSP), (i.e.) E[PU_{TN}, H(UID)].

(DSP_{CBE}aDistrict Service Provider of Coimbatore, SSP_{TN}aState Service Provider of TamilNadu)

After the encrypted message reaches the proxy (SSP_{TN}), decryption with the private key of TN (SSP) is done, (i.e.) D[PR_{TN}, E[PU_{TN}, H(UID)]].

Searching of the details with the corresponding UID in the database of the proxy will be carried out. If the data is present, the corresponding details for that UID will be sent to the client. If the data is not found in the TN proxy (SSP_{TN}) go to Step 2.

2. E[PU_{NSP}, H(UID)]
D[PR_{NSP}, E[PU_{NSP}, H(UID)]]

Step 2: The request is forwarded to the National Service Provider (NSP). The CIDR has all the details maintained in its database.

Encryption with the public key of India (NSP) with the H(UID) is done and the request is forwarded to the main server. The server (NSP) receives the request and decrypts the message with its private key. Searching is again carried out in the main database (CIDR) for retrieving details of the corresponding UID.

3. E[e_{1TN}, data]
E[e_{2CBE}, Re-encryption]

Step 3: In the response phase of the server (NSP), encryption is done twice and the decryption is done once. The security of the authenticated data of the requested client comes into picture.

So the data is encrypted with the public key (e1) and forwarded to SSP_{TN}. A copy of data is kept here. Then the data is re-encrypted with e2 at $SSP_{TN}$and the response is forwarded to the DSP_{CBE}.In client side step 4 takes place.

4. D[PR_{CBE},E[e_{2CBE}, E[e_{1TN}, data]]]

Step 4: In DSP the data is decrypted using decryption key (PR_{CBE}) to get back the client requested details.

Therefore, using Multi-key RSA, the data is being encrypted many times using different public keys, whereas, decryption is done only once using the private key to get the desired response.

The public keys of all the districts (DSPs) will be kept in their respective SSPs. Similarly, all state's (SSPs) public key will be maintained in the NSP. The private key of DSP is kept secret with the concerned DSP, so

that only that DSP can decrypt the data. Similarly SSP and NSP also keep their private keys with themselves.

One important feature of the proposed scheme is that clients only need to perform a single decryption operation to recover the original data even though the data packets may have been encrypted by multiple proxies along the delivery path.

5. CONCLUSION AND FUTURE WORK

The main goal of this paper is to propose architecture for providing solutions to the technological challenges, such as security and speed for the Online Identity Management System. Instead of using the traditional client-server architecture, client-server proxy mechanism is used for fast retrieval of data. To provide data confidentiality during transmission, Multi-key RSA algorithm is used. By the emerging internet trends this architecture could be definitely implemented in a great level and thus it brings out a change in the society which is sophisticated in technical aspects.Therefore among the major technological challenges like speed, security, volume and biometrics, the implementation finds a solution for improving speed and confidentiality (security) for the data. Further, the NSP could be used to connect with other countries to provide online identity service on an international level. By implementing this system, the process of applying and getting the UID through online can be made automated. The only manual work involved is at the client side to help the users during enrollment process.

The future work is concentrating on key generation and key exchange issues. The overhead of RSA algorithm will be discussed and to identify a suitable and simple encryption algorithms for multimedia data. To provide other security services like authentication, integrity and digital signatures using suitable algorithms.

REFERENCES

[1] "Online Identity Management needs a Universal Answer", White Paper, Verizon, 2012.

[2] "Technological Challenges Before the UIDAI", The Hindu (October, 2009), Information Technology, pp.12.

[3] Gordon W. Romney and Donald W. Parry, "A Digital Signature Signing Engine to Protect the Integrity of Digital Assets", IEEE, 2006.

[4] Narendra Kumar Menta and T.Sivakumar, "Component Based Architecture for Online Identity Management System", International Conference on Innovations in Engineering and Technology for Sustainable Development (IETSD-2012), Bannari Amman Institute of Technology, Vol.3, 3-5 September 2012, pp.28-32.

[5] O.SamiSaydjari and Vijay Varadharajan, "The Evolution of Online Identity", Co-published by the IEEE Computer and Reliability Societies, October 2009.

[6] Robert Kofler, Robert Krimmer, Alexander Prosser and Martin-Karl Unger, "The Role of Digital Signature Cards in Electronic Voting", Proceedings of the IEEE 37th Hawaii International Conference, 2004.

[7] William Stallings, "Cryptography and Network Security Principles and Practices", Prentice Hall, 2011.

