International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 395 - 399, 2014, ISSN: 2319–8656

A Review of Constraint Programming

Poonam Dabas
Department of CSE
U.I.E.T, Kurukshetra University
Kurukshetra, India

Vaishali Cooner
Department of CSE
U.I.E.T, Kurukshetra University
Kurukshetra, India

Abstract: A constraint is defined as a logical relation among several unknown quantities or variables, each taking a value in a given domain. Constraint Programming (CP) is an emergent field in operations research. Constraint programming is based on feasibility which means finding a feasible solution rather than optimization which means finding an optimal solution and focuses on the constraints and variables domain rather than the objective functions. While defining a set of constraints, this may seem a simple way to model a real-world problem but finding a good model that works well with a chosen solver is not that easy. A model could be very hard to solve if it is poorly chosen.

Keywords: Constraint Programming; Optimization; feasibility; problems; relations

1. INTRODUCTION

The development of high-tech systems is very difficult without mathematical modeling and analysis of the system behavior. For this, mathematical models are revealed in order to solve the tasks in many areas like in the modern engineering sciences like control engineering, communications engineering, and robotics. Therefore, the main focus is that without neglecting mathematical accuracy on comprehensibility and real-world applicability. Mathematical engineering has various methods to find the optimal and feasible solution like: Linear programming, Non-Linear programming, stochastic programming and Constraint programming.

Linear programming is effective only if the real world is reflected in the model used. They also sometimes give results that don't make sense in the real world. Even some situations have many possibilities to fit into linear programming. A constraint is a logical relation among several unknown quantities (or variables), each taking a value in a given domain.

2. CONSTRAINT PROGRAMMING

A logical relation among several unknown variables is known as a constraint, where each variable takes a value in a given domain. The basic idea behind constraint programming framework is to model the problem as a set of variables with domains and a set of constraints [16]. The possible values that the variables can take are restricted by the constraints.

In operations research constraint programming (CP) is an emergent field. It is based on finding a feasible solution i.e. feasibility rather than finding an optimal solution i.e. optimization. Basic CP constructs, the interface for advanced scheduling applications, and search specification are provided which are essential to a language supporting constraint programming and are represented as discrete variables [1].

The focus is not done on objective function rather than the constraints and variables domain. It possesses a strong theoretical foundation though it is quite new, a widespread and very active community around the world and an arsenal of different solving techniques. In problems with heterogeneous constraints CP has been successfully applied in planning and scheduling.

A programming paradigm where relations between variables are stated in the form of constraints is known as constraint programming. In other programming languages step or sequence of steps is not specified to execute. Because of this constraint programming a known as a form of declarative programming.

Various kinds of constraints are used in constraint programming: one is those used in constraint satisfaction problems for example- A or B is true, other one is those solved by the simplex algorithm for example- $x \leq 5$, and others.

To solve scheduling problems constraint programming is an interesting approach. Activities are defined by their starting date in cumulative scheduling; their duration and the amount of resource necessary are also defined for their execution.

Constraints are defined as just relations and which relation should hold among the given decision variables is stated by a constraint satisfaction problem (CSP). It may seem a simple while defining a set of constraints as a way to model a real-world problem but it is not easy to find a model that works well with a chosen solver. It is really hard to solve a poorly designed model. To take advantage of the features of the model such as symmetry solvers can be designed to save time in finding a solution. As many are over constrained this may exist as another problem with modeling real-world problems. Any language can be used to implement constraint solver.

For all the constraints to be satisfied there must exist an assignment of values to variables. To reduce the computational effort this technique is used which is needed to solve combinatorial problems. Constraints are used in a constructive mode to deduce new constraints, not only to test the validity of a solution. Constraints also detect inconsistencies rapidly.

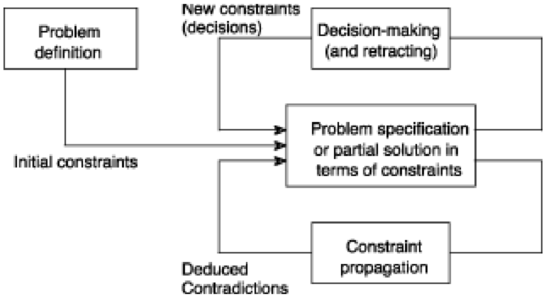

Problem definition
New constraints (decisions)
Decision-making (and retracting)
Problem specification or partial solution in terms of constraints
Initial constraints
Constraint propagation
Deduced Contradictions

Figure. 1 Behavior of a Constraint Programming system

Constraint propagation is usually incomplete for complexity reasons. So, not all but some of the consequences of constraints are deduced. All inconsistencies cannot be detected by constraint propagation.

To determine if the CSP instance is consistent or not tree search algorithms must be implemented. The figure depicts the overall behavior of a constraint-based system.

First, variables and constraints are defined as terms of the problem

Then, constraint propagation algorithms are specified. Some pre-defined constraints can be used by the constraint programming tool like scheduling constraints for which the corresponding propagation algorithms have been pre-implemented.

Finally, at last the decision-making processes. It is the way the search tree is built, and is specified. How new constraints are added to the system are shown in it like ordering a pair of activities.

3. REVIEW ANALYSIS

Over the past few years, there has been lot of research going on in the field of mathematical engineering to find the optimal solutions for the problems. Researchers have done a lot in this field which is discussed below:

Willem-Jan van Hoeve[1] has presented the modeling language for basic constraint programming and advanced scheduling constructs and specify how search can be controlled. It provides easy development of hybrid approaches such as CP based column generation. Focus here is done on the constraint programming interface of AIMMS which is based on an algebraic syntax and offers access to integer linear programming, quadratic programming (QP) and nonlinear programming (NLP).

Arnaud Lallouet, M. lopez, L. Martin, C. Vrain [2] have made an algorithm which is designed combining the major qualities of traditional top-down search and bottom-up search techniques. The contributions of this paper are setting the framework of learning CSP specifications, then the choice of the rule language, and it's rewriting into CSP and the learning algorithm which allows guiding search when traditional method fails. In this the activity of finding the constraints that are to be stated is considered as a crucial part and a lot of work has been spent on the understanding and automation of modeling tasks for the novice users who have a limited knowledge regarding how to choose the variables. A framework is designed to bridge the gap between constraint programming modeling language and ILP (Inductive Logic Programming). The very first step of the framework consists in learning a CPS (Constraint Problem Specification) describing the target problem. ILP framework and its applications to learning problems are presented.

Barry O'Sullivan [3] has presented technical challenges in the area of constraint model acquisition, formulation and reformulation algorithms for global constraints and automated solving and it also presents the metrics by which success and progress can be measured. The motivation here is to reduce the burden on constraint programmers and to increase the scope of problems that can be handled alone by domain experts. Modeling defines the problem, in terms of variables that can take different values. Progress is evaluated empirically in constraint programming. A model for practical problem as a constraint satisfaction problem (CSP) is preferred and available constraint programming tools are used to solve it. Generic methods from the machine learning field can be applied to learn an appropriate formulation of the target problem as a CSP. The filtering algorithm is difficult to design and this is considered the major challenge that one faces when designing a new global constraint.

Christian Bessiere, R. Coletta, T. petit [4] have presented a framework for learning implied global constraints which is presented in a constraint network assumed to be provided by a non-expert user. As global constraints are key feature of constraint programming learning global constraints is important. A motivation example is considered and it is shown that if it is required that the model is to be solved with more tasks then the need to improve model is needed. Constraint network is defined by a set of variables and a set of domains of values for the variables. The tighter the learned constraint is, the more promising its filtering power is. A general process to learn the parameters of implied global constraints is given. The focus is made on global constraints and set of parameters. Efficient algorithm exists to propagate when the cardinalities of the value are parameters that take values in a range. A model was generated to minimize the sum of preference variables. This was considered the first approach that derives implied global constraints according to the actual domains. Experiments show that a very small effort spent learning implied constraints with this technique can improve the solving time.

Steven J. Miller [5] has described linear programming as an important generalization of linear algebra. Various real world situations are modeled successfully using programming. The problems that can be solved by linear programming are discussed. Binary integer linear programming is also discussed which is an example of a more general problem is called Integer Linear Programming. The difficulty here due to the fact that a problem may have optimal real solutions and optimal integer solutions but both the solutions need not be closed to each other. The simplex method is used for solving the linear problems to find the optimal solutions. It has two phases, one is to find a basic feasible solution and other one is to find a basic optimal solution, given a basic feasible solution. If no optimal solution exists this phase produces a sequence of solutions that are feasible with their cost tending to minus infinity. Algorithms are defined for them. The time for finding the optimal solution is also considered as a major factor here.

Nicholas Nethercte, P J. Stukcey, R. Becket, S. Brand, G J. Duck and Guido Tack [6] have presented MiniZinc as a simple and expressive CP modeling language. It is known that there is no standard modeling language for constraint programming problems so most solvers have their own language for modeling. The experimentation and comparison between different solvers is encouraged with a standard language for modeling CP. This MiniZinc problem has two parts- model and data which may be in separate files. The assignments to parameters declared in the model are contained in the data file. The model file is not attached to any particular data file. Boolean, integers, and floats are the three scalar types provided and sets and arrays are two compound types provided. The MiniZinc is translated to FlatZinc in two parts as flattening and the rest. Flattening is done in a number of steps to reduce the model and data as much as possible. The order of the steps is not fixed. After flattening, post flattening steps are applied. Different MiniZinc to FlatZinc converters are used. The main goal here was to define a language which is not too big but expressive.

Alan M. Frisch, M. Grum, C. Jefferson, B.M. Hernandez, Ian Miguel [7] have discussed a new formal language ESSENCE for specifying combinatorial problems which provides a high level of abstraction. This language was a result of attempt to design a formal language that enables abstract problem. For this language no expertise in CP should be needed, it is accessible to anyone with knowledge of discrete mathematics as it is based on the notation and concepts of discrete mathematics. It provides high level of abstraction stating that the language should not force a specification to provide unnecessary information. This language provides an exceptionally rich set of constructs for expressing quantification. It also supports complex, nested types and also its result can be specified without modeling them.

Adrian Petcu [8] has discussed in brief about efficient optimization techniques that are essential to coordinate to business companies and distributed solution processes are desirable as they allow the participating actors to keep control on their data and also offer privacy.

Many key issues are presented that are present in this domain like the actors involved in the distributed decision processes do not have the global knowledge and overview. The goal of constraint optimization is to find the best assignment of values to the variables so that utilities are maximized and cost is minimized. A new technique based on dynamic programming was developed for distributed optimization which was a utility propagation mechanism and works on constraint problems.

It requires only a linear number of messages for finding the optimal solution. These algorithms for distributed constraint optimization have not been applied to large scale due to complexity reason.

Brahim Hnich, S.D. Prestwich, E. Selensky, B.M. Smith[9] have developed models for constraint programming for finding an optimal covering array. It is shown that the compound variables that represent tuples of variables in the original model, allow the constraints of the problem to be represented more easily, propagating better. The optimality of existing bounds is proved for finding the optimal solutions for moderate size array. In covering test problems instances are used with coverage strengths. Number of parameters here is varied. It has shown that for moderate problem size one can find provably optima solution using CP approach. One of the advantages of CP is easy handling of side constraints i.e. simply by adding them to the model.

C. Bessiere, J. Quinqueton, G. Raymond [10] have proposed an automated model to generate different viewpoints for the problem we are to model. The main idea here is to build a viewpoint enough to describe many different solutions of problems also describes a solution of the target problem. Historical data is with which it is started and historical data is used as solutions to problems close to the target problems. From this data candidate variables are extracted. So this can be seen that these viewpoints are capable of describing the historical solutions and also the solutions of our target problem. The goal here is to build viewpoints which match the given historical data. For this candidate variables are determined according to the history. A set of potential viewpoints are obtained out of which more relevant is selected to build constraint models efficiently.

P.E. Hladik. A.M. Deplanche, N. Jussien, H. Cambazard [11] has presented an approach to solve hard real time allocation problem i.e. to assign periodic tasks to processors in context of fixed priority preemptive scheduling. Bendors decomposition is also used as a way of learning when the allocation yields a valid solution. The problem is distributed in systems that belongs to a class. The authors presents a decomposition based method which separates the allocation problem from the scheduling one. The three classes that the constraint allocation problem must respect are timing, resource, and allocation constraints. For solving a master problem using constraint programming, the problem needs to be translated into CSP. The subproblem is considered as to check whether a valid solution produced by master problem is schedulable or not. If no data is sent then deadlines can correspond to non-communicating tasks. The overall problem is split into a master problem for allocation and resource constraints and a subproblem for timing constraints. The learning technique is used in an effort to combine the various issues into a solution that satisfies all constraints.

Julia L.Higle [12] has presented an introduction to stochastic programming models. Stochastic linear programming is resulted when some of the data elements in a linear program are appropriately described using some random variables. An example is illustrated giving the reason why SP model is preferred and some essential features of a stochastic program are identified. Stochastic programs are difficult to solve and formulate. When the size of the problem increases we can easily see that the solution difficulties increase as well. Sensitivity analysis is done which provides a sense of security and is important. It is used to study the robustness of the solution to a linear programming model. It is done for the accuracy of the data to check whether the solution changes or not on changing the data. If the solution remains same it is believed that the solution is appropriate and vice versa. All the uncertainties should be included in the model.

Philippe Refalo [13] has presented a new general purpose strategy for constraint programming which is inspired from integer programming technique. The importance of a variable for the reduction of the search space is measured by the impact. Designing the search strategy is difficult in integer programming whereas the concept of domain reduction is easier to understand and the use design of a search strategy is easier in constraint programming. In the impact based search strategy, by storing the observed importance of variables impacts permit us to benefit from the search effort made up to a certain node. With some standard strategies some instances remain unsolved which are solved by this technique. Certain principles are defined here for reducing the search effort. When a value is assigned to a variable in constraint programming, constraint propagation reduces the domains of other variables defined.

Y.C Law, J.H.M. Lee [14] has introduced model induction which is a systematic transformation of constraints in an existing model to constraints in another viewpoint. Three ways of combining redundant models are proposed using model induction, another way is model channeling, and the last is model intersection. It is also investigated how the problem formulation and reformulation affect execution efficiency of constraint solving algorithms. For the formulation process the variables and the domain of the variables is to be determined. The induced model is result of the model induction. The three ways of combining the redundant models are proposed so as to utilize the redundant information in enhancing constraint propagation. Alternate ways of generating models in a different viewpoint from existing model are made.

H.Y. Benson, D.F. Shanno, R.J. Vanderbei [15] have analyzed the performance of several optimization codes on large-scale nonlinear optimization problems. The size of problem is defined to the number of variables and the number off constraints. Some of the codes are tested and presented available for solving large scale NLP's. To identify the features of these codes that are efficient is the goal. Infeasibilities and unboundedness are detected in the problem as early as possible. Performance of the algorithms running on the same set of problems is compared to simple compute an estimate of the probability that an algorithm performs. A number of conclusions concerning specific algorithm details exits if various algorithms are compared. Numerical result for solving large scale nonlinear optimization problems is presented. The performance of each solver is explained easily and predicted based on the characteristics.

4. CONCLUSION

There are many challenges faced by mathematical engineering approaches to find the optimal solution. The common weakness to all of the approaches is the assumption that the input data are perfectly accurate. Many benefits of using this approach are discussed as visualization of results using activities and resources. From an existing model another model of a different viewpoint can be generated in a systematic way. Experiments show that we can improve the solving time by very small effort is spent in learning. But at the same time it is too expensive. There are many challenges as: these rely heavily on supervision of an expert and also they are not capable of acquiring a description of the problem class.

5. REFERENCES

[1] Willem-Jan van Hoeve, “Developing Constraint Programming Applications with AIMMS,” in CP,2013.

[2] Arnaud Lallouet, Matthieu Lopez, Lionel Martin, Christel Vrain, “On Learning Constraint Problems,” in ICTAI, 2010.

[3] Barry O’Sullivan, “Automated Modelling and Solving in Constraint Programming,” in proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence(AAAI-10), 2010.

[4] C. Bessi`ere, R. Coletta, and T. Petit, “Learning implied global constraints,” in IJCAI, 2007, pp. 44–49.

[5] Steven J. Miller,” An Introduction to Linear Programming,” in mathematics,2007.

[6] N. Nethercote, P. J. Stuckey, R. Becket, S. Brand, G. J. Duck, and G. Tack, “Minizinc: Towards a standard cp modelling language,” in CP, 2007, pp. 529–543.

[7] A. M. Frisch, M. Grum, C. Jefferson, B. M. Hern´andez, and I. Miguel,“The design of essence: A constraint language for specifying combinatorial problems,” in IJCAI, M. M. Veloso, Ed., 2007, pp. 80–87.

[8] Adrian Petcu,” Recent Advances in Dynamic, Distributed Constraint Optimization,” in infoscience, 2006.

[9] Brahim Hnich, Steven D. Prestwich, Evgeny Selensky, Barbara M. Smith, “Constraint Models for the Covering Test Problem,” in CP, 2006, pp. 199-219.

[10] C. Bessiere, J. Quinqueton, and G. Raymond, “Mining historical data to build constraint viewpoints,” in Proceedings CP’06 Workshop on Modelling and Reformulation, 2006, pp. 1–16.

[11] Pierre-Emmanuel Hladik, Hadrien Cambazard, Anne-Marie Deplanche, Narendra Jussien,” in ECRTS, 2005.

[12] Julia L. Higle,” Stochastic Programming: Optimization When Uncertainty Matters,” in operations research informs-New Orleans 2005, 2005.

[13] Philippe Refalo,” Impact-Based Search Strategies for Constraint Programming,” in peasant IBS, 2004.

[14] Y.C. Law, J.H.M. Lee,” Model Induction: a New Source of CSP Model Redundancy,” in AAAI, 2002.

[15] Hande Y. Benson, David F. Shanno, Robert J. Vanderbei,” A Comparative Study of Large-Scale Nonlinear Optimization Algorithms,” in NLP, 2002.

[16] Roman Bartak,” Constraint-Based Scheduling: An Introduction for Newcomers,” in SOFSEM, 2002.

[17] R. Bart´ak. Constraint programming: In pursuit of the holygrail. In *Proc. of WDS99*, 1999.

[18] J. Charnley, S. Colton, and I. Miguel, “Automatic generation of implied constraints,” in ECAI, 2006, pp. 73–77.

[19] J.-F. Puget, “Constraint programming next challenge : Simplicity of use,” in International Conference on Constraint Programming, ser. LNCS, M. Wallace, Ed., vol. 3258. Toronto, CA: Springer, 2004, pp. 5–8, invited paper.

[20] A. M. Frisch, M. Grum, C. Jefferson, B. M. Hern´andez, and I. Miguel, “The design of essence: A constraint language for specifying combinatorial problems,” in IJCAI, M. M. Veloso, Ed., 2007, pp. 80–87.

International Journal of Computer Applications Technology and Research
Volume 3-Issue 7,400-402, July 2014

Finger Vein Detection using Gabor Filter, Segmentation and Matched Filter

Poonam Dabas
Computer Science Department
UIET,Kurukshetra University
Kurukshetra, India
poonamdabas.kuk@gmail.com

Amandeep Kaur
Computer Science Department
UIET, Kurukshetra University
Kurukshetra, India
amandeepg5s@gmail.com

Abstract: This paper propose a method of personal identification based on finger-vein patterns. An image of a finger captured by the web camera under the IR light transmission contains not only the vein pattern itself; but also shade produced by various thickness of the finger muscles; bones; and tissue networks surrounding the vein. In this paper; we intro-duce preliminary process to enhance the image quality worsened by light effect and noise produced by the web camera; then segment the vein pattern by using adaptive threshold method and matched them using improved template matching. The main purposes of this paper are to investigate finger-vein technology; the applicable fields and whether finger-vein detection can solve the problems fingerprint detection imposes in certain industries.

Keywords: Finger-vein detection, Gabor filter, filter, pattern recognition.

1. INTRODUCTION

Smart recognition [1][2] of human identity for security and control is a global issue of concern in our world today. The financial losses due to identity theft can be severe; and the integrity of security systems compromised. The automatic authentication systems for control have found application in criminal identification; autonomous vending and automated banking among others. Amongst the more authentication systems that have been proposed and implemented; finger vein biometrics is emerging as the foolproof method of automated personal identification. The Finger vein is a unique physiological biometric for identifying individuals based on the physical characteristics and attributes of the vein patterns in the human finger. This is a fairly recent technological advance in the field of biometrics that is being applied to different fields such as medical; financial; law enforcement facilities and other applications where high levels of security or privacy is very important. The technology is impressive because it requires only small; relatively cheap single-chip design; and has a very fast identification process that is contact-less and of higher accuracy when compared with other identification biometrics like fingerprint; iris; facial and others. This much accuracy rate of finger vein is not unconnected with the fact that finger vein patterns are virtually impossible to forge thus it has become one of the fastest growing new biometric technology that is quickly finding its way from research labs to commercial development.

Vein image is obtained from the infrared image collection. Research has proved that the absorption of infrared light in human tissues is comparatively low, i.e. the infrared light has more penetrating ability for human body. Though the hemoglobin has a strong absorption of infrared light, experiment proves that quasi-infrared light can make good imaging of subcutaneous blood vessel of 0-1 cm depth [3].

Due to the poor infrared image quality, image enhancement and Region of Interest (ROI) techniques [4] are introduced in data collection module. After the optimized image is obtained, the vein pattern is derived according to segmentation, skeletonization refinement and feature definition procedures, and then the pattern is saved to the database as template in the registration mode. In the verification mode, the pattern is compared with the templates in the database and result can be obtained. Since vein sampling is from subcutaneous image, it is hard to be destroyed and stolen so that it can be an ideal candidate of high-level verification technique.

In this paper, we studied the vein identification [6] on trial and discovered that there were several difficulties:

a. The image grabbed by the common web camera consists of salt and pepper noise and the gray level distribution among different trial is not the same, because the web camera always does the brightness adjustment.

b. On normal conditions, gray scale discrimination of vein image is very small. We need a good threshold segmentation to get the effective binary image that provides sufficient finger vein information.

c. The pressure given on our finger will cause the vein in-side shrink or changed. So, we need to build a "less-strict" finger slot to let the user's finger in a "relax" condition.