Domain Classifier Using Conceptual Granulation and Equal Partition Approach

D. Malathi[1] and S. Valarmathy[2]
[1]Department of Computer Application, [2] Department of Elctronics and Communication Engineering
Bannari Amman Institute of Technology, Sathyamangalam - 638 401, Erode District, Tamil Nadu
E-mail:malathisubbu@gmail.com, artmathy@gmail.com

Abstract

This paper presents a systematic approach for the classification of large corpus based on concept granulation and equal partition approach. The proposed work has three main processes which are, the preprocessing treatments to text documents, feature extraction and finally the classification. The proposed approach is concentrated in the feature extraction phase. Almost bird eye view like approach is the feature extraction method. So the proposed work concept granulation and equal partition approach has been named as IMmune Term (TIM), which finds the immunized terms from the information system. At first, documents are preprocessed from text to numerical form i.e., word frequency is calculated for each document. Second, sets of features are extracted using TIM. In the third step, the TIM treated feature is introduced to Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI) for global set extraction or dimension reduction. Finally, Naove Bayes (NB) and Support Vector Machine (SVM) are used to classify the documents. The proposed work seems to be fruitful when compared to the conventional word frequency approach.

Keywords: *Concept granulation, Domain classifier, Equal partition*

1. INTRODUCTION

Information retrieval for knowledge extraction is the food served from World Wide Web today. The research communities are setting mile stones with new approaches for fruitfulness of truth accessed from internet. The work we present here is a feature extraction by spreading word granules and combining by micro averaging between local spreads. By this method the concrete terminological words for topics are highlighted. Further this approach has been realized to be better performer when compared to the conventional probabilistic and statistical approaches.

1.1 Text Mining

Text mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources [1]. Text learning technique is nothing but extraction of information from words. A role of text mining is linking together the extracted information to form new hypotheses which has to be explored by the existing means of algorithms and applications.

1.2 Preprocessing

The initial preprocessing step in Text Mining is Tokenization process, i.e. a text document is split into a Bag of words by removing all punctuation and non-text characters. This tokenized representation is further reduced into set of words by the removal of stop words like the, are, is, of, and so on. These stop words are removed because they do not support our proposed method of inferring knowledge. Further the size of the words is trimmed by stemming process. The tokenization process is common for all the text mining research.

2. LITERATURE SUPPORT

Domain classifier in text mining is nothing but topic modeling. In real world, the sequential patterns in natural language usually appear without explicit boundaries but with the variations of temporal topics [2]. When compared to Vector Space Model (VSM), Topic Model (TM) represents documents in much lesser dimensional space [3].

The curse of dimensionality and its reduction is studied and surveyed [4]. Further statistical composition of patterns from text can be learned from Dimensionality Reduction methodologies. To follow up this survey, VSM is inclined with Singular Value Decomposition (SVD) and classified using Support Vector Machine (SVM) [5].

The feature extraction phase is found to be the primary cause for knowledge associate for the research on text categorization. Probabilistic Latent Semantic

Analysis (PLSA) [3] is known as a latent topic document model, which discovers the semantic structure in training documents. It calculates the word probability by

$$p(w \mid d) = \sum_{c=1}^{n} p(w \mid c) p(c \mid d)$$

where the formula denotes the distribution of word w in topic c and the distribution of topic in a document d. LDA extends PLSA [6] model in the document segmentation model [2]. It consists of a column vector of Dirichlet distribution with parameter ~ , topic word probabilities and topic sequence ¯ . The document segmentation model with dLDA [7] is formed by sequence of block bk. The distribution of word in each block and topic distribution are determined. Similarity between each block bk and bk+1 is computed using cosine formula

$$s(b_k, b_{k+1}) = \frac{\sum p(w \mid b) p(w \mid b_{k+1})}{\sqrt{\sum p(w \mid b_k)^2} \sqrt{\sum p(w \mid b_{k+1})^2}}$$

The Naive Bayes classifier [8], a simple Bayesian classification algorithm is an effective basic approach for text categorization. This has been taken up for testing the TIM algorithm with the conventional tf-idf.

$$P(c \mid d) \propto P(c). P(d \mid c)$$

By Naļve Bayes way, the words in the documents are conditionally independent and is given by

$$P(c \mid d) \propto P(c) \prod_{w \in d} P(w \mid c)$$

The SVM [13] classifier is well suited for text categorization and acknowledges the sparse data. SVM fixed data in the hyper plane. SVM is best suited for the binary classification. It separates dataset into classes with the help of Kernel functions. The Kernel functions depends of the complexity of the data. They are linear and polynomial. The details of SVM is clearly studied in [14].

3. SYSTEM STUDY

An information table [9] represents complete information and knowledge, i.e., the objects are processed, observed, or measured by the finite number of attributes.

3.1 Definition of Information System:

An Information System is a pair S = (U, A), with every a εA, a set of its values Va is associated.
It is expressed as:

$$S = (U, A, \{V_a \mid a \in A\}, \{I_a : U \rightarrow V_a \mid a \in A\})$$

where 'U' is a Universal finite non empty set of objects, 'A' is a finite non empty set of Attributes, Va is a domain of Attributes $a \in A$, and $I_a : U \rightarrow$ Va is an Information function that maps an object in U to Va.

The information system has to be presented in such a way that for any related classification problem, a correct decision can be derived.

3.2 Definition of Document Information

Document Information is defined by information system as a pair P = (D, T), set of its values V_2 is associated where $^2 \in T$. It is expressed as: $P = (D, T, \{V_2 \mid ^2 \in T\}, \{I_2 : D \rightarrow V_2 \mid ^2 \in T\})$
Where 'D' is a Document collection, 'T' is finite terms, V_2 is a domain of terms $I_2 : D \rightarrow V_2$ is an Information function that maps terms in D to V_2 .

4. DATASETS

4.1 Movie Review Dataset

The Movie Review Dataset, Polarity dataset v0.9 with 700 positive and 700 negative reviews is used. Using movie reviews as data, the problem of classifying documents using standard machine learning techniques definitively outperform human-produced baselines processed reviews[10]. The training cases are chosen randomly from each classes about 100 documents each. Which means about 200 cases is considered for training.

4.2 Reuters-21578 Data Set

The Reuters-21578 Data Set collection provides a classification task with challenging properties. There are multiple categories, the categories are overlapping and non exhaustive, and there are relationships among the categories. There are interesting possibilities for the use of domain knowledge. There are many possible feature sets that can be extracted from the text, and most plausible feature/example matrices are large and sparse [11].

5. PROPOSED APPROACH

The proposed approach **TIM (IM**mune **T**erm) has been inspired from segment based approach [2][7][12]. TIM is a local approximation approach, in which each term is treated as a granule in each document. Each document is segmented or partitioned into a constant *k*. If a term *t* is distributed throughout the partitions and is

satisfies a threshold then it is included in the global set as *t'*.

$$t' = \frac{\sum p(\Gamma / \kappa)}{\kappa} \quad \text{where} \quad \kappa = \sum_{i=1}^{k} (t_i \mid \Gamma_i > 0)$$

The algorithms TIM and conceptgranule are designed is such a way to check whether a term is immunized i.e., **t << t' (**Dimension reduction achieved)

Algorithm: TIM(Corpus)

Input: '*n*' No. of Training Documents *D*
No. of document Partitions '*k*'
No. of local terms '*t*'
gm[][] – granule matrix with '*t*' terms
im[][] – immune matrix with '*t*'' terms

Output:
Reduced Documents with local immune terms *t'*

i. get value for *k*
ii. for *i* = 1 to *n* // training documents
iii. Perprocess each document by trimming and stop word removal
iv. Initialize a granule matrix gm[][] with *t* terms
v. *t'=1*
vi. for *j* = 1 to t // terms
vii. for *l* = 1to k // partitions
viii. $\Gamma[l] = gm[j][l]$
ix. If **conceptgranule**(Γ, k) then //single document in matrix
x. i*m*[*i*][++t'] = **conceptgranule**()//using the term otherwise drop the term

Algorithm: conceptgranule()

Input: term array
- terms

Output: Non zero value for immune terms or zero for non immune terms

i. = 0
ii. for *l* = 1to k // partitions
iii. if () then
iv. ++
v. if <= 3*k/4 then // a thershold
vi. for *l* = 1to k // partitions
vii. $t' = t' + \Gamma_l$
viii. r = t'/
ix. return (r) //include in term frequency matrix
x. else
xi. return (0)

Reduction of bag of words to concept granule is the feature extraction adopted in the above said algorithm. If U is the Universal set i.e., bag of words of document, then g, the set of concept granule set is obtained by calculating the frequency of each word in each partition.

D_j is a matrix with k X i matrix with vector g_{ik} where 'i' represents term or words 'k' represents partition of a document. $t_{i,j}$ is a single vector or a set of words in a document, where i' represents immune word i.e., each word is treated for its importance in the whole document by the Algorithm1. When a word is able to withstand the treatment, then that is said to be immune word and i represents document. The TIM approach is supposed to work as mentioned in the Figure.1

6. EXPERIMENTAL RESULTS

The experiment is taken with the Reuters and Movie Review data sets. The connection between words and the respective topic are taken into consideration. At first the training to the topics such as "wheat", "trade", "ship", etc. are taken up in Reuters and star level such as "1", "1.5", "2", etc. are taken up in Movies Review datasets. This is done with random selection of documents with the respective topic. Further, testing is done for each of the trained topic. The result is studied by micro averaging the topics and presented in the Table 1.