First, a new approach for personal identification that utilizes simultaneously acquired finger-vein and finger surface (texture) images is presented. The experimental results illustrate significantly improved performance that cannot be achieved by any of these images employed individually.

2. Finger-Vein Image Pre-processing

The acquired images are noisy with rotational and translational variations resulting from unconstrained imaging.

[5]. Therefore, the acquired images are first subjected to pre-processing steps that include:

1. Segmentation of ROI,
2. Translation and orientation alignment, and
3. Image enhancement to extract stable/reliable vascular Patterns.

The enhanced and normalized ROI images [7] are employed for feature extraction. The key objective while segmenting the ROI is to automatically normalize the region in such a way that the image variations; caused by the interaction of the user with the imaging device; can be minimized. The order to make the identification process more effective and efficient; it is necessary to construct a coordinate system that is invariant or robust (or nearly) to such variations. This is judicious to associate the coordinate system with the Finger itself since we are seeking the invariance corresponding to it. As a result two webs are utilized as the reference points/line to build up the coordinate system i.e. web among the index finger and middle finger together with the web between the ring finger and little finger. These web points are easily identified in touch-based imaging (using pegs) but should be automatically generated for contactless imaging [8].

The acquired Finger images are first binarized, so that we are able to separate the Finger region from the background region. It is followed by the estimation of the distance from centre position of the binarized Finger to the boundary of Finger.

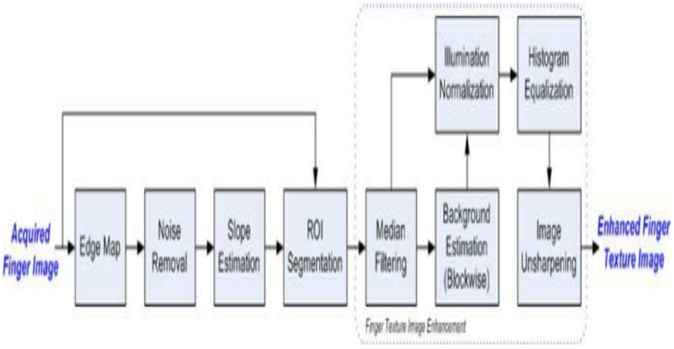


Figure 1. Block diagram illustrating key steps employed for the preprocessing of acquired finger texture images from a webcam.

3. Recognition

The same as fingerprint recognition [9], finger vein recognition also contains image pre-processing, feature extraction and matching. The image normalization, finger vein segmentation, thinning is all images pre-processing work. Then acquisition time; light intensity of acquisition environment; thickness of each finger; and intensity distribution of finger vein images are all different; hence normalization of image size and gray scale is indispensable in order to feature extraction and classification.

Finger vein contains ridge and valley lines. In addition to show paragraphs must be indented. Then all paragraphs must be justified; both left-justified and right-justified irregular shape in the minutiae and singularity regions; the ridge and valley lines show continuous and smooth change in most regions. The repeated line tracking method gives a promising result in finger-vein identification: The idea is to trace the veins in the image by chosen directions according to predefined probability in the horizontal and vertical orientations; and the starting seed is randomly selected; the whole process is repeatedly done for a certain number of times.

This process of improving the quality of a digitally damage by manipulating the image with software. This is quite easy; for example to make an image lighter or darker or to increase or decrease contrast.

In paper; Gabor filter was used to enhance finger vein images. In this paper, the systematical development of a new approach for the finger-vein feature extraction using Gabor filters is introduced. And in addition; we also investigate a new feature extraction approach using matched filters as the matched filters have been successfully utilized for the enhancement of retinal features in [10].

The Gabor filters are inspired by the multichannel processing of visual information in the biological model of human visual system and are known to achieve the maximum possible joint resolution in the spatial and spatial-frequency domains [11], which have been effectively utilized by researchers to develop object segmentation paradigm. This paper; proposed the framework for the finger vein feature extraction using multi orientation Gabor filters. We will be using GABOR Filter, Median Filter and Repeated Line. Tracking method for recognition.

4. Conclusions

This paper will present a complete and fully automated Finger image matching and Finger subsurface features, i.e., from Finger-vein images. This will present a new algorithm for the Finger-vein identification; which can more reliably extract the Finger-vein shape features and achieve much higher accuracy than previously proposed Finger-vein identification approaches. The Finger -vein matching scheme will work more effectively in more realistic scenarios and leads to a more accurate performance; as will be demonstrated from the experimental results. At last; examine a complete and fully automated approach for the identification of low resolution Finger-surface for the performance improvement.

5. References

[1] Jain, A.; Ross, A.; Prabhakar, S. An introduction to biometric recognition. IEEE Trans. Circ. Syst. Video Tech. 2004, 14, 4-20.

[2] Jain, A.K.; Fengs, J.; Nandakumar, K. Fingerprint matching. Computer 2010, 43, 36-44.

[3] Guo, Z.; Zhang, D.; Zhang, L.; Zuo, W. Palm print verification using binary orientation co-occurrence vector. Patt. Recogn. Lett. 2009, 30, 1219-1227.

[4] Ito, K.; Nakajima, H.; Kobayashi, K.; Aoki, T.; Higuchi, T. A fingerprint matching algorithm using phase-only correlation. IEICE Trans. Fundament. Electron. Commun. Comput. Sci. 2004, E87-A, 682-691.

[5] Zhang, L.; Zhang, L.; Zhang, D.; Zhu, H. Ensemble of local and global information for finger-knuckle-print recognition. Patt. Recogn. 2011, 44, 1990-1998.

[6] Miura, N.; Nagasaka, A.; Miyatake, T. Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification. Mach. Vision Appl. 2004, 15, 194-203.

[7] Yanagawa, T.; Aoki, S.; Ohyama, T. Human finger vein images are diverse and its patterns are useful for personal identification. MHF Preprint Ser. 2007, 12, 1-7.

[8] Y. Yang and M. Levine, “The Background Primal Sketch: An Approach for Tracking Moving Objects,” Machine Vision and Applications, vol. 5, pp. 17-34, 1992.

[9] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, “Detection of blood vessels in retinal images using two-dimensional matched filters,” IEEE Trans. Med. Imag., vol. 8, no. 3, pp. 263–269, Sep. 1989.

[10] J. G. Daugman, “High confidence visual recognition of persons by a test of statistical independence,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 11, pp. 1148–1161, Nov. 1993.

[11] Y. Kuno, T. Watanabe, Y. Shimosakoda, and S. Nakagawa, “Automated Detection of Human for System.

International Journal of Computer Applications Technology and Research
Volume 3-Issue7,403-408, July 2014

A Review on “Privacy Preservation Data Mining (PPDM)

Dwipen Laskar
Assistant Professor, Dept. of CSE
Girijananda Chowdhury Institute of
Management & Technology,
Guwahati, Assam, India

laskardwipen@gmail.cm

Geetachri Lachit
Assistant Professor, Dept. of MCA
Girijananda Chowdhury Institute of
Management & Technology,
Guwahati, Assam,India

lachit.geetashri@gmail.cm

Abstract: It is often highly valuable for organizations to have their data analyzed by external agents. Data mining is a technique to analyze and extract useful information from large data sets. In the era of information society, sharing and publishing data has been a common practice for their wealth of opportunities. However, the process of data collection and data distribution may lead to disclosure of their privacy. Privacy is necessary to conceal private information before it is shared, exchanged or published. The privacy-preserving data mining (PPDM) has thus has received a significant amount of attention in the research literature in the recent years. Various methods have been proposed to achieve the expected goal. In this paper we have given a brief discussion on different dimensions of classification of privacy preservation techniques. We have also discussed different privacy preservation techniques and their advantages and disadvantages. We also discuss some of the popular data mining algorithms like association rule mining, clustering, decision tree, Bayesian network etc. used to privacy preservation technique.. We also presented few related works in this field.

Keywords: perturbation data mining, bayesian network, privacy preservation, association rule mining, clustering

1. INTRODUCTION

data mining aims to extract useful information from multiple sources, whereas privacy preservation in data mining aims to preserve these data against disclosure or loss. Privacy preserving data mining (PPDM) [1,2] is a novel research direction in data mining and statistical databases [3], where data mining algorithms are analyzed for the side-effects they incur in data privacy. The main consideration of the privacy preserving data mining is two-fold. First, sensitive raw data like identifiers, name, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person’s privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy. The main objective of privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and the private knowledge remain private even after the mining process. In this paper, we provide a classification and description of the various techniques and methodologies that have been developed in the area of privacy preserving data mining. Agarwal and Srikant [3] and Lindell and Pinkas [4] introduced the first Privacy-preserving data mining algorithms which allow parties to collaborate in the extraction of knowledge, without any party having to reveal individual items or data. The goal of this paper is to give an review of the different dimension and classification of privacy preservation techniques used in privacy preserving data mining. Also aim is to give different data mining algorithms used in PPDM and related research in this field.

2. DIMENSIONS OF PRIVACY PRESERVATION DATA MINING

Different techniques are used in privacy preserving data mining. They can be classified based on the following six dimensions [5]: *Data Mining Scenario, Data Mining Tasks, Data Distribution, Data Types, Privacy Definition, Protection Method.*

The first dimension refers to the different data mining scenarios used in privacy preservation. They are basically of two major classes presently used. In the first one organization release their data sets for data mining and allowing unrestricted access to it. Data modification is used to achieve the privacy in this scenario. In the second one organization do not release their data sets but still allow data mining tasks. Cryptographic techniques are basically used for privacy preserving

The second dimension refers to the different data mining tasks due to the data set containing various patterns. Different types of data mining tasks used are like classification, association rule mining, outlier analysis, and clustering and evolution analysis [6]. The basic need of a privacy preservation technique is to maintain data quality to support all possible data mining tasks and statistical analysis.

The third dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to the cases where different database records reside in different places, while vertical data distribution, refers to the cases where all the values for different attributes reside in different places.

The fourth dimension refers to different types of data types which are basically of three classes: Numerical and Categorical and Boolean. Boolean data are the special case of categorical data which takes two possible values 0 and 1. Numerical data has a natural ordering inherent to them but which is lacking in Categorical data. This is the most basic difference between categorical and numerical values which forces the privacy preservation technique to take different approaches for them.

The fifth dimension refers to the different definitions of privacy in different context. The definition of privacy is context dependant. In some scenario individuals data values are private, whereas in other scenario certain group, association or classification rules are private. They are basically of two classes: *Individual privacy preservation* and *Collective privacy preservation* [7]. The primary goal of Individual privacy preservation is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. The goal of the Collective privacy preservation is to protect against learning sensitive knowledge representing the activities of a group. Depend on the privacy definition we work on different privacy preserving techniques.

The sixth dimension refers to different Protection Methods: Privacy in data mining is protected through different methods such as *data modification* and *secure multiparty computation* (SMC). In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of modification include: *data perturbation, Data swapping, Aggregation, and Suppression.*

Data perturbation, which refers to a data transformation process typically performed by the data owners before publishing their data. The goal of performing such data transformation is two-fold. On one hand, the data owners want to change the data in a certain way in order to disguise the sensitive information contained in the published datasets, and on the other hand, the data owners want the transformation to best preserve those domain-specific data properties that are critical for building meaningful data mining models, thus maintaining mining task specific data utility of the published datasets. The major challenge of data perturbation is balancing privacy protection and data quality, which are normally considered as a pair of contradictive factors Two types of data Perturbation are available: *Additive Perturbation* and *matrix multiplicative Perturbation* [3][7].

In *Additive Perturbation is* a technique for in which noise is added to the data in order to mask the attribute values of records [4][7]. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions.

In the *matrix multiplicative perturbations* [7] can also be used to good effect for privacy-preserving data mining. The data owner replaces the original data X with $Y = MX$ where M is an $n' \times n$ matrix chosen to have certain useful properties. If M is orthogonal ($n' = n$n and $M^{T}M = I$), then the perturbation exactly preserves Euclidean distances, *i.e.*, for any columns x_1 ,x_2 in X ,their corresponding columns y_1 ,y_2 in Y satisfy $x_1 - x_2 = y_1 - y_2$. If each entry of M is generated independently from the same distribution with mean zero and variance σ 2 (n' not necessarily equal to n), then the perturbation approximately preserves Euclidean distances on expectation up to constant factor 2σ n' . If M is the product of a discrete cosine transformation matrix and a truncated perturbation matrix, then the perturbation approximately preserves Euclidean distances.

In *data swapping* techniques, the values across different records are interchanged in order to perform privacy preserving in data mining. One advantage of this technique is that the lower order marginal totals of the data are completely preserved and are not perturbed at all. Therefore certain kinds of aggregate computation can be exactly performed without violating the privacy of the data [8].

In *Suppression* technique sensitive data value are removed or suppressed before published. Suppression is used to protect an individual privacy from intruders attempt to accurately predict a suppressed value. Information loss is an important issue in suppression by minimizing the number of values suppressed [9][10].

Aggregation is also known as generalization or global recording. It is used for protecting an individual privacy in a released data set by perturbing the original data set before its releasing. Aggregation change k no. of records of a data by representative records. The value of an attribute in such a representative record is generally derived by taking the average of all values, for the attributes, belonging to the records that are replaced. Another method of aggregation or generalization is transformation of attribute values. For ex- an exact birth date can be changed by the year of birth. Such a generalization makes an attribute value less informatics. For ex- if exact birth date is changed by the century of birth then the released data can became useless to data miners [11].

3. DIFFERENT TECHNIQUES OF PEIVACY PRESERVING DATA MINING

Different types of privacy preservation techniques are used. They are mainly classified into following categories: [31] *Anonymization based*, *Randomized Response based*, *Condensation approach based*, *Perturbation based*, *Cryptography based* and *Elliptic Curve Cryptographic based*.

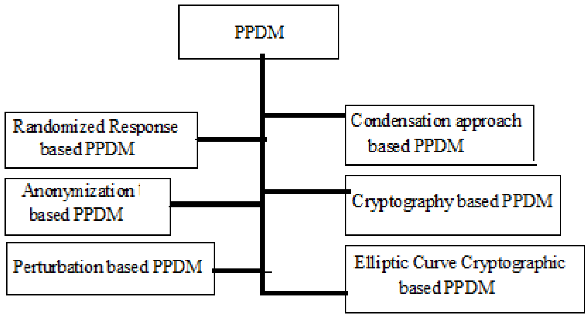
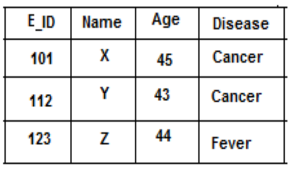

PPDM
Randomized Response based PPDM
Condensation approach based PPDM
Anonymization based PPDM
Cryptography based PPDM
Perturbation based PPDM
Elliptic Curve Cryptographic based PPDM

Fig 1: Techniques of PPDM

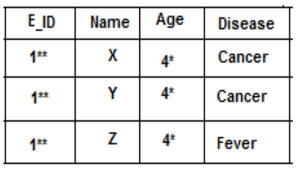
3.1 Anonymization Based PPDM

Anonymization based PPDM: Anonymization method aim is to make the individual record be indistinguishable among a group of records with using techniques of generalization and suppression [12][13][31]. Replacing a value with less specific but semantically consistent value is called as generalization and suppression involves blocking the values. K-anonymity is used to represent anonymization method. The anonymization method is ensured that after getting transformation data is true but there is some information loss in some extent. A database is *k*-anonymous with respect to quasi-identifier attributes (a set of attributes that can be used with certain external information to identify a specific individual) if there exist at least *k* transactions in the database having the same values according to the quasi-identifier attributes



E_ID	Name	Age	Disease
101	X	45	Cancer
112	Y	43	Cancer
123	Z	44	Fever

Fig: 2 (a) Original Data



E_ID	Name	Age	Disease
1**	X	4*	Cancer
1**	Y	4*	Cancer
1**	Z	4*	Fever

Fig: 2 (a) (b) K-Anonymous data

Advantages: [14]

- This method is protects identity disclosure when it is releasing sensitive information.

Disadvantages:

- It is prone to homogeneity attack and the background knowledge attack.
- Does not protect attribute disclosure to sufficient extent
- It has the limitation of *k*-anonymity model which fails in real scenario when the attackers try other methods.

3.2 Randomized Response Based PPDM

In Randomized response [14][31], the data is muddled in such a way that the central place cannot let know with probabilities better than a pre-defined threshold, whether the data from a customer contains truthful information or false information. The information received from each individual user is scrambled and if the number of users is significantly large, the aggregate information of these users can be predictable with good amount of accuracy. One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. The process of data collection in randomization method encompasses two steps [14]. In the first step, the data providers transmit the randomized data to the data receiver. In second step, reconstruction of the original distribution of the data is done by the data receiver by employing a distribution reconstruction algorithm

Advantages:

- It is a simple technique which can be easily implemented at data collection time.
- It is more efficient. However, it results in high information loss.

Disadvantages:

- It is not required for multiple attribute databases
- It results in high information loss

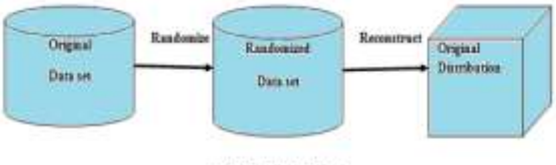


Fig 3: Randomization Technique

3.3 Condensation based PPDM

In a condensation approach, [15][31] constrained clusters are constructed in the data set, and pseudo-data from the condensed statistics of these clusters are generated. The constraints on the clusters are defined in terms of the sizes of the clusters which are chosen in such a way to preserve *k* anonymity. Some of the advantages and disadvantage of this method is [14]:

Advantages:

- This approach works with pseudo-data rather than with modifications of original data
- It is a better preservation of privacy compared to the techniques which simply use modifications of the original data.

Disadvantages:

- The pseudo-data have the same format as the original data.
- So, it is no longer necessitates the redesign of data mining algorithms

3.4 Perturbation Based PPDM

The perturbation approach the data service is not allowed to learn or recover precise records. This restriction naturally leads to some challenges. This method does not reconstruct the original data. But it can do only distributions. So, new algorithms need to be developed which use these reconstructed distributions in order to perform mining of the underlying data. This means that for each individual data problem, a new distribution based data mining algorithm needs to be developed [3]. Some of the advantages and disadvantage of this method is [14]:

Advantages:

- It is very simple technique.
- Different attributes are treated independently.

Disadvantages:

- Does not reconstruct the original vale rather than only distortion
- The perturbation approach does not provide a clear understanding of the level of indistinguishability of different records

3.5 Cryptography Based PPDM

In many cases, multiple parties may require to share private data. They want to share information without leakage at their end. For example, different branches in an educational institute wish to share their sensitive sales data to co-ordinate themselves without leaking privacy. This requires secure and cryptographic protocols for sharing the information across the different parties. Cryptography [31], in the presence of a intruder extends from the traditional tasks of encryption and authentication .In an ideal situation, in addition to the original parties there is also a third party called "trusted party". All parties send their inputs to the trusted party, who then computes the function and sends the appropriate results to the other parties. The protocol that is run in order to compute the function does not leak any unnecessary information. Sometimes there are limited leaks of information that are not dangerous. This process requires high level of trust. Some of the advantages and disadvantage of this method is [14]:

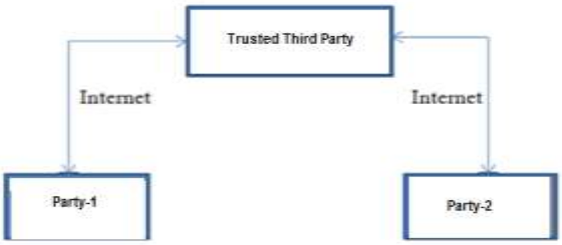


FIG 4: System using Semi Trusted Third Party

Advantages:

- Cryptography offers a well-defined model for privacy for proving and quantifying it.
- There exit a vast range of cryptographic algorithms

Disadvantages:

- It is difficult to scale when more than a few parties are involved
- It does not guarantee that the disclosure of the final data mining result may not violate the privacy of individual records.

3.6 Elliptical Curve Cryptographic Based PPDM

Elliptic Curve Cryptography (ECC) is an smart alternative to conservative public key cryptography, such as RSA. ECC are useful in the implementation on constrained devices where the major computational resources such as speed, memory is limited and low-power wireless communication protocols are used. That is because it attains the same security levels with traditional cryptosystems using smaller parameter sizes as discussed in [16][31]. Some of the advantages and disadvantage of this method is:

Advantages:

- Very few attributes are required compared to traditional cryptographic approaches

Disadvantages:

- Implementation is a complex task.

4. PRIVACY PRESERVING DATA MINING ALGORITHM

The followings are some of the data mining algorithms that have been used for privacy preservation:

4.1 Association Rule Mining

The association rule mining problem can be formally stated as follows [17]: Let I = {i_1, i_2, ---, i_m} be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that T $\subseteq$ I. Associated with each transaction is a unique identifier, called its TID. We say that a transaction T contains X, a set of some items in I, if X $\subseteq$ T. An association rule is an implication of the form, X $\Rightarrow$ Y, where X $\subset$ I, Y $\subset$ I, and X ∩ Y = φ. The rule X $\Rightarrow$ Y holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y . The rule X $\Rightarrow$ Y has support s in the transaction set D if s% of transactions in D contains X U Y. To find out if a particular itemset is frequent, it count the number of records where the values for all the attributes in the itemset are 1.

4.2 Clustering

Clustering [18] is a data mining method that has not taken its real part in the works already quoted although, the most important algorithm of this method was very studied in the context of privacy preserving, which is k-means algorithm [19]. Surveying privacy preserving k-means clustering approaches apart from other privacy preserving data mining ones is important due to the use of this algorithm in important other areas, like image and signal processing where the problem of security is strongly posed [20]. Most of works in privacy preserving clustering are developed on the k-means algorithm by applying the model of secure multi-party computation on different distributions (vertically, horizontally and arbitrary partitioned data). Among the formulations of Partition clustering based on the minimization of an objective function, k-means algorithm is the most widely used and studied. Given a dataset D of n entities (objects, data points, items,…) in real p-dimension space R^p and an integer k. The k-means clustering algorithm partitions the dataset D of entities into k-disjoint subsets, called clusters. Each cluster is represented by its center which is the centroid of all entities in that subset. The need to preserve privacy in k-means algorithm occurs when it is applied on distributed data over several sites, so called “parties” and that it wishes to do clustering on the union of their datasets. The aim is to prevent a party to see or deduce the data of another party during the execution of the algorithm. This is achieved by using secure multi-party computation that provides a formal model to preserve privacy of data.