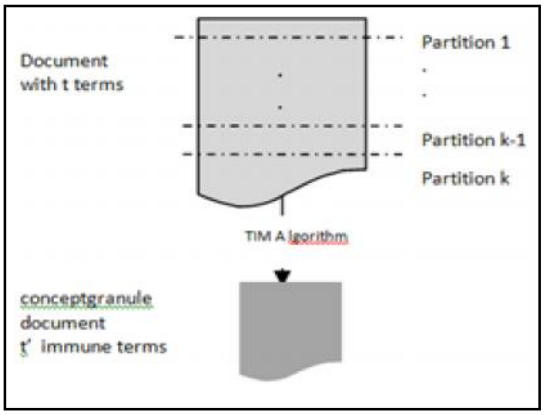


Fig.1 TIM

Table 1 Micro Average of PCA and LSI Reduced NB and SVM Classified Reuters and Movie Review Datasets

Dataset	Classifiers	tf-idf		TIM	
		PCA	LSI	PCA	LSI
Reuters	NB	0.785	0.791	0.912	0.921
	SVM	0.824	0.892	0.939	0.932
Movie Review	NB	0.722	0.706	0.811	0.792
	SVM	0.788	0.745	0.852	0.874

The TIM algorithm which has been proposed, shown positive play in dimension reduction (using PCA and LSI) and classification (using NB and SVM). This could be comparably seen in the dimension reduction scheme represented in the Figure 2.

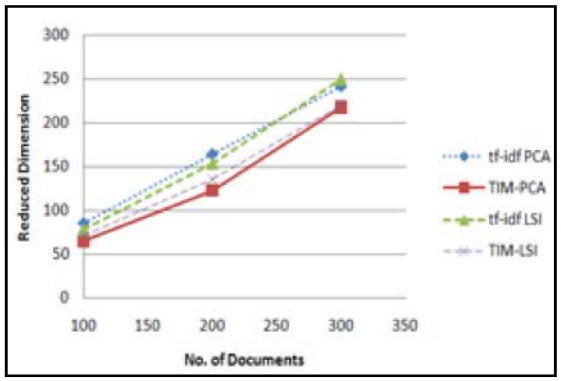


Fig.2 Dimension reduction of tf-idf and TIM preprocessing

On the Reuters the topical word support is much when compared to the Movie Reviews. This is reflected in the results of the classification. Reuters could be classified at higher end whereas the Movie Review could be able to respond to level less. The main concentration is given to the preprocessing level of documents. TIM algorithm is introduced as next phase to tf-idf. The results shows much better improvement in dimension reduction and classification.

One advantage of using TIM approach is, documents are treated much in lower level, than the inclusion of ontology or Parts of speech tagging. Because of this, much of the time is reduced in preprocessing level.

One difficulty of using TIM approach is that, each of the training document is partitioned into k segments in the training level. Because of this the training time is fat, but testing time is much reduced by looking only for the topical words.

7. CONCLUSION

TIM algorithm developed with the inspiration of concept granulation and equal partition based approach has found to be yielding better performance result with the support of PCA, LSI, NB and SVM approaches. The magnitude of terms is playing the major role of the algorithm.

Though PCA, LSI, NB and SVM are highly repeated techniques, they seem to be much beneficial in understanding the proposed work. The TIM-PCA and TIM-LSI have marked computationally changes in the dimension. In spite of this one important point is noted. The datasets play the major role in classification. The Reuters dataset is topic oriented and has been classified with more than ten topics. The Movie Review is not topically distributed. It is proposed for sentimental classification. i.e., positive and negative sentiments. The major role is played by the adjective and adverb terminologies, rather than the topical words.

NB and SVM is showing low performance in the Movie Review dataset than the Reuters dataset. It has been understood that the preprocessing of documents much in the lower level or local optimization is very essential for unstructured data.

REFERENCES

[1] Ronen Feldman, James Sanger, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data", Cambridge University Press, 2007.

[2] Jen-Tzung Chien and Chuang-Hua Chueh, "Topic-Based Hierarchical Segmentation", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 1, January 2012.

[3] T. Hofmann, "Probabilistic Latent Semantic Indexing", in Proceedings of the ACM SIGIR, 1999, pp.50-57.