4.3 Classification Data Mining

Classification is one of the most common applications found in the real world. The goal of classification is to build a model which can predict the value of one variable, based on the values of the other variables. For example, based on financial, criminal and travel data, one may want to classify passengers as security risks. In the financial sector, categorizing the credit risk of customers, as well as detecting fraudulent transactions is classification problems. Decision tree classification is one of the best known solution approaches. The decision tree in ID3 [21] is built top-down in a recursive fashion. In the first iteration it finds the attribute which best classifies the data considering the target class attribute. Once the attribute is identified in the given set of attributes algorithm creates a branch for each value. This process is continued until all the attributes are considered. In order to calculate which attribute is the best to classify the data set information gain is used. Information gain is defined as the expected reduction in entropy. Another most actively developed methodology in data mining is the Support Vector Machine (SVM) classification [22]. SVM has proven to be effective in many real-world applications [23]. Like other classifiers, the accuracy of an SVM classifier crucially depends on having access to the correct set of data. Data collected from different sites is useful in most cases, since it provides a better estimation of the population than the data collected at a single site.

4.4 Bayesian Data Mining

Bayesian networks are a powerful data mining tool. A Bayesian network consists of two parts: the network structure and the network parameters. Bayesian networks can be used for many tasks, such as hypothesis testing and automated scientific discovery. A Bayesian network (BN) is a graphical model that encodes probabilistic relationships among variables of interest [24].

Formally, a Bayesian network for a set V of m variables is a pair (Bs, Bp). The network structure Bs = (V,E) is a directed acyclic graph whose nodes are the set of variables. The parameters Bp describe local probability distributions associated with each variable. The graph Bs represents conditional independence assertions about variables in V: An edge between two nodes denotes direct probabilistic relationships between the corresponding variables. Together, Bs and Bp define the joint probability distribution for V.

5. RELATED WORKS

R. Agrawal, T. Imielinski, A. N. Swami [17] present a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions.

Kiran.P, Kavya N. P. [25] propose a SW-SDF based privacy preserving data classification technique. They uses sensitive weight to differentiate between sensitive attribute values.

Kiran.P, Kavya N. P. [26] proposed a method SW-SDF personal privacy for k means clustering. This algorithm groups objects in to k-clusters. Each item is placed in to the closest cluster based on the distance measures computed. In this method they propose an algorithm in such a way that the resultant clusters are almost equal to the original cluster and the privacy is retained.

J. Vaidya, C.W. Clifton, [27] proposed an association rule mining algorithm based on the Apriori algorithm. The Apriori algorithm was selected to extract the candidate set.

Vaidya J., Clifton C. [28] proposed a Privacy-Preserving k-means clustering over vertically partitioned Data based on the k-means algorithms. The k-means algorithm was selected for partitions of the clusters based on their similarity.

Vaidya J., Clifton C. [29] provides solution for privacy-preserving decision trees over vertically partitioned data.

Yu H., Vaidya J., Jiang X. [22] provides solution for privacy-preserving SVM classification on vertically partitioned data (PP-SVMV). It securely computes the global SVM model, without disclosing the data or classification information of each party to the others (*i.e.*, keeping the *model privacy* as well as the *data privacy*)

Vaidya J., Clifton C. [30] provides solution for two parties owning confidential databases to learn the Bayesian network on the combination of their databases without revealing anything else about their data to each other.

6. CONCLUSION

Classical data mining algorithms implicitly assume complete access to all data. However, privacy and security concerns often prevent sharing of data, thus devastating data mining projects. Recently, researchers have gaining more interested on finding solutions to this problem. Several algorithms have been proposed to do knowledge discovery, while providing guarantees on the non-disclosure of data. In this paper we have given a brief discussion on different dimensions of classification of privacy preservation techniques. We have also discussed different privacy preservation techniques and their advantages and disadvantages. We also discuss some of the popular data mining algorithms like association rule mining, clustering, decision tree, Bayesian network etc. used to privacy preservation technique. We also presented few related works in this field.

7. REFERENCES

[1] Chris Clifton and Donald Marks, “Security and privacy implications of data mining”, In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1996), 15–19.

[2] Daniel E. O’Leary, “Knowledge Discovery as a Threat to Database Security”, In Proceedings of the 1st International Conference on Knowledge Discovery and Databases (1991), 107–516.

[3] R. Agrawal and R. Srikant, “Privacy-preserving data mining”, In ACM SIGMOD, pages 439–450, May 2000

[4] Y. Lindell and B. Pinkas, “Privacy preserving data mining”, J. Cryptology, 15(3):177–206, 2002.

[5] M. Sharma, A. Chaudhary, M. Mathuria and S. Chaudhary, “A Review Study on the Privacy Preserving Data Mining Techniques and Approaches”, International **J**ournal of **C**omputer **S**cience and **T**elecommunications, ISSN 2047-3338, Vol.4, Issue. 9, September 2013, pp: 42-46

[6] J. Han and M. Kamber. “Data Mining Concepts and Techniques”. Morgan Kaufmann Publishers, San Diego, CA 92101-4495, USA, 2001.

[7] Xinjing Ge and Jianming Zhu (2011). “Privacy Preserving Data Mining, New Fundamental Technologies in Data Mining”, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-547-1, InTech, DOI: 10.5772/13364. Available from: http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/privacy-preserving-data-mining

[8] Divya Sharma,” A Survey on Maintaining Privacy in Data Mining”, International Journal of Engineering Research and Technology (IJERT), Vol. 1 Issue 2, April – 2012,p.26.

[9] S. Rizvi and J.R Hartisa. “Maintaining data privacy in association rule mining”. In Proc. of the 28th VLDB Conference, pages 682-693, Hong-Kong, China, 2002.

[10] Y. Saygin, V. S. Verykios and A. K. Elmagarmid. “Privacy preserving association rule mining”. In RIDE, pages 151-158, 2002.

[11] V. S. Iyenger. “Transforming data to satisfy privacy constraints”. In Proc. Of SIGKDD’02, Edmonton, Alberta, Canada, 2002.

[12] Sweeney L, "Achieving k-Anonymity privacy protection using generalization and suppression" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 571-588, 2002.

[13] Sweeney L, "k-Anonymity: A model for protecting privacy" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 557-570, 2002.

[14] Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", International Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.

[15] Charu C. Aggarwal and Philip S. Yu,(2004) “A condensation approach to privacy preserving data mining”, In EDBT, pp. 183–199.

[16] Ioannis Chatzigiannakis, Apostolos Pyrgelis, Paul G. Spirakis, Yannis C. Stamatiou “Elliptic Curve Based Zero Knowledge Proofs and Their Applicability on Resource Constrained Devices” University of Patras Greece, arXiv: 1107.1626v1 [cs.CR] 8 Jul 2011

[17] R. Agrawal, T. Imielinski, and A. N. Swami. “Mining association rules between sets of items in large database”s. In P. Buneman and S. Jajodia, editors, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207{216, Washington, D.C., May 26{28 1993}.

[18] Jain A., Murty M., and Flynn P.” Data Clustering: A Review”, ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[19] MacQueen J., “Some Methods for Classification and Analysis of Multivariate Observations,” in Proceedings of 5th Berkley Symposium Math. Statistics and Probability, California, USA, pp. 281-296, 1967.

[20] Erkin Z., Piva A., Katzenbeisser S., Lagendijk R., Shokrollahi J., Neven G., and Barni M., “Protection and Retrieval of Encrypted Multimedia Content: When Cryptography meets Signal Processing,” EURASIP Journal of Information Security, vol. 7, no. 17, pp. 1-20, 2007.

[21] Lindell Y. , Pinkas B., “Privacy Preserving Data mining*”, International Journal of Cryptology, Citesheer, 2000

[22] Yu H., Vaidya J., Jiang X.: “Privacy-Preserving SVM Classification on Vertically Partitioned Data”, PAKDD Conference, 2006.

[23] V. N. Vapnik, “Statistical Learning Theory”, John Wiley and Sons, 1998.

[24] G. Cooper and E. Herskovits, “A Bayesian Method for the Induction of Probabilistic Networks from Data,” Machine Learning, vol. 9, no. 4, pp. 309-347, 1992.

[25] Kiran.P, Kavya N. P.,”SW-SDF based privacy preserving data classification”, International Journal of Computers & Technology, Volume 4 No. 3, March-April, 2013.

[26] Kiran.P, Kavya N. P.,”SW-SDF based privacy preserving for k-means clustering”, International Journal of Scientific & Engineering Research, Volume 4, issue 6,pp. 563-566, ISSN 2229-5518, June, 2013.

[27] J. Vaidya, C.W. Clifton, “Privacy preserving association rule mining in vertically partitioned data”, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002. URL citeseer.nj.nec.com/492031. html.

[28] Vaidya J., Clifton C.: “Privacy-Preserving k-means clustering over vertically partitioned Data. ACM KDD Conference, 2003.

[29] Vaidya J., Clifton C.: “Privacy-Preserving Decision Trees over vertically partitioned data”. Lecture Notes in Computer Science, Vol 3654, 2005.

[30] Vaidya J., Clifton C. “Privacy-Preserving Naive Bayes Classifier over vertically partitioned data”. SIAM Conference, 2004.

[31] Shrivastava A., Dutta U.: “An Emblematic Study of Different Techniques in PPDM”. International Journal of Advanced Research in Computer science and Software Engineering (IJARCSSE), Vol.3, Issue.8, pp.443-447, 2013.

Review on Clustering and Data Aggregation in Wireless Sensor Network

Pooja Mann
M.Tech(Computer Science & Engineering),
Geeta Institute of Management and Technology,
Kanipla, Kurukshetra

Tarun Kumar
Deptt. of Computer Science & Engineering,
Geeta Institute of Management and Technology,
Kanipla, Kurukshetra

Abstract: Wireless Sensor Network is a collection of various sensor nodes with sensing and communication capabilities. Clustering is the process of grouping the set of objects so that the objects in the same group are similar to each other and different to objects in the other group. The main goal of Data Aggregation is to collect and aggregate the data by maintaining the energy efficiency so that the network lifetime can be increased. In this paper, I have presented a comprehensive review of various clustering routing protocols for WSN, their advantages and limitation of clustering in WSN. A brief survey of Data Aggregation Algorithm is also outlined in this paper. Finally, I summarize and conclude the paper with some future directions.

Keywords: Wireless Sensor Network, Clustering, Data Aggregation, LEACH

1. INTRODUCTION

A wireless sensor network (WSN) is an ad-hoc network composed of small sensor nodes deployed in large numbers to sense the physical world. Wireless sensor networks have very broad application prospects including both military and civilian usage. They include surveillance [1], tracking at critical facilities [2], or monitoring animal habitats [3].

In general, a WSN consists of a large number of tiny sensor nodes distributed over a large area with one or more powerful sinks or base stations (BSs) collecting information from these sensor nodes. All sensor nodes have limited power supply and have the capabilities of information sensing, data processing and wireless communication [4].

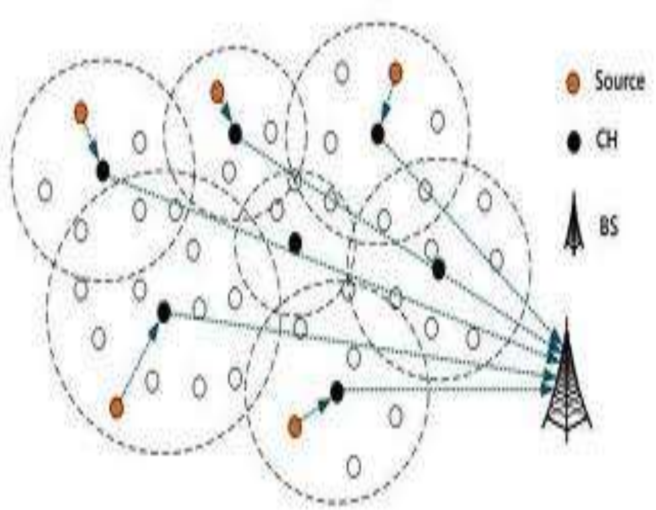


Figure 1 Sensor Network Architecture

WSN has various characteristics like Ad Hoc deployment, Dynamic network topology, Energy Constrained operation, Shared bandwidth, large scale of deployment. Despite of these characteristics routing in WSN is more challenging. Firstly, resources are greatly constrained in terms of power supply, processing capability and transmission bandwidth. Secondly, it is difficult to design a global addressing scheme as Internet Protocol (IP). Furthermore, IP cannot be applied to WSNs, since address updating in a large-scale or dynamic WSN can result in heavy overhead. Thirdly, due to the limited resources, it is hard for routing to cope with unpredictable and frequent topology changes, especially in a mobile environment. Fourthly, data collection by many sensor nodes usually results in a high probability of data redundancy, which must be considered by routing protocols. Fifthly, most applications of WSNs require the only communication scheme of many-to-one, *i.e.*, from multiple sources to one particular sink, rather than multicast or peer to peer. Finally, in time-constrained applications of WSNs, data transmissions should be accomplished within a certain period of time. Thus, bounded latency for data transmissions must be taken into consideration in this kind of applications.

Based on network structure, routing protocols in WSNs can be coarsely divided into two categories: flat routing and hierarchical routing. In a flat topology, all nodes perform the same tasks and have the same functionalities in the network. Data transmission is performed hop by hop usually using the form of flooding. In small-scale networks flat routing protocols are relatively effective. However, it is relatively undesirable in large-scale networks because resources are limited, but all sensor nodes generate more data processing and bandwidth usage. On the other hand, in a hierarchical topology, nodes perform different tasks in WSNs and typically are organized into lots of clusters according to specific requirements or metrics. Generally, each cluster comprises a leader referred to as cluster head (CH) and other member nodes (MNs) or ordinary nodes (ONs), and the CHs can be organized into further hierarchical levels. In general, nodes with higher energy act as CH and perform the task of data processing and information transmission, while nodes with low energy act as MNs and perform the task of information sensing.

2. CLUSTERING

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A good clustering algorithm is able to identify clusters irrespective of their shapes. Other requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitivity to the order of input records, etc [5]. In the hierarchical network structure each cluster has a leader, which is also called the cluster head (CH) and usually performs the special tasks referred above (fusion and aggregation), and several common sensor nodes (SN) as members. The cluster formation process eventually leads to a two-level hierarchy where the CH nodes form the higher level and the cluster-member nodes form the lower level. The sensor nodes periodically transmit their data to the corresponding CH nodes. The CH nodes aggregate the data and transmit them to the base station (BS) either directly or through the intermediate communication with other CH nodes. However, because the CH nodes send all the time data to higher distances than the common (member) nodes, they naturally spend energy at higher rates. A common solution in order to balance the energy consumption among all the network nodes is to periodically re-elect new CHs each cluster. CH nodes aggregate the data and transmit them to the base station (BS) either directly or through the intermediate communication with other CH nodes. A typical example of the implied hierarchical data communication within a clustered network is illustrated in Figure 2.

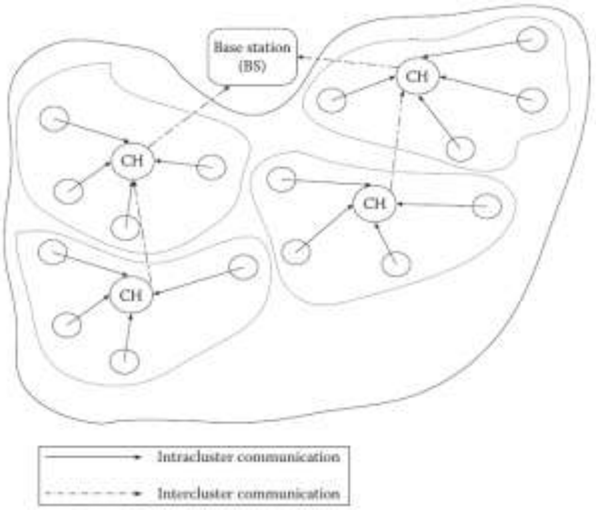


Figure 2 Data Communication in Clustered Network

2.1 Classification of Clustering

- **One Hop Model**

This is the simplest approach and represents direct communication. In these networks every node transmits to the base station directly. This communication implies not only to be too expensive in terms of energy consumption, but it is also infeasible because nodes have limited transmission range [6],[7],[8]. Most of the nodes in networks with large area coverage usually are far enough thus their transmissions cannot reach the base station. Direct communication is not a feasible model for routing in WSN.

- **Multi-hop Planar Model**

In this model, a node transmits to the base station by forwarding its data to one of its neighbors, which is closer to the base station. The latter passes on it to neighbors that is even closer to the base station. Thereby the information travels from source to destination by hop from one node to another until it reaches the destination. Regarding to the energy and transmission range node limitations, this model is a viable approach. A number of protocols employ this approach like [9][10][11][12], and some use other optimization techniques to enhance the efficiency of this model. One of these techniques is data aggregation used in all clustering-based routing protocol, for instance in [13] and [14]. Even though these optimization techniques improve the performance of this model, it is still a planar model. In a network composed by thousands of sensors, this model will exhibit high data dissemination latency due to the long time needed by the node information to arrive to the base station [15], [16].

- **Clustering-based Hierarchical Model**

A hierarchical approach for the network topology breaks the network into several areas called clusters as shown in figure 3. Nodes are grouped depending on some parameter into clusters with a cluster head, which has the responsibility of routing the data from the cluster to other cluster heads or base stations. Data travels from a lower clustered layer to a higher one. Data still hops from one node to another, but since it hops from one layer to another it covers larger distances and moves the data faster to the base station than in the multi hop model [17],[18],[19],[20].

The latency in this model is theoretically much less than in the multi-hop model. Clustering provides inherent optimization capabilities at the cluster heads, what results in a more efficient and well structured network topology. This model is more suitable than one hop or multi hop model. The multi-hop model is a more practical approach than in one hop. In this case, data is forwarded by hops from one node to another until it reaches the base station.

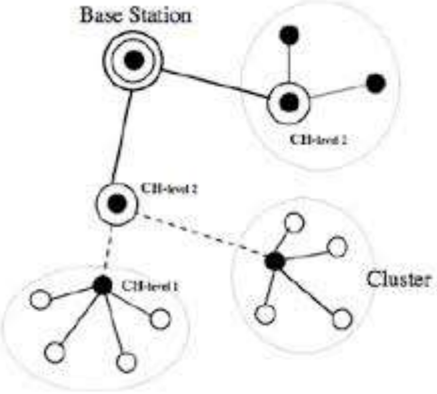


Figure 3 Hierarchical clustering-based Model

Some drawbacks of this model are the high latency in networks comprised of thousands of sensors and the serious delay that data experiences. Perhaps the most important drawback is that the closest nodes to the base station would have to act as intermediaries to all traffic being sent to the base station by the rest of the network.

3. DATA AGGREGATION

Data aggregation is a process of aggregating the sensor data using aggregation approaches. The general data aggregation algorithm works as shown in fig 4. The algorithm uses the sensor data from the sensor node and then aggregates the data by using some aggregation algorithms such as centralized approach, LEACH(low energy adaptive clustering hierarchy),TAG(Tiny Aggregation) etc. This aggregated data is transfer to the sink node by selecting the efficient path [21].

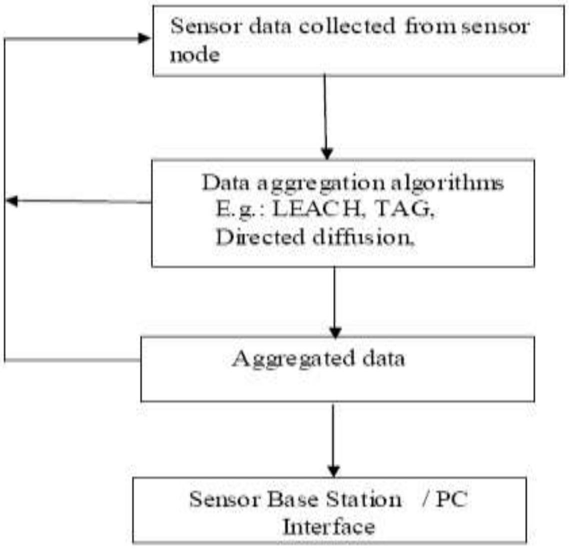

Sensor data collected from sensor node
Data aggregation algorithms
E.g.: LEACH, TAG,
Directed diffusion.
Aggregated data
Sensor Base Station / PC Interface

Figure 4 General architecture of the data aggregation algorithm

Data aggregation, which is the process of aggregating the data from multiple nodes to eliminate redundant transmission and provide fused data to BS, is considered as an effectual technique for WSNs to save energy. The most popular data aggregation algorithms are cluster-based data aggregation algorithms, in which the nodes are grouped into clusters and each cluster consists of a cluster head (CH) and some members, each member transmits data to its CH, then, each CH aggregates the collected data and transmits the fused data to BS. The cluster-based WSNs have an inherent problem of unbalanced energy dissipation. Some nodes drain their energy faster than others and result in earlier failure of network. Some researchers have studied this problem and proposed their algorithms which have both advantages and disadvantages. Our motivation is to propose a novel solution to this problem in the cluster-based and homogeneous WSNs, in which the CHs transmit data to BS by one-hop communication, with an objective of balancing energy consumption by an energy efficient way and prolonging network lifetime.

4. DATA AGGREGATION PROTOCOLS BASED ON NETWORK ARCHITECTURE

4.1 Flat Networks

In flat networks, each sensor node plays the same role and is equipped with approximately the same battery power. In such networks, data aggregation is accomplished by data centric routing where the sink usually transmits a query message to the sensors, e.g, via flooding and sensors which have data matching the query send response messages back to the sink. The choice of a particular communication protocol depends on the specific application at hand.