[4] D. Malathi and S.Valarmathy, "A Comprehensive Survey on Dimension Reduction Techniques for Concept Extraction from a Large Corpus", IJCIS, International Journal of Computing Information Systems, ISSN No.2229-5208, Vol.3, No.5, 2011, pp.1-6,

[5] D. Malathi and S.Valarmathy, "Conceptually Co-occurring Words Included as Feature Selection in Text Document Classification using SVD and SVM", International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697, Vol. 3, No. 7, Nov-Dec, pp.145-148.

[6] D. M. Blei, Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning, Vol. 3, No. 5, 2003, pp.993-1022.

[7] T.Brants, F. Chen and I. Tsochantaridis, "Topic-based Document Segmentation with Probabilistic Latent Semantic Analysis", In the Proceeding of International Conference on Information and Knowledge Management, 2002, pp.21-218.

[8] D. D. Lewis, "Representation and Learning in Information Retrieval", Ph.D. Dissertation, Amherst, MA, USA, 1992.

[9] Z. Pawlak, "Rough Sets", International Journal of Information and Computer Science, Vol. 11, No.5, 2002, pp.341-356.

[10] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning

Techniques", in the Proceedings of EMNLP, 2002, pp.79-86.

[11] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[12] Andrea Tagarelli, George Karypis, "A Segment-based Approach To Clustering Multi-Topic Documents, Text Mining Workshop", in the Proceeding of SIAM Data mining Conference, 2008.

[13] Simon Tong and Daphne Koller, "Support Vector Machine Active Learning with Applications to Text Classification", Journal of Machine Learning Research, 2001, pp.45-66.

[14] D. Christopher Manning, Prabhakar Raghavan and Hinrich Schьtze, "Introduction to Information Retrieval", Cambridge University Press, 2008.

Indian Journal of Engineering, Science, and Technology (IJEST)

(ISSN: 0973-6255)
(A half-yearly refereed research journal)

Information for Authors

1. All papers should be addressed to The Editor-in-Chief, Indian Journal of Engineering, Science, and Technology (IJEST), Bannari Amman Institute of Technology, Sathyamangalam - 638 401, Erode District, Tamil Nadu, India.
2. Two copies of manuscript along with soft copy are to be sent.
3. A CD-ROM containing the text, figures and tables should separately be sent along with the hard copies.
4. Submission of a manuscript implies that : (i) The work described has not been published before; (ii) It is not under consideration for publication elsewhere.
5. Manuscript will be reviewed by experts in the corresponding research area, and their recommendations will be communicated to the authors.

Guidelines for submission

Manuscript Formats

The manuscript should be about 8 pages in length, typed in double space with Times New Roman font, size 12, Double column on A4 size paper with one inch margin on all sides and should include 75-200 words abstract, 5-10 relevant key words, and a short (50-100 words) biography statement. The pages should be consecutively numbered, starting with the title page and through the text, references, tables, figure and legends. The title should be brief, specific and amenable to indexing. The article should include an abstract, introduction, body of paper containing headings, sub-headings, illustrations and conclusions.

References

A numbered list of references must be provided at the end of the paper. The list should be arranged in the order of citation in text, not in alphabetical order. List only one reference per reference number. Each reference number should be enclosed by square brackets.

In text, citations of references may be given simply as "[1]". Similarly, it is not necessary to mention the authors of a reference unless the mention is relevant to the text.

Example

[1] M.Demic, "Optimization of Characteristics of the Elasto-Damping Elements of Cars from the Aspect of Comfort and Handling", International Journal of Vehicle Design, Vol.13, No.1, 1992, pp. 29-46.

[2] S.A.Austin, "The Vibration Damping Effect of an Electro-Rheological Fluid", ASME Journal of Vibration and Acoustics, Vol.115, No.1, 1993, pp. 136-140.

SUBSCRIPTION

The annual subscription for IJEST is Rs.600/- which includes postal charges. To subscribe for IJEST a Demand Draft may be sent in favour of IJEST, payable at Sathyamangalam and addressed to IJEST. Subscription order form can be downloaded from the following link http:// www.bitsathy.ac.in./ ijest.html.

For subscription / further details please contact:

IJEST
Bannari Amman Institute of Technology
Sathyamangalam - 638 401, Erode District, Tamil Nadu Ph: 04295 - 226340 - 44
Fax: 04295 - 226666 E-mail: ijest@bitsathy.ac.in Web:www.bitsathy.ac.in

Indian Journal of Engineering, Science, and Technology

Volume 7, Number 1, January - June 2013

CONTENTS