4.1.1 Push Diffusion

In the push diffusion scheme, the sources are active participants and initiate the diffusion while the sinks respond to the sources. The sources flood the data when they detect an event while the sinks subscribe to the sources through enforcements. The *sensor protocol for information via negotiation* (SPIN) [22] can be classified as a push based diffusion protocol.

4.1.2 Two Phase Pull Diffusion

Directed diffusion is a representative approach of two phase pull diffusion. It is a data centric routing scheme which is based on the data acquired at the sensors. The attributes of the data are utilized message in the network. Figure 5 illustrates the interest propagation in directed diffusion. If the attributes of the data generated by the source match the interest, a gradient is set up to identify the data generated by the sensor nodes. The sink initially broadcasts an interest message in the network. The gradient specifies the data rate and the direction in which to send the data. Intermediate nodes are capable of caching and transforming the data. Each node maintains a data cache which keeps track of recently seen data items. After receiving low data rate events, the sink reinforces one particular neighbor in order to attract higher quality data. Thus, directed diffusion is achieved by using data driven local rules.

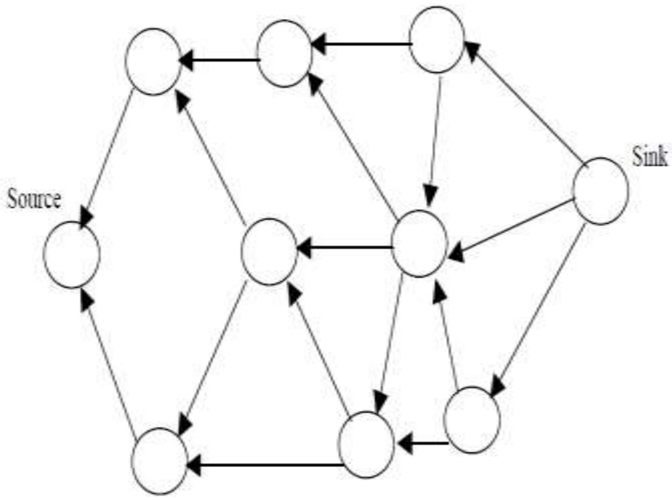


Figure 5 Interest propagation in directed diffusion

4.1.3 One Phase Pull Diffusion

Two phase pull diffusion results in large overhead if there are many sources and sinks. Krishnamachari et al. [23] have proposed a one phase pull diffusion scheme which skips the flooding process of directed diffusion. In one phase pull diffusion, sinks send interest messages that propagate through the network establishing gradients. However, the sources do not transmit exploratory data. The sources transmit data only to the lowest latency gradient pertinent to each sink. Hence, the reverse route (from the source to the sink) has the least latency. Removal of exploratory data transmission results in a decrease in control overhead conserving the energy of the sensors.

4.2 Hierarchical Networks

A flat network can result in excessive communication and computation burden at the sink node resulting in a faster depletion of its battery power. The death of the sink node breaks down the functionality of the network. Hence, in view of scalability and energy efficiency, several hierarchical data aggregation approaches have been proposed. Hierarchical data aggregation involves data fusion at special nodes, which reduces the number of messages transmitted to the sink. This improves the energy efficiency of the network.

4.2.1 Data Aggregation in Cluster based networks

In energy constrained sensor networks of large size, it is inefficient for sensors to transmit the data directly to the sink. In such scenarios, sensors can transmit data to a local aggregator or cluster head which aggregates data from all the sensors in its cluster and transmits the concise digest to the sink. This results in significant energy savings for the energy constrained sensors. Figure 6 shows a cluster based sensor network organization. The cluster heads can communicate with the sink directly via long range transmissions or multi hopping through other cluster heads.

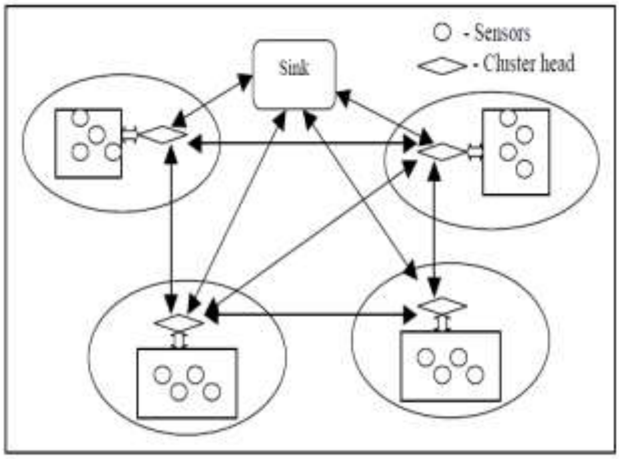


Figure 6 Cluster based Network

4.2.2 Chain based Data Aggregation

In cluster-based sensor networks, sensors transmit data to the cluster head where data aggregation is performed. However, if the cluster head is far away from the sensors, they might expend excessive energy in communication. Further improvements in energy efficiency can be obtained if sensors transmit only to close neighbors. The key idea behind chain based data aggregation is that each sensor transmits only to its closest neighbor. Lindsey et al. [24] presented a chain based data aggregation protocol called power efficient data gathering protocol for sensor information systems (PEGASIS). In PEGASIS, nodes are organized into a linear chain for data aggregation. The nodes can form a chain by employing a greedy algorithm or the sink can determine the chain in a centralized manner. Greedy chain formation assumes that all nodes have global knowledge of the network. The farthest node from the sink initiates chain formation and at each step, the closest neighbor of a node is selected as its successor in the chain. In each data gathering round, a node receives data from one of its neighbors, fuses the data with its own and transmits the fused data to its other neighbor along the chain. Eventually the leader node which is similar to cluster head transmits the aggregated data to the sink. Figure 7 shows the chain based data aggregation procedure in PEGASIS.

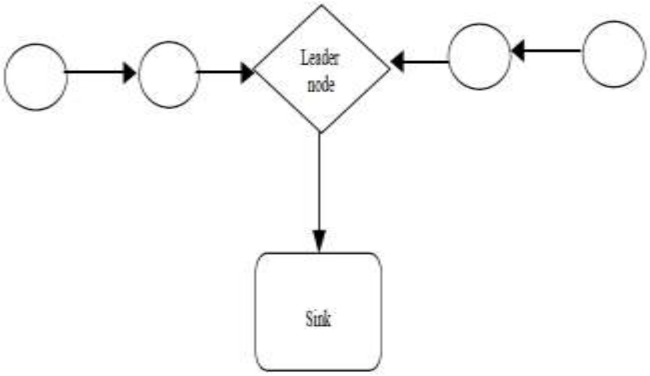


Figure 7 Chain based organization in a sensor network

The PEGASIS protocol has considerable energy savings compared to LEACH. The distances that most of the nodes transmit are much less compared to LEACH in which each node transmits to its cluster head. The leader node receives at most two data packets from its two neighbors. In contrast, a cluster head in LEACH has to perform data fusion of several data packets received from its cluster members. The main disadvantage of PEGASIS is the necessity of global knowledge of all node positions to pick suitable neighbors and minimize the maximum neighbor distance.

4.2.3 Tree based Data Aggregation

In a tree based network, sensor nodes are organized into a tree where data aggregation is performed at intermediate nodes along the tree and a concise representation of the data is transmitted to the root node. Tree based data aggregation is suitable for applications which involve in-network data aggregation. An example application is radiation level monitoring in a nuclear plant where the maximum value provides the most useful information for the safety of the plant. One of the main aspects of tree-based networks is the construction of an energy efficient data aggregation tree.

4.2.4 Grid based Data Aggregation

In grid-based data aggregation, a set of sensors is assigned as data aggregators in fixed regions of the sensor network. The sensors in a particular grid transmit the data directly to the data aggregator of that grid. Hence, the sensors within a grid do not communicate with each other. In this aggregation, the data aggregator is fixed in each grid and it aggregates the data from all the sensors within the grid. This is similar to cluster based data aggregation in which the cluster heads are fixed. Grid-based data aggregation is suitable for mobile environments such as military surveillance and weather forecasting and adapts to dynamic changes in the network and event mobility. Figure 8 shows that in grid based data aggregation, all sensors directly transmit data to a predetermined grid aggregator.

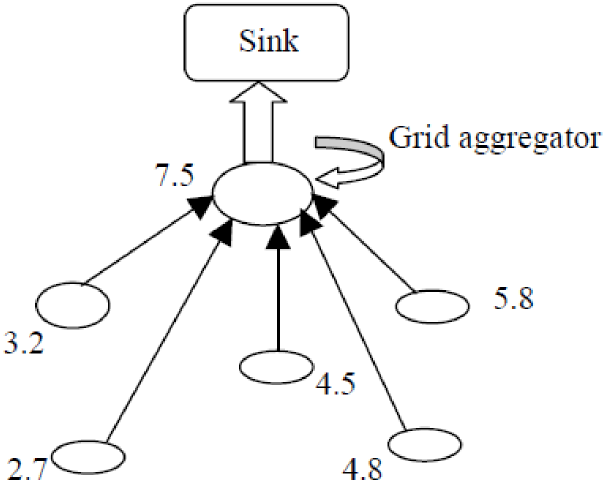


Figure 8 Grid based Data Aggregation

5. LEACH PROTOCOL

LEACH performs local data fusion to “compress” the amount of data being sent from the clusters to the base station, further reducing energy dissipation and enhancing system lifetime. Sensors elect themselves to be local cluster-heads at any given time with a certain probability. These cluster head nodes broadcast their status to the other sensors in the network. Each sensor node determines to which cluster it wants to belong by choosing the cluster-head that requires the minimum communication energy. Once all the nodes are organized into clusters, each cluster-head creates a schedule for the nodes in its cluster. This allows the radio components of each non-cluster-head node to be turned off at all times except during its transmit time, thus minimizing the energy dissipated in the individual sensors. Once the cluster-head has all the data from the nodes in its cluster, the cluster-head node aggregates the data and then transmits the compressed data to the base station.[25]
LEACH is self adaptive and self-organized. This protocol uses round as unit, each round is made up of cluster set-up stage and steady-state stage, for the purpose of reducing unnecessary energy costs, the steady state stage must be much longer than the set-up stage. The process of it is shown in Figure 9.

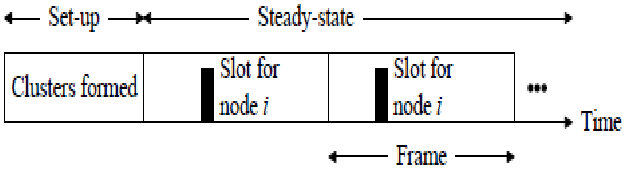


Fig.9 LEACH Protocol process

At the stage of cluster forming, a node randomly picks a number between 0 to 1, compared this number to the threshold values ***t***(***n***) , if the number is less than ***t***(***n***) , then it become cluster head in this round, else it become common node. Threshold ***t***(***n***) is determined by the following:

$$\mathbf{t}(n) = \begin{cases} \dfrac{p}{1 - p * (r \bmod \dfrac{1}{p})} & \text{if } n \in G \\ 0 & \text{if } n \notin G \end{cases}$$

Where p is the percentage of the cluster head nodes in all nodes, r is the number of the rounds, G is the collections of the nodes that have not yet been head nodes in the first 1/P rounds. Using this threshold, all nodes will be able to be head nodes after 1/P rounds.[26]

6. ACKNOWLEDGEMENT

I would like to thanks the faculty member of the Computer Science & Engineering Department of Geeta Institute of Engineering and Techology, kanipla (District Kurukshetra).

7. REFERENCES

[1] D. Culler, D. Estrin, and M. Srivastava, “Overview of Sensor Networks,” *IEEE Computer*, August 2004.

[2] N. Xu, S. Rangwala, K. Chintalapudi, D. Ganesan, A. Broad, R. Govindan, and D. Estrin, "A Wireless Sensor Network for Structural Monitoring," *Proceedings of the ACM Conference on Embedded Networked Sensor Systems, Baltimore, MD*, November 2004.

[3] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "Wireless Sensor Networks for Habitat Monitoring," *WSNA'02, Atlanta,Georgia*, September 2002.

[4] Xuxun Liu, "A Survey on Clustering Routing Protocols in Wireless Sensor Networks,"*Sensors* 2012, *12*, 11113-11153; doi:10.3390/s120811113, 9 August 2012.

[5] Amandeep Kaur Mann, Navneet Kaur, "Survey Paper on Clustering Techniques," International Journal of Science, Engineering and Technology Research (IJSETR) Vol 2, Issue 4, April 2013

[6] S. Meguerdichian, S.Slijepcevic, V.Karayan and M.Potkonjak, "Localized Algorithms in Wireless Ad-Hoc Networks: Location Discovery and sensor Exposure," in Proc. Of Mob adhoc, Long Beach, CA, USA, pp. 106-116, 2001.

[7] Wang Wei, "Study on Low Energy Grade Routing Protocols of Wireless Sensor Network," Dissertation, Hang Zhou, Zhe Jiang University, 2006.

[8] Wei Bo, Hu Han-ying, Fu Wen, "A pseudo LEACH algorithm for Wireless Sensor Network" in Proc. IAENG, March 2007.

[9] Fan Xiangning, "Improvement on LEACH Protocol on Wireless Sensor Network," mt. Conference on Sensor Technologies and Applications, 7 July, 2007.

[10] Haiming Yang, Biplab Sikdar, "Optimal Cluster Head Selection in the LEACH Architecture," Performance, Computing and communications Conference, Int. 2, 2007.

[11] Haosong Gou, "An Energy Balancing LEACH Algorithm for Wireless Sensor Networks," 7th Conference on Information Technology, 3 October, 2010.

[12] Hu Junping, "A Time-based Cluster-Head Selection Algorithm for LEACH", IEEE, 1 August, 2008.

[13] A. Manjeshwar and D.Agrawal, "TEEN: A Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks," in Proc. 1st Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, San Francisco, CA, pp. 2009-2015, April 2001.

[14] Bilal Abu Bakr, "Extending Wireless Sensor Network Lifetime in the LEACH-SM Protocol by Spare Selection," 5th Int. Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, July,2011.

[15] W.R. Heizelman, A.Chandrakasan and H.Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks", Proc. 33rd Hawaii Int. Conference on System Science, Vol. 2, 4-7 Jan,2000.

[16] Ye W, Heidenman J Esrtrin D, "An Energy-Efficient MAC Protocol for Wireless Sensor Network," in proc. IEEE INFOCOM, http://www.isi.edu/div7/publication files/yeO2a.pdf , 2002.

[17] M. Dong, K. Yung and W. Kaiser, "Low Power Signal Processing Architectures for Network Microsensors," in Proc. Int. Symposium on Low Power Electronics and Design, pp 173-177, Aug, 1997.

[18] Mo Xiaoyan, "Study and Design on Cluster Routing Protocols of Wireless Sensor Networks," Dissertation, hang Zhou, Zhe Jiang University, 2006.

[19] Mu Tong, "LEACH-B: An Improved LEACH Protocol for Wireless Sensor Network," IEEE, 2010.

[20] O. Younis and S. Fahmy, "HEED: A Hybrid,Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks," Trans. Mobile Computing, Vol. 3, No. 4, pp 336-379, Oct-Dec, 2004

[21] Nandini. S. Patil, Prof. P. R. Patil, "Data Aggregation in Wireless Sensor Network," IEEE International Conference on Computational Intelligence and Computing Research,2010

[22] J. Kulik, W.R. Heinzelman and H. Balakrishnan, "Negotiation-based protocols for disseminating information in wireless sensor networks," *Wireless Networks*, vol. 8, March 2002, pp. 169-185.

[23] B. Krishnamachari and J. Heidemann, "Application specific modeling of information routing in wireless sensor networks", *Proc. IEEE international performance, computing and communications conference,* vol. 23, pp. 717-722, 2004.

[24] S. Lindsey, C. Raghavendra, and K.M. Sivalingam, "Data gathering algorithms in sensor networks using energy metrics," *IEEE Trans. Parallel and Distributed Systems*, vol. 13, no. 9, September 2002, pp. 924-935.

[25] Wendi Rabiner Heinzelman, Anantha Chandrakasan, and Hari Balakrishnan, "Energy efficient Communication protocol for Wireless Microsensor Networks," Published in the Proceedings of the Hawaii International Conference on System Sciences, January 4-7, 2000, Maui, Hawaii.

[26] Chunyao FU, Zhifang JIANG, Wei WEI2and Ang WEI, "An Energy Balanced Algorithm of LEACH Protocol in WSN," IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 1, January 2013.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 415 - 419, 2014, ISSN: 2319–8656

RED: A HIGH LINK UTILIZATION AND FAIR ALGORITHM

Neha
Department of CE
UCOE, Punjabi University Patiala
Punjab, India

Abhinav Bhandari
Department of CE
UCOE Punjabi University Patiala
Punjab, India

Abstract— Internet and its applications are an integral part of our daily life .These days they are widely used for various purposes such as communication, public services, entertainments, distant educations, etc., each possessing different quality of service (QoS) requirements. How to provide finer congestion control for network emerges as a major problem. To prevent the problem of congestion control and synchronization various active queue management (AQM) techniques are used. AQM algorithms execute on network routers and detect initial congestion by monitoring some functions. When congestion occurs on the link the AQM algorithms detects and provides signals to the end systems. Various algorithms have been proposed in recent years but RED is one of the most influential techniques among all the existing ones. This review paper provides the functioning mechanism of the RED technique with the help of its algorithm & its variants.

Keywords- AQM ,RED ,RED parameters, RED algorithm, RED variants

1 . INTRODUCTION

The Internet intended for openness and scalability in its beginning. In Internet congestion occurs when the total demand for a resource (e.g. link bandwidth) exceeds the existing capacity of the resource. It is necessary to stay away from high packet loss rates in the Internet. In communication networks, available bandwidth and network routers plays an most important role during transmission of data packets. As a result, the routers have to be designed in such a way that they can survive large queues in their buffers in order to accommodate for transient congestion. Earlier the congestion in the network is detected after the packet has been dropped. When a packet is dropped before it reaches its destination, all of the resources it has consumed in transit are wasted. In extreme cases, this situation can lead to congestion collapse [4].But with the time to meet the demands of the network and for providing better throughput various active queue management techniques come into existence. The basic idea of active queue management is used to detect congestion as well as to control the queue size before the buffer overflows. When transmission control protocol (TCP) detects packet losses in the network TCP decreases packet sending rate .since the congestion at a router is relieved due to this decrease of the packet sending rate of TCP, AQM mechanism helps to prevents Buffer flow. AQM mechanisms control the queue length (i.e., the number of packets in the router's buffer) by actively discarding arriving packets before the router's buffer becomes full. [11] Figure 1 shows the AQM mechanism. To solve the problem of congestion in internet the Internet Engineering Task Force (IETF) has suggested the use of RED [2], [3], an active queue management scheme which is able to achieve high throughput and low average delay (for TCP traffic) by spreading randomly packets drops between flows.

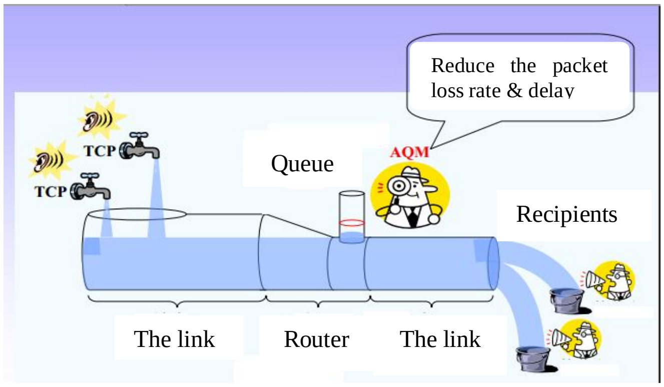


Fig.1. Active Queue Management Mechanism

Among various AQM scheme, Random Early Detection (RED) [8] is probably the most extensively studied. RED is shown to effectively tackle both the global synchronization problem and the problem of bias against bursty sources. Due to its popularity, RED or its variants has been implemented by many router vendors in their products.

This paper is organized as follow; section 1 gives the introduction about congestion control with AQM technique (RED) Section 2 gives a view about the RED technique. Section 3 describes the RED algorithm with its parameters and diagrammatically explains the working of RED algorithm .Section 4 gives a view about the different variants of RED. Section 5 Conclusion & Section 6 References.

2. RED (RANDOM EARLY DETECTION)

Random Early Detection (RED) was first proposed AQM mechanism and is also promoted by Internet Engineering Task

Force (IETF) as in [2]. Random Early Detection (RED) was introduced in 1993 [3] by Floyd and Jacobson. RED provides congestion avoidance by controlling the queue size at the gateway [3]. RED is customized for TCP connection across IP routers it's considered to avoid congestion It notifies the cause before the congestion truly happens rather than wait till it actually occurs. It provides a method for the gateway to provide some feedback to the resource on congestion status. In order to solve the problem of passive queue management technique this section proposes and evaluates a primarily different queue management algorithm called RED. RED is designed to be used in conjunction with TCP, which currently detects congestion by means of timeouts (or some other means of detecting packet loss such as duplicate ACKs) [10]. RED has been designed with the objective to:

- reduce packet loss and queuing wait,
- Avoid global synchronization of sources
- Maintain high link utilization, and
- Remove biases against bursty sources

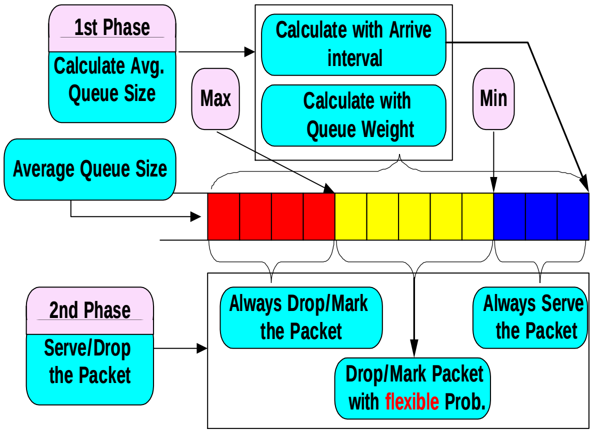

1st Phase
Calculate Avg. Queue Size
Calculate with Arrive interval
Calculate with Queue Weight
Max
Min
Average Queue Size
2nd Phase
Serve/Drop the Packet
Always Drop/Mark the Packet
Always Serve the Packet
Drop/Mark Packet with flexible Prob.

Fig.2. RED mechanism

Fig 2 represents that RED monitors the average queue size and marks packets. If the buffer is almost empty, all incoming packets are accepted. As the queue grows, the probability for dropping an incoming packet grows too. When the buffer is full, all incoming packets are dropped. Thus RED buffer mechanism works with constant bit rate (CBR) traffic can be used at an early stage to know the effect of change of network parameters over system performance. The main aim of RED is to control the queue size and indicating the end hosts when to slow down their packet transmission rate. It takes benefit of the congestion control mechanism of TCP by randomly dropping packets earlier to periods of high congestion, RED tells the packet source to reduce its transmission rate. Assuming the packet source is using TCP, it will reduce its transmission rate until all the packets reach their destination, representing that the congestion is cleared.

3. RED ALGORITHM

The RED algorithm is congestion avoidance scheme used in communication network to keep away from congestion. Compared to other algorithms this avoids congestion at common network bottlenecks, where the system triggers before any congestion actually occurs. RED performance is highly sensitive to its parameter settings ([10]). Table1 represents the various parameters included in RED algorithm.

Table1: Parameters of RED algorithm

Parameter	Meaning
Maxth	Maximum threshold
Minth	Minimum threshold
Maxp	Maximum packet Dropping/Marking Probability
Wq	Weighting factor
P_a	Probability
Avg	Average queue length

RED algorithm differentiate between the temporary and persistent congestion in the network, so as to propose the network to hold busty traffic, rather than shaping busty traffic, in which the average queue length is calculated for each packet arrival. RED has two types of queue length threshold activities: Minth and Maxth. As the packet arrives at a router, RED calculates the new average queue size Avg and compares it with the these two thresholds, and then takes actions according to the given rules: if the average queue length is smaller than the minimum threshold (Avg < Minth), no action is taken; if the average queue length is larger than the maximum threshold (Avg > Maxth), the packet is always dropped; if the average queue length lies between the two thresholds , the newly arriving packet is dropped with some probability(P_a) as shown in figure 3:

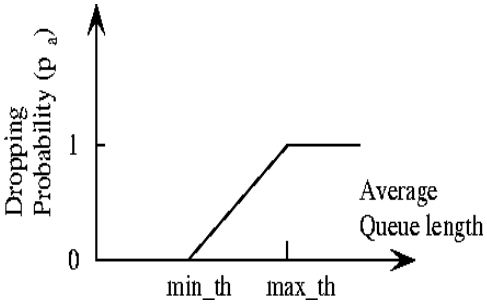


Fig.3 Droping probability in RED

The droping probability of RED algorithm Can be represented as shown in the Figure3. The general RED algorithm can be presented as follows in the figure 4 and the parameters used in the algorithm are defined in Table 1

For each new arrival of packet:
Compute the average queue length;

```
 If  (  Minth ≤ Avg < Maxth)
 {
   Calculate the  probability Pa,
    with probability Pa: mark/drop the arriving packet
 }
 Else if (Maxth ≤  Avg)
 {
   mark/drop the arriving packet
 }
 Else
{
 Do not mark/ drop the packet
 }
```

Fig.4. General Algorithm of RED

The average queue size at arrival of a new packet can be calculated by using the formula i.e.

Avg = (1 – weight) X Avg + weight x currQ

Where   0 < Weight < 1

currQ is the current queue length

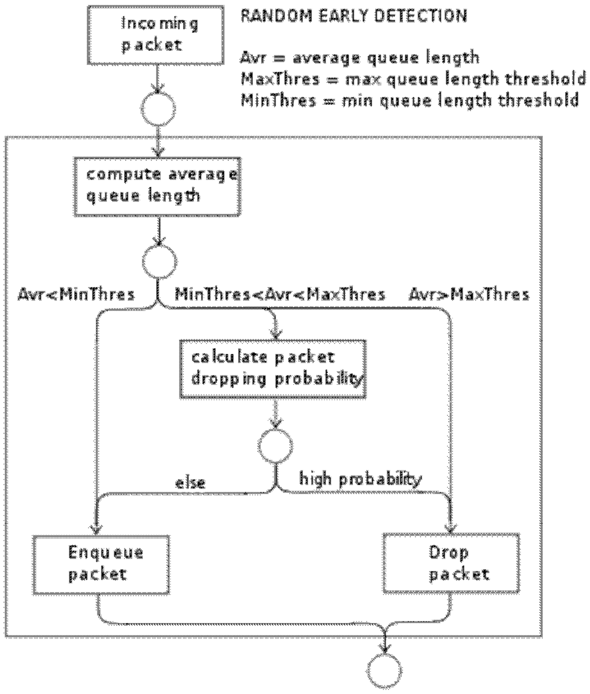

Incoming packet
RANDOM EARLY DETECTION
Avr = average queue length
MaxThres = max queue length threshold
MinThres = min queue length threshold
compute average queue length
Avr<MinThres
MinThres<Avr<MaxThres
Avr>MaxThres
calculate packet dropping probability
else
high probability
Enqueue packet
Drop packet

Fig.5. Working of RED algorithm

Random early detection is a queue management scheme that is proposed to respond the shortcomings of drop tail. RED perfectly notifies one of sources of congestion by randomly dropping an arriving packet. The selected source is informed by the packet loss and its sending rate is reduced accordingly. Therefore, congestion is alleviated. It is an early congestion declaration. The dropping probability is a function of average queue length. When the queue tenure grows, congestion builds up. Then, the dropping probability increases in order to supply enough early congestion notifications another goal of RED is to eradicate biases against busty sources in the network. This is done by limiting the queue use so that there is always room left in the queue to buffer transient bursts. In addition, the marking purpose of RED takes into account the last packet marking time in its calculations in order to reduce the probability that successive packets belonging to the same burst are marked. Fig.5 represents the working flowchart for RED algorithm.

4. VARIANTS OF RED

Many variants of RED have been proposed in the past out of all few are briefly explained in the following section. Random Early Detection (RED) was first proposed AQM mechanism and is also promoted by Internet Engineering Task Force (IETF) as in [2]. Random Early Detection (RED) was introduced in 1993 [3] by Floyd and Jacobson and then many variants were also proposed. It is a representative AQM mechanism, have been tenderly studied by numerous researchers. When a packet arrives at a router, the RED router calculates the average queue length. The RED router drops the arriving packet with the probability which is calculated from the average queue length and a configuration of control parameters. If the average queue length becomes large, the change of the packet drop probability of RED will become unstable.

To overcome the drawbacks of RED these different variants comes into existence. All variants depends on RED parameters i.e. average queue length, minimum threshold, maximum threshold and dropping probability in dealing with congestion and achieving the maximum Quality of service .It provide a mechanism to control the congestion collapse in the internet.Following is the brief description of variants of RED that are widely studied to overcome the weaknesses of the general RED algorithm.

4.1 RED Flow Trust (REDFT)

It is a scalable AQM scheme, based on flow trust .It could be used to counter low rate DoS attacks as well as Flooding rate DoS attacks. In this each flow behaviors' are monitored and according to the data collected trust values are calculated.

4.2 Non Linear RED (NLRED)

NLRED is the same as the original RED except that the linear packet dropping probability function is replaced by a nonlinear quadratic function. While inheriting the simplicity of RED, NLRED was shown to outperform RED. In particular, NLRED is less sensitive to parameter settings, has a more predictable average queue size, and can achieve a higher throughput.

4.3 Hybrid RED (HRED)

The Hybrid RED algorithm will effectively change the transfer function of the overall control loop and thus the stability of network. Hybrid RED algorithm gives a better loss rate and link utilization compared to the existing RED and LPF/ODA algorithms. Two modifications to RED are proposed: (i) use of both instantaneous queue size and its Exponential Weighted Moving Average (EWMA) for packet marking/dropping and (ii) reducing the effect of the EWMA queue size value when the queue size is less than minth for a certain number of consecutive packet arrivals.

4.4 Flow Random Early Drop (FRED)

It introduces "per-active-flow accounting" to enforce a drop rate on each flow that is dependent on the flow's buffer occupancy. FRED has two parameters MINq(i) and MAXq(i) which are the minimum and maximum numbers of packets that each flow (i) is allowed to buffer. FRED uses a global variable (avgcq) to calculate approximately the average per-active flow buffer usage. It maintains the number of active flows for each of them, FRED maintains count of buffer packets and a count of times when the flow is responsive (Qlen>MAXq).FRED requires a certain amount of overhead for per active flow counting.

Following is the tabular representation of variants of RED as shown below in table 2

Table 2 Variants of RED

Sr.no	Name of the Variant	Year	Aut-hor	Advantages	Remarks
1	RED-FT(flow trust)[7]	2013	Jiang et.al	It employs the flow trust to safeguard legitimate flows. It improves the throughput and delay in DoS attacking scenarios.	Potential work is required to enhance the detection accuracy & propagation of flow trust in networks
2	NLRED(Non linear RED) [13]	2009	Wang et.al	In this scheme packet dropping becomes gentler than RED at light traffic load. NLRED achieves a higher and more stable throughput than RED.	Packet dropping becomes more aggressive at heavy load
3	HRED (Hybrid RED)[9]	2008	Haider et.al	HRED shows better utilization of network bandwidth and a lower packet loss rate	Hybrid RED under different traffic mixes such as exponential ON/OFF traffic over UDP with TCP, is still uncovered.
4	FRED(Flow Random Early Drop)[5]	1997	D.Lin and R. Morris	It protects from fragile flows and maintain high degree of fairness by using per active flow accounting.	It requires a certain amount of overhead for per active flow counting.

5.CONCLUSION

RED is the most known active queue management mechanism that is widely studied by many researchers. The best benefit of using RED is instead of considering instant queue length RED uses average queue length to drop a packet. The performance of RED may get worse in many situations but the different variants of RED help to avoid the problems that are faced by the general RED algorithm. Many refinements to fundamental algorithm make it more adaptive and require less variation.

6. REFERENCES

[1]V. Misra, W. B. Gong, and D. Towsley, "Fluid-based analysis of a network of AQM routers supporting TCP flows with an application to RED," ACM SIGCOMM Computer Communication Review, Vol. 30, pp. 151-160, 2000.

[2] B. Braden, D. Clark, J. Crow croft, B. Davie, S. Deering, D. Estrin, S. Floyd, V. Jacobson, G. Minshall, C. Partridge, L. Peterson, K. Ramakrishnan, S. Shenker, J. Wroclaw ski and L. Zhang RFC 2309: Recommendations on Queue Management in April 1998

[3] S. Floyd and V. Jacobson, "Random early detection gateway for Congestion avoidance,"IEEE/ACM Transaction on Networking, vol. 1, no.4, pp.397-413, Aug. 1993.

[4] V. Jacobson. Congestion Avoidance and Control. In Proceedings of ACMSIGCOMM, pages 314–329, August 1988.

[5] Dong Lin, Robert Morris, ``Dynamics of random early detection''. ACM, Computer Communication Review, vol.27, no.4, Oct. 1997, pp.127-37, USA

[6] S. Floyd, "Recommendations on using the gentle variant of RED," May 2000. Available at http://wwww.aciri.org/ floyd/red/gentle.html..

[7] Xianliang Jiang,Yang ,Jin and Wei Wei, "RED-FT:A Scalable Random Early Detection Scheme with Flow Trust against DoS Attacks,"IEEE Communication, vol. 17. No. 5 MAY2013

[8] S. Floyd, and V. Jacobson, Random early detection gateways for congestion avoidance, IEEE/ACM Transactions on Networking (TON),1(4),397-413,1993.

[9] Aun Haider, and Richard J. Harris "A Hybrid Random Early Detection Algorithm for Improving End-to-End Congestion Control in TCP/IP Networks" African Journal of Information and Communication Technology, Vol. 4, No. 1, March 2008.

[10]Larry L. Peterson, Bruce S. Davie, Computer networks: a systems approach, 2nd edition, San Francisco: Morgan Kaufmann Publishers, 2000.

[11] B. Braden et al., "Recommendations on queue management and congestion avoidance in the Internet," *Request for Comments (RFC) 2309*, Apr. 1998.

[12] Floyd, S., R. Gummadi, and S. Shenker, Adaptive RED: An Algorithm for Increasing the Robustness of RED's Active Queue Management. Preprint, available at http://www. icir.org/floyd/papers. html, August, 2001.

[13] Teresa Álvarez, Virginia Álvarez, Lourdes Nicolás, UNDERSTANDING CONGESTION CONTROL ALGORITHMS IN TCP USING OPNET, Spain, 2010.

[14] G.F.Ali Ahmed, Reshma Banu, Analyzing the performance of Active Queue Management Algorithms, International journal of Computer Networks & Communications (IJCNC), Vol.2, No.2, March 2010.

[15] Chengyu Zhu, O.W.W. Yang, J. Aweya, M. Ouellette, Montuno, A comparison of active queue management algorithms using the OPNET Modeler, Communications Magazine, IEEE, volume 40, Pages: 158 – 16, June 2002.

[16] G.Thiruchelvi et al., A Survey on Active Queue Management Mechanisms, IJCSNS International Journal of
Computer Science and Network Security, VOL.8 No.12, December 2008.

[17] Sunitha Burri, BLUE: Active Queue Management CS756 Project Report, May 5, 2004.

[18] Michael Welzl, Leopold Franzens Network Congestion Control Managing Internet Traffic, University of Innsbruck.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 420 - 426, 2014, ISSN: 2319–8656

Router Based Mechanism for Mitigation of DDoS Attack- A Survey

Tamana
Department of CE
UCOE, Punjabi University
Patiala, India

Abhinav Bhandari
Department of CE
UCOE, Punjabi University
Patiala, India

Abstract*:* Today most of the activities like trade, e-commerce are dependent on the availability of Internet. The growing use of internet services in the past few years have facilitated increase in distributed denial of service attack. Due to DDos attacks, caused by malicious hosts secured data communication over the internet is very difficult to achieve and is the need of the hour. DDos attacks are one of the most widely spread problems faced by most of the internet service providers (ISP's). The work which had already been done was in the direction of detection, prevention and trace-back of DDos attack. Mitigation of these attacks has also gained an utmost importance in the present scenario. A number of techniques have been proposed by various researchers but those techniques produce high collateral Damage so more efforts are needed to be done in the area of mitigation of DDos attacks.

This paper focuses on **Distributed Denial of Service attack, surveys, classification and also proposed mitigation techniques revealed in literature by various researchers.**

Keywords: ***distributed denial of service, congestion control, flooding attack, legitimate traffic;***

1. INTRODUCTION

The current Internet is vulnerable to attacks and failures. The past events have illustrated the Internet's vulnerability to distributed denial of service (DDos) attacks. The number of Denial of Service (DoS) and Distributed Denial of Service (DDos) attacks on the Internet has risen sharply in the last several years. Distributed Denial of Service (DDos) attacks have become an increasingly frequent disturbance. Internet Service providers are routinely expected to prevent, monitor and mitigate these types of attacks which occur daily on their networks. Denial of service attack is an active type of attack that affects availability infrastructure of the internet. The DoS which is considered here creates flood which uses bandwidth of the channel to be used by clients for legitimate work from server machine.

DDos attacks are often launched by a network of remotely controlled, well organized, and widely scattered Zombies or Botnet computers that are simultaneously and continuously sending a large amount of traffic and/or service requests to the target system that occupy a significant proportion of the available bandwidth. Hence, DoS attacks are also called bandwidth attacks. The aim of a bandwidth attack is to consume critical resources in a network service. Possible target resources may include CPU capacity in a server, stack space in network protocol software, or Internet link capacity. By exhausting these critical resources, the attacker can prevent legitimate users from accessing the service. A crucial feature of bandwidth attacks is that their strength lies in the volume rather than the content of the attack traffic. This has two major implications:

(1) Attackers can send a variety of packets. The attack traffic can be made arbitrarily similar to legitimate traffic, which greatly complicates defense.
(2) The volume of traffic must be large enough to consume the target's resources. The attacker usually has to control more than one computer to generate the attack traffic. Bandwidth attacks are therefore commonly DDos attacks.
They are very hard to defend against because they do not target specific vulnerabilities of systems, but rather the very fact that the target is connected to the network. All known DDos attacks take advantage of the large number of hosts on the Internet that have poor or no security; the perpetrators break into such hosts, install slave programs, and at the right time instruct thousands of these slave programs to attack a particular destination.

2. DDoS ARCHITECTURE

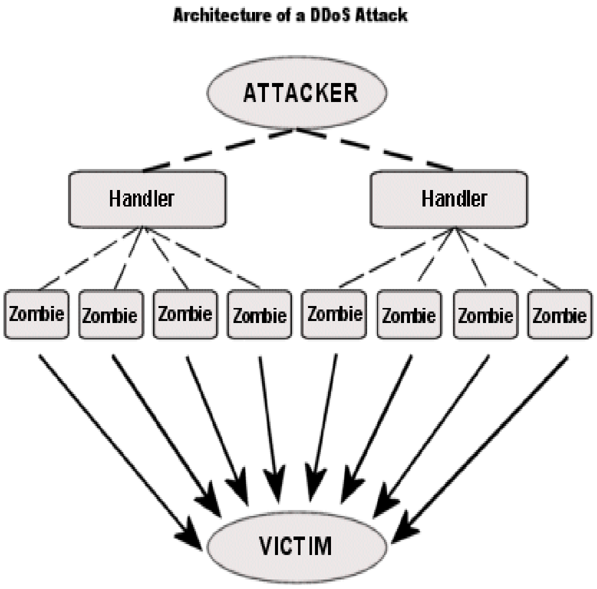


Figure 1 Architecture of DDoS

The remaining part of this paper is organized as follows: Section 2 represents the architecture of DDoS, Section 3 Classifies attacks on various parameters, Section 4 describes Related work, Section 5 elucidates mitigation technique, Section 6 Architecture of explained technique i.e pushback. Section 7 is our conclusion.

3. ATTACK CLASSIFICATION

The types of attack are categorized as following:

• **Classification by degree of autonomy**, that is divided to manual, semi- automatic, or automatic. The automatic methods could be further classified by their communication mechanism (direct, indirect), host scanning strategy (random, hitlist, signpost, permutation, local subnet), vulnerability scanning strategy (horizontal, vertical, coordinated, stealthy), and propagation mechanism (central, back-chaining, autonomous).
• **Classification by exploited weakness**, that is either semantic or brute-force.
• **Classification by source address validity**, that is either spoofed or valid. The spoofed mechanisms could be further divided into routable or non-routable based on address routability or random, subnet, enroute, fixed based on spoofing technique.
• **Classification by possibility of characterization** and if it is characterizable, then whether the traffic is filterable or non-filterable.
• **Classification by attack rate dynamics**, which is either constant, increasing, or fluctuating rate.
• **Classification by the impact on the victim**, which is either disruptive (self, human- or non-recoverable) or degrading.

• **Classification by victim type**, which is application, host, resource, network, or infrastructure.

• **Classification by persistence of agent set**, that can be constant or variable.
The categorization for either known or expected defense mechanisms in [16] is summarized below:
• **Classification by activity level**, which was divided to preventive and reactive. Preventive defense mechanisms were further partitioned to attack prevention (system and protocol security) and DoS prevention (resource accounting and multiplication). Reactive methods were split to either classification by attack detection strategy (pattern, anomaly or third-party) or classification by attack response strategy (agent identification, rate-limiting, filtering, or reconfigration).
• **Classification by cooperation degree**, that can be autonomous, cooperative, or interdependent.
• **Classification by deployment location**, that can be victim network, intermediate network, or source network.
• **Classification by attack response strategy**, which had the following subcategories: agent identification, rate limiting, filtering, and reconfigration.

4. RELATED WORK

Distributed Denial of Service attacks have been a real problem for less than three years, and not much published work exists on the subject. Related work falls into two categories: old work that can also be used in countering DDos attacks, and new work specifically aimed at this task. Originally, it was suggested that DDos attacks could be countered by applying resource allocation techniques on network bandwidth. Integrated Services and Differentiated Services are two approaches aimed at isolating flows with specific quality of service (QoS) requirements from lower-priority traffic. It is not clear if this approach would help; Web traffic, which is a significant fraction of network traffic, is likely to remain best-effort, so it will not be protected by QoS requirements. It is also not clear to what extent compromised sources could fake traffic to show it belonged to QoS-protected flows. There is also an approach that is similar to pushback that was described in an Active-Networks-based defense against flooding attacks. There are many congestion-control mechanisms, which might alleviate some of the effects of congestion due to DDos attacks if only they were globally deployed. Random Early Detect (RED) and its variants tries to identify flows that do not obey TCP-friendly end-to-end congestion control, and preferentially drop them. There is also a large body of work (*e.g.*, Fair Queuing, Class-Based Queuing) aimed at allocating specific fractions of the available bandwidth to each flow so that they all get served. The main problem with these approaches is that packets belonging to DDos attacks do not have readily-identifiable flow signatures, and can thus not be identified by these mechanisms. This is the reason why the concept of Aggregate-based Congestion Control was developed. The common problem that all the tracking techniques are trying to solve is that source addresses in attack packets cannot be trusted, because they are very easy to forge. If all edge routers in the entire Internet were implementing source address filtering, this task would be greatly simplified. Of course, most machines where the packets are originating have been compromised by an attacker, and their owners do not even know that they are being

used for an attack. Also, even if the hundreds or thousands of machines that an attack is coming from were known, it is not clear what could be done about them. Finally, it has been suggested that intrusion detection systems or firewalls be used to detect an attack in progress, and notify upstream elements accordingly. We view Aggregate-based Congestion Control and Pushback as complimentary to many of these approaches. For example, a good map of the network with reliable historical traffic profiles from traces can be used to determine sudden changes in traffic profiles that could signal an attack, or help determine how to allocate rate limits in pushback messages.

A number of useful related techniques of mitigation have been reported in this literature. Abraham presented a new packet marking approach i.e. Pi (short for Path identifier) in which path fingerprint is embedded in each packet which enables a victim to identify packets traversing same paths[10]. In this scheme each packet traversing the same path carries the same identifier. Path identifier fits in each single packet so the victim can immediately filter traffic after receiving just one attack packet [10]. Xiuli Wang proposed Pushback to mitigate DDos attacks. It is based on improved Aggregate based congestion control (IACC) algorithm and is applied to routers to defend against bandwidth consumption attacks [1]. In this scheme we first match the attack signature of the packet, if it is matched packet is sent to the rate limiter which will decide whether to drop the packet or not. From the rate limiter the packet is sent to the Pushback daemon which will drop these packets with the help of upstream routers. Ruiliang Chen and Jung- Min Park combined the packet marking and pushback concepts to present a new scheme called as Attack Diagnosis. In this scheme an Intrusion Detection System is installed at the victim that detects the attack. The victim instructs the upstream routers to start marking packets with trace back information based on which victim reconstructs the attack paths and finally upstream routers filter the attack packets. Abraham[17] in 2003 and Raktim[2] in 2010 proposed mitigation techniques based on Path identification and attestation; Nicholas[10] in 2007 proposed Client puzzles to mitigate DDos attacks whereas Antonis Michalas[4] 2010. Ruiliang Chen[15] proposed Throttling or rate limit to mitigate these attacks.

5. MITIGATION TECHNIQUE

PUSHBACK - A technique in which routers learn a congestion signature to tell good traffic from bad traffic based on the volume of traffic to the target from different links. The router then filters the bad traffic according to this signature. A pushback scheme is given to let the router ask its adjacent routers to filter the bad traffic at an earlier stage. By pushing the defense frontier towards the attack sources, more legitimate traffic is protected.

6. PUSHBACK ARCHITECTURE

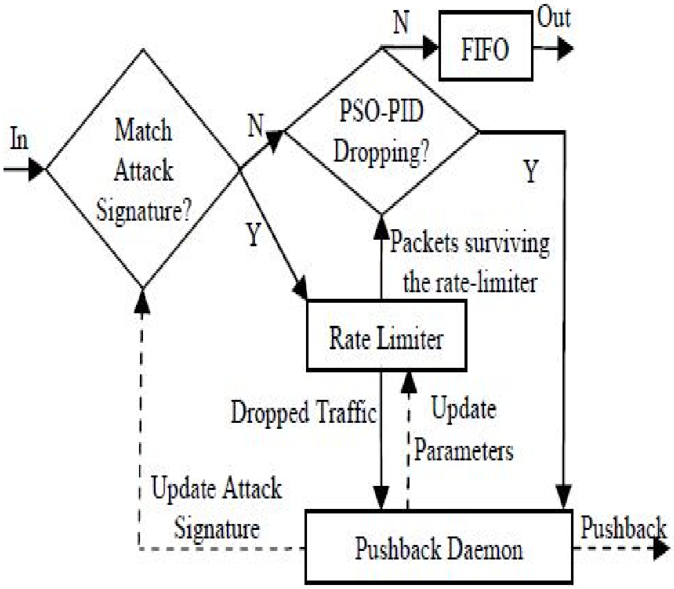


Figure 2 Pushback based on ACC

The input traffic consists of all incoming links of a router. Pushback can be expressed as the following steps:

Step 1 Whether a packet matches attack signature (congestion signature)?

Step 2 *If so*, the packet is sent to the rate limiter, which decides whether a packet is dropped or forwarded according to congestion level. The surviving packets are sent to the PSO-PID drop, go to Step4.

Step 3 *If not*, the packet is sent to the PSO-PID drop directly.

Step 4 The PSO-PID drop decides whether to drop the packet or add the packet to the FIFO output queue.

Step 5 All dropped packets from both the rate limiter and the PSO-PID drop are sent to the Pushback daemon. The daemon requires the upstream routers to drop these packets, periodically updates the parameters of the rate limiter and the attack signature, and also informs the upstream daemons to update theirs.

6.1 Congestion control as a DDos defense and mitigation key:

If we could unequivocally detect packets belonging to an attack and drop just those, the problem would be solved. However, routers cannot tell with total certainty whether a packet actually belongs to a ‘good’ or a ‘bad’ flow; our goal will be to develop heuristics that try to identify most of the bad packets, while trying not to interfere with the good ones. Again, Mahajan et al. introduce the concept of Aggregate-based Congestion Control (ACC); in this context, an aggregate is defined as a subset of the traffic with an identifiable property. For example, “packets to destination D", "TCP SYN packets", or even "IP packets with a bad checksum" are all potential descriptions of aggregates. The task is to identify aggregates responsible for congestion, and preferentially drop them at the routers.

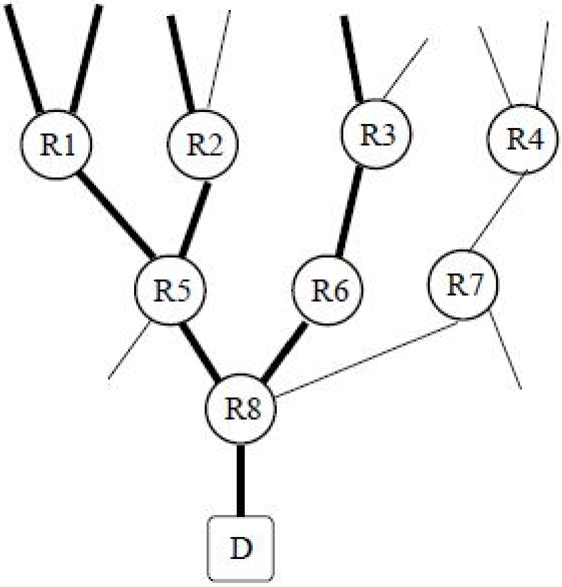


Figure 3 A DDos attack in progress

To illustrate Pushback, consider the network in Figure 3. The server D is under attack; the routers Rn are the last few routers by which traffic reaches D. The thick lines show links through which attack traffic is flowing; the thin lines show links with no bad traffic. Only the last link is actually congested, as the inner part of the network is adequately provisioned. In the absence of any special measures, hardly any non-attack traffic would be reaching the destination. Some non-attack traffic is flowing through the links between R2-R5, R3-R6 R5-R8, R6-R8, and from R8 to D, but most of it is dropped due to congestion in R8 D. Throughout this paper we shall be referring to 'good', 'bad', and 'poor' traffic and packets. Bad packets are those sent by the attackers. Bad traffic is characterized by an attack signature, which we strive to identify; what can be really identified is the congestion signature, which is the set of properties of the aggregate identified as causing problems. Poor traffic consists of packets that match the congestion signature, but are not really part of an attack; they are just unlucky enough to have the same destination, or some other properties that cause them to be identified as belonging to the attack. Good traffic does not match the congestion signature, but shares links with the bad traffic and may thus suffer.

In figure 3, some of the traffic entering R4 is good (the part exiting R7 that is not going to R8), and some is poor, as it is going to D. There may be some good traffic entering R5 from the links above, and exiting from the lower left link, but depending on how congested the links R1-R5 and R2-R5 are, it may suffer. The other links have a mixture of bad and poor traffic. Now, no matter how smart filters R8 could employ, it cannot do anything to allow more good traffic originating from the left side of the graph to reach D. All it can do is preferentially drop traffic arriving from R5 and R6, hoping that more good traffic would flow in via R7. With Pushback, R8 sends messages to R5 and R6 telling *them* to rate-limit traffic for D. Even though the links downstream from R5 and R6 are not congested, when packets arrive at R8 they are going to be dropped anyway, so they may as well be dropped at R5 and R6. These two routers, in turn, propagate the request up to R1, R2, and R3, telling *them* to rate-limit the bad traffic, allowing some of the 'poor' traffic, and more of the good traffic, to flow through.

Table1: Comparative analysis of existing techniques

Name of the technique	Year	Description	Advantages	disadvantages
Egress filtering	IEEE-2010	The IP header of packet leaving are checked for filtering criteria, if criteria is met packet is routed otherwise it is not sent to destination host.	Egress filtering prevents information leaks due to misconfigration, as well as some network mapping attempts.	Decreases performance.
Ingress Filtering	IEEE-2010	In this method filters identify the packets entering the domain and drops the traffic with IP address that does not match the domain prefix connected to a ingress router.	It checks the source IP field of IP packets it receives, and drops packets if the packets don't have an IP address in the IP address block that the interface is connected to.	Keeping track of the many legitimate addresses that can go through a large ISP is next to impossible. It is better to have security as close to the source as possible.
Pushback	IEEE-2008	In this method when the congestion level reaches a certain threshold, sending router starts dropping the packets and illegitimate traffic can be calculated by counting the number of packets dropped for a particular IP address as attackers change their IP address constantly.	Pushback works on aggregates i.e packets from one or more flows carrying common traits. Most effective when attack is non isotrophic. Promising way to combat DDos attack and flash crowds.	The deployment of filters in upstream routers really depends on the downstream router's ability to estimate what fraction of the aggregate comes from each upstream router.
IP Trace Back- Rate Limiting	IEEE-2006	In this Internet traffic is trace back to the true source rather spoofed IP address which helps in identifying attackers traffic and possibly the attacker.	Reduced marking overhead due to ingress filtering. No need path construction algorithm.	Difficult to set threshold values for accurate results.
Path Fingerprint	IEEE-2003	Path Fingerprint represents the route an IP packet takes and is embedded in each IP packet, IP packet with incorrect path fingerprint are considered spoofed.	Path Fingerprint moves Pushback filters close to the attack.	Path Fingerprint is a per-packet deterministic mechanism.

Attack Diagnosis	IEEE-2003	In this scheme an Intrusion Detection System is installed at the victim which detects the attack. The victim instructs the upstream routers to start marking packets with trace back information based on which victim reconstructs the attack paths and finally upstream routers filter the attack packets.	Attack Diagnosis effectively thwarft attacks involving a moderate number of zombies.	It is not appropriate for large scale attacks. Attack Diagnosis trace back and throttles the traffic of one zombie at a time.

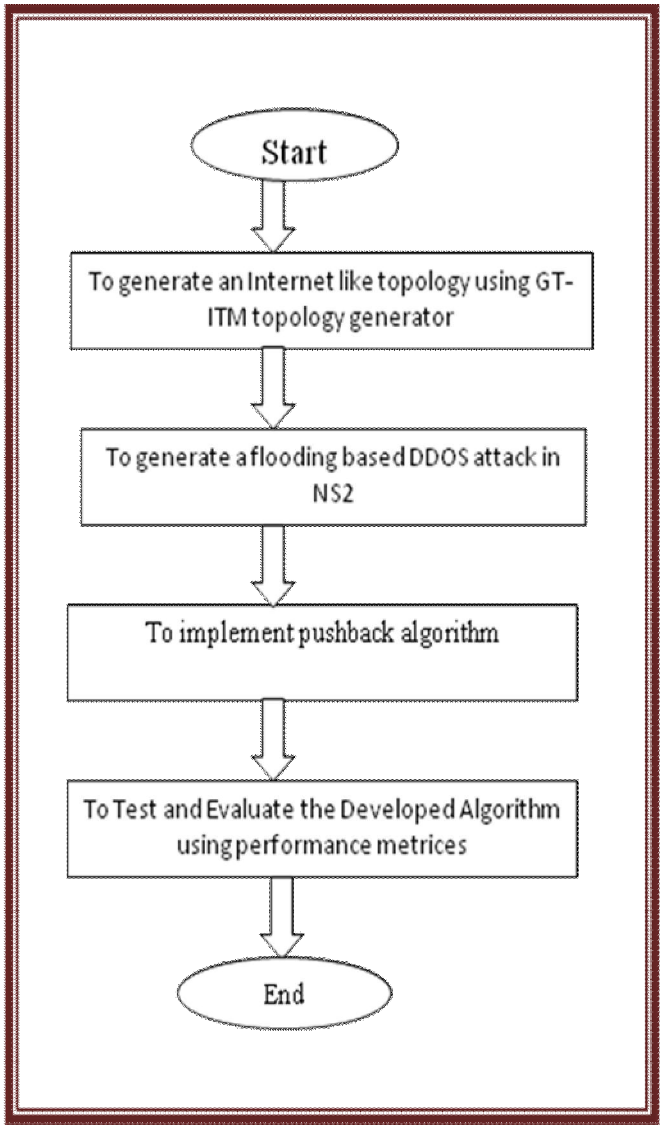

Start
To generate an Internet like topology using GT-ITM topology generator
To generate a flooding based DDOS attack in NS2
To implement pushback algorithm
To Test and Evaluate the Developed Algorithm using performance metrices
End

Figure 4 Generic Methodology

The topology generator can generate topology using any standard graph generator (GT-ITM, Tiers etc). Currently it supports GT-ITM topology generator only and converts the topology graph into ns format.

7. CONCLUSION

In this paper, we presented a review on Distributed Denial of Service attack and defense techniques with an emphasis on pushback technique based on router based mechanism. With such enriched attacks, the defense is even more challenging especially in the case of application layer DDoS attacks where the attack packets are a form of legitimate-like traffic mimicking in the events of flash crowds. The major challenge is to distinguish between actual DDoS attack from flash crowd.

Major challenge in the area of mitigation is that testing and evaluation of mitigation technique have not been done in a comprehensive manner. So various experimental Scenario's are needed to be considered for the same.

8. REFERENCES

[1] Xiuli wang "Mitigation of DDos Attacks through Pushback and Resource Regulation" International Conference on Multimedia and Information Technology 2008.

[2] Raktim Bhattacharjee, S. Sanand, and S.V. Raghavan. "Path Attestation Scheme to avert DDos Flood Attacks" International Federation for Information Processing,2010.

[3] Antonis Michalas, Nikos Komninos, Neeli R. Prasad, Vladimir A. Oleshchuk"New Client Puzzle Approach for DoS Resistance in Ad hoc Networks" IEEE 2010.

[4] Nicholas A. Fraser, Douglas J. Kelly, Richard A. Raines, Rusty O. Baldwin and Barry E. Mullins "Using Client Puzzles to Mitigate Distributed Denial of Service Attacks in the Tor Anonymous Routing Environment" ICC, 2007.

[5] Ruiliang Chen, Jung-Min Park "Attack Diagnosis: Throttling Distributed Denial of Service Attacks Close to the Attack Sources" IEEE 2005.

[6] Abraham Yaar, Adrian Perrig, Dawn Song "Pi: A Path Identification Mechanism to Defend against DDos Attacks" IEEE 2003.

[7] http://home.gwu.edu/~ecamposn/software.html Nscript NS-2 scripting tool

[8] http://www.isi.edu/nsnam/vint/ Virtual Internetwork Testbed collaboration

[9] http://www.isi.edu/nsnam/ns/ NS-2 website

[10] Yinan Jing, Xueping Wang, Xiaochun Xiao, Gendu Zhang"Defending Against Meek DDos Attacks By IP Trace-back based Rate Limiting"IEEE 2006.

[11] John Ioannidis, Steven M. Bellovin," Implementing Pushback: Router-Based Defense Against DDos Attacks" AT&T Labs Research

[12] DHWANI GARG," A Comprehensive Survey of Distributed Defense Techniques against DDos Attacks" International Journal of Computer Science and Network Security VOL.9 No.12, December 2009

[13] Chen, S. and Song, Q(2005).Perimeter-Based Defense against High Bandwidth DDos Attacks. IEEE Transactions on Parallel and Distributed Systems 16(6): 526-537

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 427 - 433, 2014

Comparative Analysis of Quality of Service for Various Service Classes in WiMAX Network using NS-3

Gaurav Sharma
CSE and IT Department
BBSB Engineering College
Fatehgarh Sahib, India

Rishma Chawla
CSE and IT Department
RIET College
Phagwara, India

Abstract: Broadband access is an important requirement to satisfy user demands and support a new set of real time services and applications. WiMAX, as a Broadband Wireless Access solution for Wireless Metropolitan Area Networks, covering large distances with high throughput and is a promising technology for Next Generation Networks. Nevertheless, for the successful deployment of WiMAX based solutions, Quality of Service (QoS) is a mandatory feature that must be supported. Quality of Service (QoS) is an important consideration for supporting variety of applications that utilize the network resources. These applications include voice over IP, multimedia services, like, video streaming, video conferencing etc. In this paper the performances of the MPEG-4 High quality video traffic over a WiMAX network using various service classes has been investigated. To analyze the QoS parameters, the WiMAX module developed based on popular network simulator NS-3 is used. Various parameters that determine QoS of real life usage scenarios and traffic flows of applications is analyzed. The objective is to compare different types of service classes with respect to the QoS parameters, such as, throughput, packet loss, average delay and average jitter.

Keywords: BE, ertPS, nrtPS, QoS, rtPS, UGS, WiMAX, NS-3

1. INTRODUCTION

WiMAX [8], Worldwide Interoperability for Microwave Access are designed to deliver a metro area broadband wireless access (BWA) service. It is based on Institute of Electrical and Electronics Engineers (IEEE) 802.16 standard. The technology provides basic Internet Protocol (IP) connectivity to the user. The rapid growth of new services based on multimedia applications such as Voice over IP, Audio and Video Streaming, Video Conferencing, File Transfer, e-mail etc. has created a demand for last mile broadband access. The various advantages of BWA include rapid deployment, high scalability, and lower maintenance and upgrade costs, and granular investment to match market growth. Various multimedia applications along with the common email, file transfer and web browsing applications are becoming increasingly popular. These applications send large audio and video streams with variable bandwidth and delay requirements.

On the other hand, remote monitoring of critical services, electronic commerce and banking applications, as well as, network control and signaling do not need strict bandwidth guarantees due to the burst nature of the data transfer. These applications also require reliable and prompt packet routing. The presence of different kinds of applications in a network, results in heterogeneous traffic load. The traffic from different applications may require certain type of quality of service. IEEE 802.16/WiMAX provides different service flow classes for different applications to enhance the performance. In this paper, the Quality of Service (QoS) asprescribed in the WiMAX networks is studied and performance analysis of different service flows is compared with respect to QoS parameters like throughput, packet loss, average delay and average jitter.

2. WiMAX NETWORK IP-BASED ARCHITECTURE

The WiMAX End-to-End Network Systems Architecture document [2] defines the WiMAX Network Reference Model (NRM). It is a logical representation of the network architecture. The NRM identifies functional entities and reference points over which interoperability is achieved. The architecture has been developed with the objective of providing unified support of functionality needed in a range of network deployment models and usage scenarios.

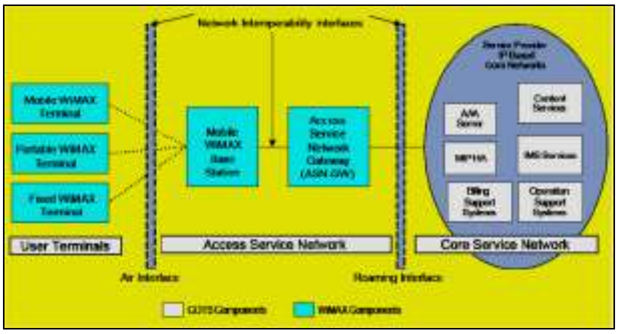


Figure 1: WiMAX Network IP-Based Architecture [3]

Figure 1 shows basic components of a WiMAX network. The subscriber stations (SS) are connected over the air interface to the base station (BS). The base station is part of the Access Service Network (ASN) and connects to the Connectivity Service Network (CSN) through the ASN Gateway. In generic telecommunication terminology, ASN is equivalent to RAN (Radio Access Network) and CSN is equivalent to Core.

The ASN performs the following main functions:

i. WiMAX Layer 2 and Layer 3 connectivity with the subscriber stations, including IP address allocation.
ii. Network discovery and selection of an appropriate network service provider that the subscriber station accesses.

iii. Radio Resource Management

The CSN performs the following main functions:

i. Provides internet access
ii. Authentication, Authorization and Accounting
iii. Policy and admission control based on user subscription profiles

3. QUALITY OF SER VICE IN IEEE 802.16/WiMAX NETWORK

In 1994, QoS, in the field of telephony, was defined in the International Telecommunication Union (ITU) Recommendation E.800. This definition is very broad, listing 6 primary components: Support, Operability, Accessibility, Retainability, Integrity and Security. In 1998, the ITU published a document discussing QoS in the field of data networking. The term Quality of Service refers to the probability of the telecommunication network meeting a given traffic contract. In the field of packet-switched networks and computer networking it is used informally to refer to the probability of a packet succeeding in passing between two points in the network. Although the name suggests that it is a qualitative measure of how reliable and consistent a network is, there are a number of parameters that can be used to measure it quantitatively. These include throughput, transmission delay or packet delay, delay jitter, percentage of packets lost etc. QoS mechanism is added in the layer 2 i.e. Media Access Control (MAC) layer architecture of IEEE 802.16 standards. MAC layer is responsible for scheduling of bandwidth for different users. The MAC layer performs bandwidth allocation based on user requirements as well as their QoS profiles. The standard is designed to support a wide range of applications. These applications may require different levels of QoS. To accommodate these applications, the 802.16 standard has defined five service flow classes where each service flow is characterized by a mandatory set of QoS parameters, which is tailored to best describe the guarantees required by the applications that the service flow class is designed for. Furthermore, for uplink connections, it also specifies which mechanisms to use in order to request bandwidth. This enables end-to-end IP based QoS. The service flows can be created, changed, or deleted by the issuing Dynamic Service Addition (DSA), Dynamic Service Change (DSC), and Dynamic Service Deletion (DSD) messages.

Table 1. Five service flows defined in MAC layer architecture of IEEE 802.16 standards

Service Class	Application	Delay Sensitivity
UGS	Voice	No Delay
rtPS (real-time Polling Service)	Streaming video	High
ertPS (extended real- time Polling Service)	Voice over IP with silence suppression	Very high
nrtPS (non real-time Polling Service)	FTP, messaging, games	Moderate
Best Effort	Email, Web Browsing	Low

Each of these actions can be initiated by the Subscriber Station (SS) or the Base Station (BS) and are carried out through a two or three-way-handshake. The five service flows defined in MAC layer architecture of IEEE 802.16 standards are summarized in Table 1.

3.1 Unsolicited Grant Services (UGS):

UGS is designed to support constant bit rate (CBR) services, such as T1/E1 emulation, and voice over IP (VoIP) without silence suppression. It offers transmission authorization on a periodic basis. UGS traffic is scheduled in a way that SS has a dedicated slot (of fixed size) in which it transmits, and never has to ask for bandwidth for this service (except when creating flow). This guarantees the data rate for the connection.

3.2 Real-Time Polling Services (rtPS)

rtPS is designed to support real- time services that generate variable size data packets on a periodic basis, such as MPEG video or VoIP with silence suppression. In opposition to UGS, the SS should perform explicit requests, which will simply an increase of the overhead and latency i.e. in rtPS class, BS provides periodic uplink request opportunities that match the requested real-time needs in which a SS can specify the desired bandwidth.

3.3 Extended Real Time Polling Service (ertPS)

Another service type called ertPS (Extended rtPS) was introduced to support variable rate real-time services such as VoIP and video streaming. It has an advantage over UGS and rtPS for VolP applications because it carries lower overhead than UGS and rtPS.

3.4 Non-Real-Time Polling Services (nrtPS)

The nrtPS is designed to support non-real-time services that require variable size data grant burst types on a regular basis. It is very similar to rtPS but SSs can ask for bandwidth in a random fashion. In the nrtPS class, BS polls on a regular basis (minimum traffic rate is achieved, but not latency).

3.5 Best Effort (BE) Services

BE services are typically provided by the Internet today for Web surfing. By definition this is a class of service that does not provide any guarantees in terms of throughput and / or delays. For BE, SS may use contention request opportunities, as well as unicast polls when the BS sends them. Since BS doesn’t need to poll for BE traffic, a long period may pass before BE packets are sent, especially when network is congested

4. RELATED WORK

Talwalkar and Iliyas [3] have analyzed quality of service in WiMAX networks. In their analysis, a WiMAX module is developed based on popular network simulator NS-2. In their simulation they have used VoIP and video traffic for the analysis of QoS for three service classes (BE, UGS, rtPS) in WiMAX networks, while the QoS for nrtPS and ertPS has not been analyzed because the WiMAX module developed does not support the nrtPS and ertPS service classes.

In 2010 QoS deployment over a cellular WiMAX network [4] was examined. The author has compared the performance obtained using two different QoS configurations differing from the delivery service class used to transport VoIP traffic, i.e. UGS or ertPS. OPNET modeller version 14.5

with WiMAX module capability was been used for the simulation.

The analysis of various critical QoS parameters like throughput, average jitter and average delay for VoIP using NOAH as protocol in NS-2 simulator has also been done [5]. Their simulation focuses on the QoS parameters for BE service class.

H. Abid, H. Raja, A. Munir, J. Amjad, A. Mazhar and D. Lee [6] performed a performance analysis when multimedia contents are transferred over WiMAX network using BE and ertPS service classes.The analysis of QoS service for WiMAX network using MATLAB for simulation of WiMAX network and AODV as the routing protocol has also been done. This time the performance analysis focuses on UGS service class only [7].

Anouari and Haqiq [8] in 2012 investigated the performances of the most common VoIP codecs, which are G.711, G.723.1 and G.729 over a WiMAX network using various service classes and NOAH as a transport protocol. To analyze the QoS parameters, the popular network simulator NS-2 was used. Various parameters that determine QoS of real life usage scenarios and traffic flows of applications is analyzed. Their objective was to compare different types of service classes with respect to the QoS parameters, such as, throughput, average jitter and average delay. The service classes for which the QoS was analyzed includes BE, UGS and rtPS, again nrtPS and ertPS service classes are not included in the work.

Table 2 Comparison of various techniques, services classes and simulation tools used

Technique	Service	Simulator
Performance analysis of Video Conferencing and Multimedia application Services over WiMAX [11]	BE, UGS, rtPS	NS-2
Measuring data and voip traffic in wimax networks [4]	UGS, rtPS	OPNET
Analysis of VoIP traffic in WiMAX using NS2 simulator [5]	BE	NS-2
Performance Analysis of WiMAX Best Effort and ertPS Service Classes for Video Transmission [6]	BE, ertPS	NS-2
Performance Analysis of QoS Parameters for Wimax Networks [7]	UGS	MATLAB
Performance Analysis of VoIP Traffic in WiMAX using Various Service Classes [8]	BE, UGS and rtPS	NS-2

From the Table 2 it is concluded that the performance analysis for service classes like BE, UGS, rtPS have been carried out by different authors using different simulating tools like NS-2, MATLAB, OPNET. This gave us the idea to analyze the performance analysis of nrtPS service class for WiMAX network and analyze the comparative results for the same.

This work focuses on comparative performance analysis of different service classes which includes BE, rtPS, nrtPS and UGS when video traffic is transferred over the WiMAX network. To analyze the QoS parameter, simulation based on popular network simulator NS-3 is used. Various parameters that determine QoS of real life usage scenarios and traffic flows of applications are analyzed. The goal is to compare different types of service flows with respect to QoS parameters, such as, throughput, packet loss, average delay and average jitter.

5. SIMULATION DETAILS

To analyze QoS in a network it is necessary to study real life scenarios. The simulation set up would reflect the actual deployment of the WiMAX network. Based on the network reference model described earlier Figure 2 shows the setup that will be used. There are multiple SS's in the range of a base station. The base station is connected to the core network. The focus of analysis will be the point-to-multipoint connection between the subscriber stations and the base station. Various types of traffic can be set up from one or many subscriber station(s) and a base station to mimic real life scenarios. Video streaming will be used to provide real-time variable bit rate traffic i.e. watching videos online.

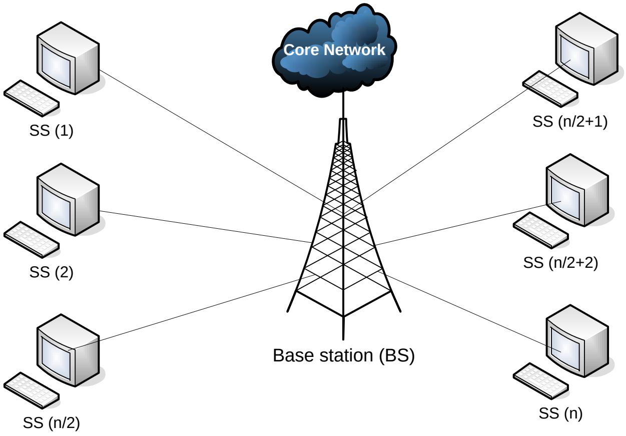


Figure 2. Simulation Setup

NS-3 is a discrete event simulator, that can execute C++ and Python based scripts. NS-3 is likely to be the most popular tool for simulation purposes because of being open source, having well managed source code and easy to use languages like C++. It provides support for UDP and MAC protocols over wired and wireless networks.

Amine Ismail [9] worked on the WiMAX module for NS-3, that suits the requirements of this thesis work and with the help of the basic WiMAX scripts available in the module it became possible to write the scripts. Most network elements in NS-3 simulator are developed as classes, in object-oriented fashion. It is freely distributed and all the source code is available.

Figure 3 shows the basic structure of NS-3 architecture and setup used in this thesis work. The network topology and traffic scenarios etc. are specified in the C++ script. The C++ library has all the implementation details. When NS-3 is run, the resulting simulation data can be obtained in a text file format trace. This file contains time stamp and information about each packet that is sent, received or dropped. It also has information about the packet size, type of packet etc. A base station and subscriber station can be set up as a node in NS-3. As the number of nodes in the simulation increase, the packets that are sent and received increases. This makes the trace file very large. To aid in extracting the right data out from the file, Php script is written. This extracted data is used to prepare charts and graphs for the results of the simulation.

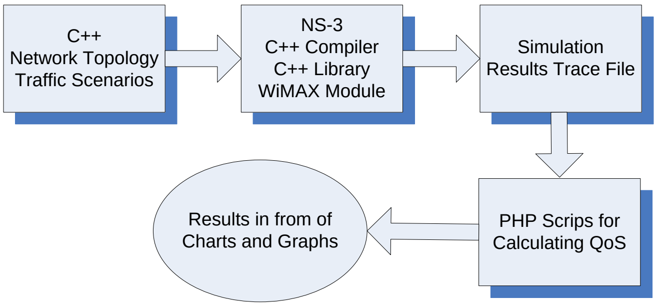


Figure 3 Ns-3 Architecture and set up

The NS-3 WiMAX module has the distinction between nodes. While setting up the node, it could be classified as a base station as opposed to a subscriber node.

The WiMAX module developed by Amine Ismail is used, that work on IEEE 802.16e standard. The WiMAX module simulates both physical (PHY) and MAC layers based on IEEE 802.16e standard.

After setting up NS-3 and compiling the WiMAX module it was discovered that there are several memory leaks in the simulation. For a single node the simulation would run properly but as the nodes increased, so would the packet traffic. This would eventually cause the simulation to stop to run as it would run out of dynamic memory necessary to allocate new packets. Several memory leaks issues were fixed during the course of experimentation, to make the simulation traffic.

As mentioned earlier, simulation scenario for NS-3 is defined by a C++ script. In a typical model, the nodes in the network are setup, whose properties are changed to base station and subscriber stations. The subscriber stations are associated with the base station. Application traffic agent(s) are created and are attached to the source node(s). An example of a traffic Agent could be a VBR traffic agent. On the top of the traffic agent, an application which generates required traffic is created. Half of the subscriber stations will act as the traffic sources and another half of the subscriber stations will act as traffic sinks. For that purpose only an even number of subscriber stations like, 2, 4, 6, 8 and 10 are considered. The focus of the study will be the measurement and analysis of the QoS parameters using the information gathered from the total downlink packets, from base station to subscriber station, and uplink packets, from subscriber station to base station.

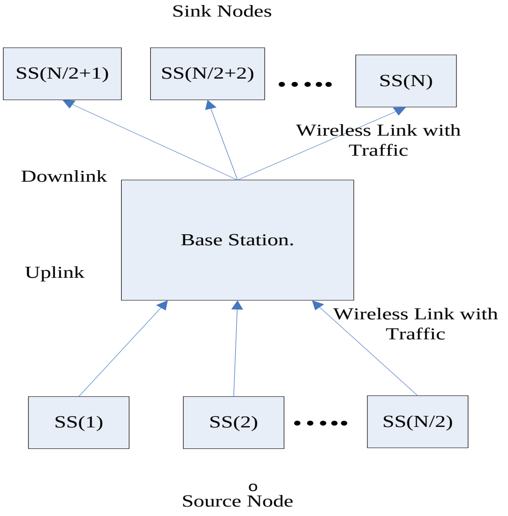


Figure 4 NS-3 Simulation Example

Figure 4 displays a typical example of the NS-3 setup. In the WiMAX module, the simple nodes are created, that are converted into Base Station (BS) nodes and Subscriber Station (SS) nodes using the WiMAX module that creates Subscriber Stations as DEVICE_TYPE_SUBSCRIBER_STATION and the base station is created as DEVICE_TYPE_BASE_STATION. Thereafter, the SS and BS nodes are converted into network devices and IP addresses are assigned to them. This is done to provide the network facilities to the SS & BS nodes. Mobility can also be applied to the nodes or they can be kept static. After that half of the SS nodes are converted into UDP servers and reset half of the SS nodes are converted into UDP clients. The UDP trace file (video streaming traffic) is applied on the UDP server(s) that is transmitted by the UDP servers and received by UDP client(s),for the duration for which the UDP application runs. The service flows can be changed as per the requirement. The types of service flows that are supported by the WiMAX module are BE, rtPS, nrtPS and UGS. At this point, ertPS service flow is not supported.

The C++ script was parameterized. The input parameters that were varied, for example, number of nodes in the simulation, service flow classes were passed in as a parameter while running the simulation for easy execution.

6. QUALITY OF SERVICE PARAMETERS

QoS provisioning encompasses providing Quality of Service to the end user in terms of several generic parameters. The perceived quality of service can be quantitatively measured in terms of several parameters. In the analysis, the throughput, average delay, average jitter and packet loss were considered.

6.1 Throughput

Throughput is the amount of number of packets effectively transferred in a network, in other words throughput is data transfer rate that are delivered to all terminals in a network. It is measured in terms of packets per second or per time slot It is a measure of the date rate (bits per second) generated by the application. Equation 1 shows the calculation for throughput TP, where $PacketSize_i$ is the packet size of the *ith* packet reaching the destination, $PacketStart_o$ is the time when the first packet left the source and $PacketArrival_n$ is the time when the last packet arrived.

$$TP = \frac{\sum_i PacketSize_i}{PacketArrival_n - PacketStart_0}$$

Equation 1: Throughput Calculation

From the trace file, based on the packet ID, each data packet was kept track of. The time a packet is sent, the time when the packet was received and the packet size was stored for all packets that reached the destination. To calculate throughput, the size of each packet was added. This gave the total data that was transferred. The total time was calculated as the difference between the time the first packet started and the time the last packet reached the destination. Thus throughput is equal to the total data transferred divided by the total time it took for the transfer.

6.2 Average Delay or latency

Delay or latency would be time taken by the packets to transverse from the source to the destination. The main sources of delay can be further categorized into source-processing delay, propagation delay, network delay and destination processing delay. Equation 2 show the calculation for Average Delay, where *PacketArrival*$_i$ is the time when packet "i" reaches the destination and *PacketStart*$_i$ is the time when packet "i" leaves the source. "*n*" is the total number of packets.

$$AverageDelay = \frac{\sum_i PacketArrival_i - PacketStart_i}{n}$$

Equation 2: Average Delay

From the trace file, difference between the start time of the packet and time when the packet reaches destination is calculated. The average of all these times gives the average delay.

6.3 Jitter or Delay Variation

Jitter can be observed as the end-to-end delay variation between two consecutive packets. The value of jitter is calculated from the end to end delay. Jitter reveals the variations in latency in the network caused by congestion, route changes, queuing, etc. Delay variation is the variation in the delay introduced by the components along the communication path. It is the variation in the time between packets arriving. Jitter is commonly used as an indicator of consistency and stability of a network. Measuring jitter is critical element to determining the performance of network and the QoS the network offers. Equation 3 shows the steps for calculation of average jitter. It is the average of the absolute difference in the time it took for successive packets to reach the destination.

$$AverageJitter = \frac{\sum_i \left| \begin{array}{l} \left(PacketArrival_{i+1} - PacketStart_{i+1}\right) - \\ \left(PacketArrival_i - PacketStart_i\right) \end{array} \right|}{n-1}$$

Equation 3: Average Jitter

There are five packets that arrive at the destination with slightly varying delays. Note that the numbers mentioned here do not represent any of the actual data. They are thought about to make the calculations clear. In the last column, the absolute value of difference in successive packets is calculated. Thus average jitter is the average of delay difference in successive packets.

The packets should arrive at the same delay. The delay entries in "Delay" column will be identical. This will have the delay variation to be zero, implying no jitter.

6.4 Packet loss or corruption rate

Packet loss affects the perceived quality of the application. Several causes of packet loss or corruption would be bit errors in an erroneous wireless network or insufficient buffers due to network congestion when the channel becomes overloaded. Equation 4 shows the simple equation to calculate packet loss. It is the sum of all the packets that do not reach the destination over the sum of the packets that leave the destination.

$$PacketLoss = \frac{\sum LostPacketSize_i}{\sum PacketSize_j} \times 100$$

Equation 4: Packet Loss

Packet loss calculation is relatively simple. The sum of the packet size of all packets that are sent is calculated. Next, the sum of all packets that are received is calculated. The difference in the two values gives the data that was lost. The ratio of total data lost and the total data that was sent gives the packet loss.

A php script was written to analyze the ns-3 trace files and calculate the four parameters.

7. SIMULATION RESULTS

To analyze the quality of service in WiMAX network, real life use case is considered. WiMAX provides basic IP connectivity. With the availability of a larger data pipe, viewing videos over the internet is common application these days. So video streaming is analyzed. In this section, the simulation results for QoS parameters, obtained for video traffic, are presented.

A number of subscriber stations generating video traffic are set up in this experiment. The performance of QoS service flows in terms of the QoS parameters was analyzed. Video streaming is Variable Bit Rate (VBR) traffic. Unlike Constant Bit Rate (CBR) the packet size varies based on each frame type e.g. in case of a MPEG traffic, the 'I' and 'B' frames are smaller in size than a 'P' frame. Similarly for H.263 traffic, there are 'I' frames, 'P' frames and 'PB' frames. For the analysis using the simulation, a high quality MPEG-4 data stream was used. MPEG-4 is designed as a simple and straight forward video coding with enhanced compression performance and to provide a network friendly video representation. MPEG-4 video streaming is need to be packetized for transportation over networks. The transport protocol for MPEG-4 is Real-time Transport Protocol (RTP). The VBR frame stream for the high quality MPEG-4 encoded Mr. Bean movie was obtained from the following website in [10]. Using this information, a trace file is generated. This trace file is attached to the UDP agent as a traffic source. The trace file is in binary format and contains information of the time, frame number, frame type and packet size. Analysis is done using BE, nrtPS, rtPS and UGS service flows for the video traffic. The parameters analyzed are throughput, packet loss, average jitter and average delay. These parameters are observed for each service flow as the number of nodes with the video traffic increases.

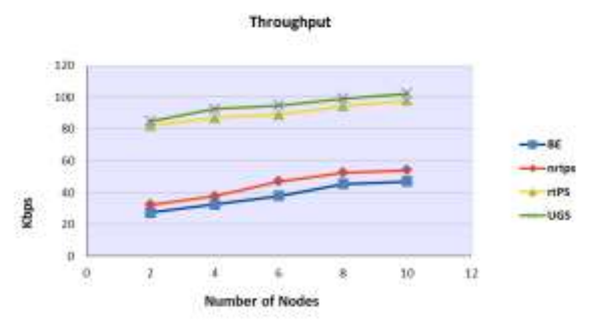


Figure 5: Overall throughput for video traffic as the number of nodes increases

Figure 5 shows the variation of overall throughput for video traffic as the number of nodes increases. The overall throughput of rtPS and UGS service flow is higher than BE and nrtPS service flow. For 10 nodes the value of throughput for BE service flow is lowest. If the throughput using UGS service flow is compared with three other service flows i.e. BE, nrtps and rtPS, the UGS service flow provides much higher throughput than BE and nrtPS service flow and nearly equal throughput to the rtPS service flow for the same number of nodes.

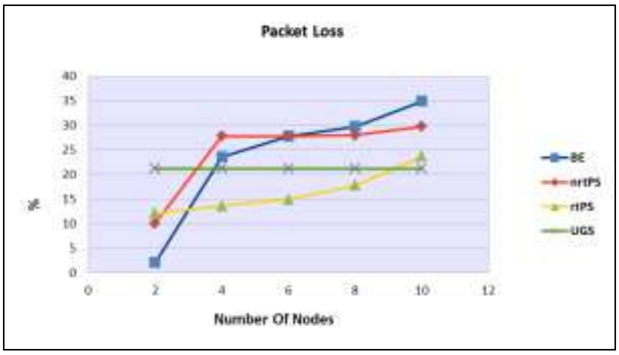


Figure 6 Variation of packet loss for video traffic with all four service flows.

Figure 6 shows the variation in packet loss for video traffic with all four service flows. UGS service flow provides constant packet loss. In case of BE and nrtPS service flow the packet loss increases gradually as the number of nodes streaming video traffic in network increases. The value of packet loss for 2 nodes is much less than the value for 10 nodes for both the service flow i.e. BE and nrtPS. However in case of rtPS service flow there is very slight variation in packet loss as the number of nodes increases from 2 to 10 and the value of packet loss is lowest than all three service flows for increasing number of nodes.

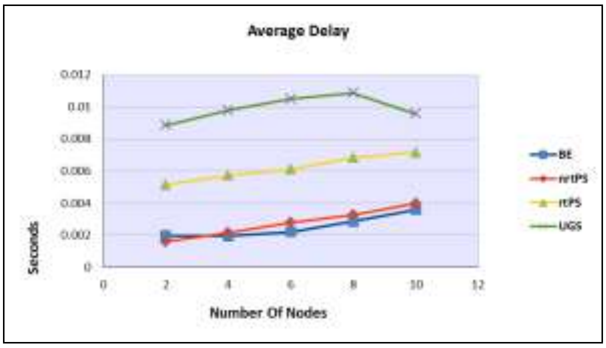


Figure 7 Average delay for four service flows

Average Figure 7 shows the average delay for the four service flows i.e. BE, nrtPS, rtPS and UGS as the number of nodes increase. UGS has higher delay as compared with BE, nrtPS and rtPS flows, although the delay variation is less for UGS. This is because of UGS offers fixed grants on a periodic basis. There is no bandwidth request mechanism in BE traffic. Data is sent whenever resources are available and not required by any other scheduling-service classes. As the number of nodes increase the average delay increases rapidly for BE service flow as compared to rtPS service flow. The value of average delay is higher in rtPS service flow as compared to nrtPS service flow but in case of rtPS the variation in delay value is comparatively much less than the nrtPS sevice flow. Similar to BE service flow the value of delay increases rapidly in nrtPS service flow as the number of nodes streaming video traffic increases.

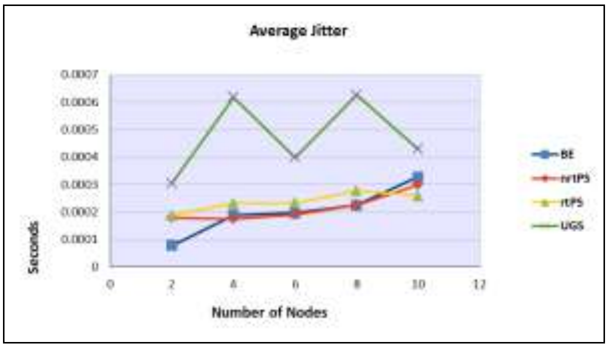


Figure 8 Variation in average jitter with all four service flows

Figure 8 shows the variation in average jitter with all four service flows. From the Figure 8 it can be seen that the value of jitter does not show any relation with number of nodes for UGS flow. However, rtPS service flow has lowest jitter value than all the four service flows i.e. BE, nrtPS and UGS for maximum number of nodes i.e. 10 nodes and it remains more or less constant. For small number of nodes, BE service flow has the lowest jitter. As the number of nodes increases, the jitter value starts increasing rapidly. Thus when lot of network resources are available video traffic over BE service flow introduces the least jitter. However, as the nodes increase, the network resources get divided between all the nodes. Hence, it results into the increased jitter. However in case of rtPS service flows the average jitter is having a small variation with the increase in the number of nodes and is almost constant.

8. CONCLUSION

The performance analysis of different service flows namely BE, nrtps, rtPS and UGS, on QoS parameters like throughput, packet loss, average delay and average jitter was analyzed and compared when video traffic is passed with increasing number of nodes over WiMAX network. During the analysis rtPS service flow comes out to be better than all other three service flow for average jitter and packet loss. The variation in value of average jitter is very less in case of rtPS service flow and it has least jitter value for maximum number of nodes i.e. 10 than all other service flows. In case of average delay the value for rtPS service flow is high than the BE and nrtPS service flow but there is very slight variation in value with increasing number of nodes which increases rapidly in case of other two. The delay is maximum in UGS service flow for video traffic as it is a VBR traffic. rtPS service flows shows the least packet loss while streaming video traffic with increased number of nodes and throughput is much higher as compared to BE and nrtPS service flow and is nearly equal to UGS service flow with increased number of nodes. As the UGS service flow does not utilize the network resources effectively when the traffic is not Constant Bit Rate (CBR) traffic and streaming video traffic is Variable Bit Rate (VBR) traffic. The bandwidth can be periodically requested in rtPS service flow instead of fixed bandwidth already being allocated, which may or may not get used. Thus, by considering all the results, it can be concluded that for streaming video traffic rtPS service flow is best suited.

9. FUTURE SCOPE

Video streaming was considered in the current analysis. Further analysis could be done for other applications including VOIP, video telephony which combines video traffic and VOIP traffic, File Transfer Protocol (FTP) traffic etc. The

WiMAX module used in the analysis did not support ertPS service flow that is defined by the IEEE 802.16 standards. ertPS service flow is designed for applications which generate variable rate traffic which are delay dependent. An example of such traffic is VOIP with silence suppression.

10. REFERENCES

[1]. Retnasothie, F. E., Ozdemir, M. K., Cek, T.Y., Celebi, H., Zhang, J., and Muththaiah, R., “Wireless IPTV over WiMAX: Challenges and Applications”,Proceedings of IEEE WAMICON, [2006].

[2]. WiMAX End-to-End Network Systems Architecture - Stage 2: Architecture Tenets, Reference Model and Reference Points,” WiMAX Forum, December, [2005]

[3]. Talwalkar, R. A. and Ilyas, M.,“Analysis of Quality of Service (QoS) in WiMAX Networks”, IEEE International Conference on Networking, [2008].

[4]. Adhicandra, I., "Measuring data and voip traffic in wimax networks," Arxiv Preprint arXiv:1004.4583, [2010].

[5]. Joshi, D. and Jangale, S., “Analysis of VoIP traffic in WiMAX using NS2 simulator” International Journal of Advanced Research in Computer Science and Electronics Engineering, Vol. 1, Issue 2, April [2012].

[6]. Abid, H., Raja, H., Munir, A., Amjad, J., Mazhar, A. and Lee, D., “Performance Analysis of WiMAX Best Effort and ertPS Service Classes for Video Transmission”, ICCSA, Issue 3, Page(s): 368-375, [2012].

[7]. Vikram, M. and Gupta, N.,"Performance Analysis of QoS Parameters for Wimax Networks." International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Page(s): 105-110, [2012].

[8]. Anouari, T. and Haqiq, A., "Performance Analysis of VoIP Traffic in WiMAX using Various Service Classes", International Journal of Computer Applications (0975 - 8887), Volume 52, No. 20, August [2012].

[9]. Ismail, M. A., Piro, G., Grieco, L. A. and Turletti, T. “ An Improved IEEE 802.16 WiMAX Module for the NS-3 Simulator”, International Conference on Simulation Tools and Techniques (SIMUTools’10), March, [2010].

[10]. “MPEG-4 and H.263 Video Traces for Network Performance Evaluation” (Master's thesis) Athamneh K. http://trace.eas.asu.edu/TRACE/pics/FrameTrace/mp4/Verbose_bean.dat

[11]. Kaarthick, Yeshwenth, V. J., Nagarajan. N. and Rajeev, “Performance analysis of Video Conferencing and Multimedia application Services over WiMAX”, IEEE International Advance Computing Conference, Page(s):1109-1123, March [2009].

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 434 - 438, 2014

A Fault tolerant system based on Genetic Algorithm for Target Tracking in Wireless Sensor Networks

Venkatesh S
Department of Computer Science and Engineering,
B.S.Abdur Rahman University,
Vandalur, Chennai, India.

K.M.Mehata
Department of Computer Science and Engineering,
B.S.Abdur Rahman University,
Vandalur, Chennai, India.

Abstract- In this paper, we explored the possibility of using Genetic Algorithm (GA) being used in Wireless Sensor Networks in general with specific emphasize on Fault tolerance. In Wireless sensor networks, usually sensor and sink nodes are separated by long communication distance and hence to optimize the energy, we are using clustering approach. Here we are employing improved K-means clustering algorithm to form the cluster and GA to find optimal use of sensor nodes and recover from fault as quickly as possible so that target detection won't be disrupted. This technique is simulated using Matlab software to check energy consumption and lifetime of the network. Based on the simulation results, we concluded that this model shows significant improvement in energy consumption rate and network lifetime than other method such as Traditional clustering or Simulated Annealing.

Keywords: Genetic Algorithm, Wireless sensor network, Fault tolerant, energy efficiency, Network lifetime, K-means clustering, Reliability

1. INTRODUCTION

Nowadays, there are numerous applications in which Wireless Sensor networks being used like Military and civilian as well. This includes target tracking, surveillance, and security management etc. Since a Wireless sensor is a small, lightweight, untethered, battery-powered device, it has limited energy, processing power, memory [15]. Hence, energy consumption is a critical issue in sensor networks. Since these devices are less expensive, It is possible to design and construct the deployment of large scale wireless sensor networks (WSN) with potentially thousands of nodes [1]. WSN and their applications have tremendous potential in both commercial and military environments due to their low cost and pervasive surveillance [2]. For critical environment with high degree of dependability is required, WSN should offer characteristics such as: reliability, availability and maintainability. Availability to a large extend depends on fault tolerance to keep the system working as expected. Availability on the service level means that the service delivered by a WSN (or part of it) is not affected by failures and faults in underlying components such as single nodes or node subsystems. In WSNs, the failure of such components is almost unavoidable. Most of the detection and recovery techniques therefore aim at reducing MTTR (the amount of time required for detecting and recovering from a failure) as much as possible. In these WSN deployments, it is common to have a node providing functionality to its neighbors. In a typical WSN, all data obtained by member sensors must be transmitted to a sink or data collector. More energy will be consumed during transmission, if the communication distance is longer. It is estimated that to transmit a *k*-bit message across a distance of *d*, the energy consumed can be represented as follows:

$$E(k,d)=E_{elec}* k + E_{amp}*k*d^2 \qquad (1)$$

where E_{elec} is the radio energy dissipation and E_{amp} is transmit amplifier energy dissipation. In this scenario, it would be ideal if there is a cluster head and it could aggregate sensor data before it is forwarded to a base station, thereby saving energy. [4] In the cellular networks and ad hoc networks energy requirements is not a constraint as base stations or batteries can be replaced as needed, but nodes in sensor networks have very limited energy and their batteries cannot usually be recharged or replaced due to hostile or hazardous environments. Hence energy saving is one of the important factor in WSN and hence important characteristic of sensor networks is the stringent power budget of wireless sensor nodes. There are two components of a sensor node viz sensing unit and wireless transceiver, usually directly interact with the environment, which is subject to variety of physical, chemical, and biological factors. On such cases, Nodes with stronger hardware capabilities can perform operations for other nodes that would either have to spend a significant amount of energy or would not be capable of performing these operations. These services, however, may fail due to various reasons, including radio interference, de-synchronization, battery exhaustion, or dislocation due to inhospitable conditions. Such failures are caused by software and hardware faults, environmental conditions, malicious behavior, or bad timing of a legitimate action. In general, the consequence of such an event is that a node becomes unreachable or violates certain conditions that are essential for providing a service. Besides, a failure caused by a trivial software bug can be propagated to become a massive failure of the sensor network. In other cases, some sensor nodes became faulty due to transient conditions like high heat due to hostile environment as well. For example, in multiple moving objects tracking which is an active research area in WSNs due to its practical use in a wide variety of applications [3], including military or environmental monitoring applications, may missed out the target being tracked if any one of the sensor nodes that monitoring it becomes faulty. In such scenarios, prediction or node failure in these methods must be handled quickly before target slips far away. For reporting on the other hand, optimization techniques are applicable on different clustering and data fusion methods. Since the number of nodes participate in the tracking process, clustering phase can be the critical phase of the tracking from the energy consumption point of view. After clustering phase is done, it is equally important to use the member nodes in the cluster optimally so that cluster lifetime should be extended which in turn extends the network life time.

The rest of the paper is organized as follows: section II discusses about the assumption and background , section III discusses related work that has been done in WSNs with Fault tolerant and GA perspective, section IV presents the challenges involved in monitoring multiple objects with transient fault in particular and the proposed method and Section V gives the conclusion and the future work.

2. BACKGROUND

A. Assumptions

Following assumptions are made about the sensors and the sensor network in the development of the proposed target tracking algorithm:

- A set of sensors are deployed in a square terrain. The nodes possesses the following properties
- The sensor network is static
- All nodes are assumed to have the capabilities of a cluster-head and the ability to adjust their transmission power based on transmission distance.
- Two nodes communicate with each other directly if they are within the transmission range
- The sensor nodes are assumed to be homogeneous i.e. they have the same processing power and initial energy.
- The sensor nodes are assumed to use different power levels to communicate within and across clusters.
- The sensor nodes are assumed to know their location and the limits S (Number of nodes in each cluster).
- For multiple object tracking, no 2 objects are in same cluster at same time.

3. RELATED WORK

In the paper by Heinzelman et al's paper[5] "Energy-Efficient Communication Protocol for Wireless Micro -sensor Networks" which describes a clustering-based protocol called LEACH. Here the performance of LEACH with direct communication and MTE is being compared. They use a pre-determined optimal number of clusters (5% of the total number of nodes) in their simulations. In the paper [6] Heinzelman et al determine that the optimal number of clusters for a 100-node network to be 3-5 by using a computation and communication energy model; however, determining the optimal number of cluster-heads depends on several factors such as sensor densities, the position of a sink, etc.

In the paper [7] by Tillett et al, the PSO (Particle Swarm Optimization) approach is proposed which would divide the sensor node field into groups of equal sized groups of nodes. PSO is an evolutionary programming technique that mimics the interaction of ants or termites to find a good solution. Although partitioning into equal sized clusters balances the energy consumption of cluster heads, this method is not applicable to some networks where nodes are not evenly distributed.

In the paper [8] by Ostrosky et al, address a somewhat different partitioning problem: Given n points in a large data set, partition this data set into *k* (*k* is known) disjoint clusters so as to minimize the total distance between all points and the cluster-heads to which they belong. The authors use a polynomial-time approximation scheme to solve the problem.

A small number of nodes are selected to become clusterheads. They are responsible for coordinating the nodes in their clusters, for instance by collecting data from them and forwarding it to the base station. In case that a cluster head fails, no messages of its cluster will be forwarded to the base station any longer. The cluster head can also intentionally or due to software bugs forward incorrect information. Depending on the application case, the impact of such a failure can vary from quality degradation of measurements to alarm messages not being delivered to a back-end system. While forwarding messages, nodes can aggregate data from multiple other nodes in order to reduce the amount of data sent to the base station. One common simple approach is to calculate the average of correlated measured values such as temperature, humidity and pressure, sending only one message to the back-end. If a node generates incorrect data, the data aggregation results can suffer deviations from the real value. Also, if a node responsible for generating the aggregated data is subject to a value failure, the base station will receive incorrect information of an entire region of the network.

This paper [9] investigates prediction-based approaches for performing energy efficient reporting in object tracking sensor networks. A dual prediction-based reporting mechanism (called DPR) has been proposed, in which both sensor nodes and the base station predict the future movements of the mobile objects. Transmissions of sensor readings are avoided as long as the predictions are consistent with the real object movements. DPR achieves energy efficiency by intelligently trading off multi-hop/long-range transmissions of sensor readings between sensor nodes and the base station with one-hop/short-range communications of object movement history among neighbor sensor nodes. The impact of several system parameters and moving behavior of tracked objects on DPR performance has been explored, and also two major components of DPR are studied: prediction models and location models through simulations. In this paper, Profile-Based Algorithm (PBA) [10] has been proposed that aims to use the information contained in the network and in the object itself to optimize energy consumption, thus extending lifetime. Here it utilizes the regularity in the object's behavior to reduce energy consumption. In this paper [11], In Target Tracking application, the sensor nodes collectively monitor and track the movement of an event or target object. The network operations have two states: the surveillance state during the absence of any event of interest, and the tracking state which is in response to any moving targets. Thus, the power saving operations, which is of critical importance for extending network lifetime, should be operative in two different modes as well. In this paper, we study the power saving operations in both states of network operations. During surveillance state, a set of novel metrics for quality of surveillance, which suggests that atleast p-sesnor nodes required to cover any location, is proposed specifically for detecting moving objects. In the tracking state, we propose a collaborative messaging scheme that wakes up and shuts down the sensor nodes with spatial and temporal preciseness. In this paper [12], a novel approach toward Base Station (BS) oriented clustering and tracking in WSNs is introduced. Proposed method overlooks ad-hoc ability of WSNs to earn energy efficiency and fault tolerance. BS is a powerful energy and computational resource, therefore, BS is burdened with major part of clustering and tracking operations. 3-D cubic antenna is used to enable our sensors to receive BS packets from long distance. Also, BS has a good knowledge of nodes energy level, as a result, BS rotates activated nodes and CH to avoid load balancing problem.

4. PROPOSED SOLUTION

Genetic Algorithm:

Chromosome:

All living organisms consist of cells. In each cell there is the same set of chromosomes. Chromosomes are strings of DNA and serves as a model for the whole organism. Usually a chromosome consist of genes which is basically blocks of DNA. Each gene encodes a particular protein. Basically it can be said, that each gene encodes a trait, for example color of eyes. Possible settings for a trait (e.g. blue, brown etc) are called alleles. Each gene has its own position in the chromosome. This position is called locus.

Complete set of genetic material (all chromosomes) is called genome. Particular set of genes in genome is called genotype. The genotype is with later development after birth base for the organism's phenotype, its physical and mental characteristics, such as eye color, intelligence etc.

Reproduction:

During reproduction, first occurs recombination (or crossover). Genes from parents form in some way the whole new chromosome. The new created offspring can then be mutated. Mutation means, that the elements of DNA are a bit changed. This changes are mainly caused by errors in copying genes from parents. The fitness of an organism is measured by success of the organism in its life.

Search Space:

If we are solving some problem, we are usually looking for some solution, which will be the best among others. The space of all feasible solutions (it means objects among those the desired solution is) is called search space (also state space). Each point in the search space represent one feasible solution. Each feasible solution can be "marked" by its value or fitness for the problem. We are looking for our solution, which is one point (or more) among feasible solutions - that is one point in the search space. The looking for a solution is then equal to a looking for some extreme (minimum or maximum) in the search space. The search space can be whole known by the time of solving a problem, but usually we know only a few points from it and we are generating other points as the process of finding solution continues.

The problem is that the search can be very complicated. One does not know where to look for the solution and where to start. There are many methods, how to find some suitable solution (ie. not necessarily the best solution), for example hill climbing, tabu search, simulated annealing and genetic algorithm. The solution found by this methods is often considered as a good solution, because it is not often possible to prove what is the real optimum.

Problem definition:

The clustering strategy limits the number of nodes in each cluster,S. The clustering aims to associate every node with one cluster. Here clustering has been done using improved K-means algorithm which takes into account the node distance from given cluster center, energy requirement and Fault prone.

The objective is to propose a fault tolerant approach in wireless sensor networks for target tracking application with Genetic algorithm for optimal resource utilization. The idea of target tracking is that, the sensor nodes are deployed randomly in a boundary and based on given cluster center, the improved K-means clustering algorithm select the cluster head. Now a node is faulty (transient fault) and the other nodes should take care of the functionality of this node and target should not slip away from the monitored region.

In the first step, for the given nodes and cluster center, the improved k-means clustering algorithm being used[13].

K means is an exclusive clustering algorithm and it is the one of the simplest unsupervised learning algorithms that solve the clustering problem [14]. Wireless Sensor Network has number of nodes, which are randomly scattered over the sensor network. The sensor nodes which are deployed in the sensor network, knows their location information. The coordinates (xi, yi) of each sensor node are used to estimate the distance between two sensor nodes. Based on minimum distance and highest energy, the sensor nodes are clustered by using improved K means clustering algorithm.

GA Operators:

Crossover and mutation provide exploration, compared with the exploitation provided by selection. The effectiveness of GA depends on the trade-off between exploitation and exploration.

Crossover: We use one-point crossover in this paper. The crossover operation takes place between two consecutive individuals with probability specified by *crossover rate*. These two individuals exchange portions that are separated by the crossover point The following is an example of crossover:

Indv1: 1 0 1 0 0 1 0 1

Indv2: 1 0 1 1 ↑ 1 1 1 0

Crossover point

After crossover, two offspring are created as below:

Child1: 1 0 1 0 1 1 1 0

Child2: 1 0 1 1 0 1 0 1

Mutation: As discussed earlier, the mutation operator is applied to each bit of an individual with a probability of *mutation rate*. When applied, a bit whose value is 0 is mutated into 1 and vice versa. An example of mutation is as follows.

Indv: 1 1 0 1 1 1 1
↓ ↓

Indv: 1 1 1 0 1 1 0

Selection: The selection process chooses the candidate individuals based on their fitnesses from the population in the current generation. Otherwords, if the chromosome is better, then the chances of getting selected is higher. Some of the selection methods being used are roulette wheel selection, Rank selection, steady-state selection etc. Proportional selection (or roulette wheel selection) is used in this algorithm. It is implemented by using a biased roulette wheel, where each individual is assigned a slot, the size of which is proportional to the fitness value. Those individuals with higher fitness values are more likely to be selected as the individuals of population in the next generation.

Fitness Evaluation:

The main criteria we need to consider is to find out nodes with most residual energy ($E_{n)}$) and least fault prone ($F_{n)}$) in a given cluster.

$F(x) = max(E_n) + min(F_n);$

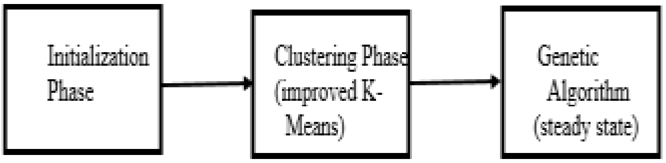

Initialization Phase
Clustering Phase (improved K-Means)
Genetic Algorithm (steady state)

Fig.1. Process Flow

In the first step, select a cluster center with their xi, yi coordinates. Then calculate the distance between each sensor node and the selected cluster center and also get the energy of each node. The node which is nearer to the cluster center
with maximum energy becomes the cluster head and other associated nodes which are nearer to this cluster head than other cluster center becomes part of this cluster. This step is repeated for setting up of all initial clusters.

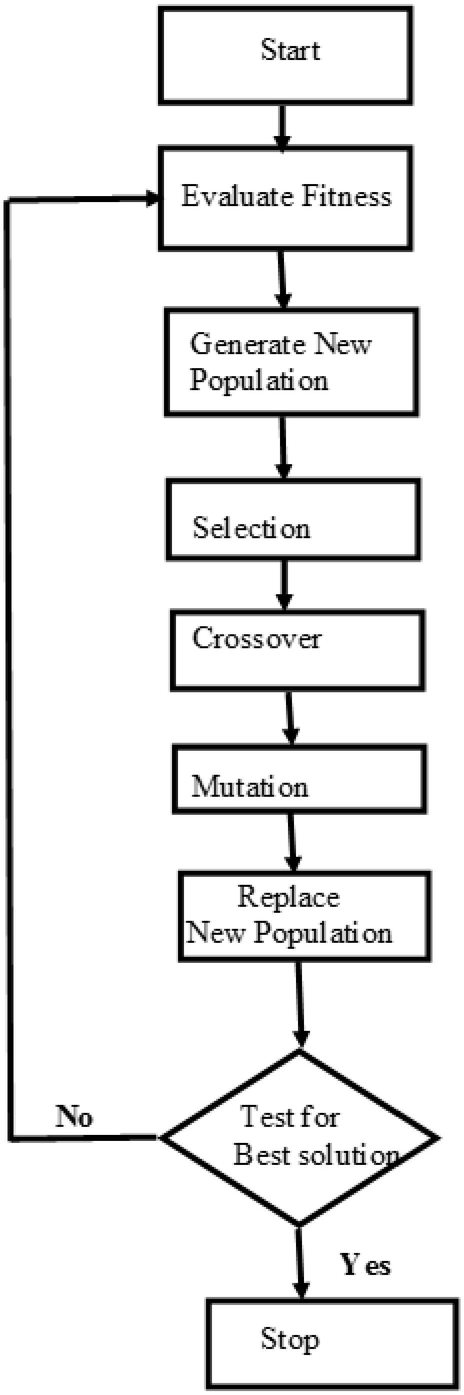


Fig.2. Flow Diagram

The distance between reference nodes is computed by using this formula,

Distance=√(x2-x1)^2+(y2-y1)^2 (2)

Where, (x1, y1) and (x2, y2) are the coordinates of the reference node.

Cluster head selection:

After the formation of cluster, re-compute the calculated distance of each cluster member with cluster head and based on maximum residual energy and least distance with less fault prone, the cluster member get ranking. This will be used when the cluster head gets down when their energy level threshold value becomes less than the fixed threshold value.

Object tracking with Fault Tolerance:

Once Object tracking started, If a cluster head gets down due to energy depletion, the next cluster head within that cluster being selected which satisfied energy requirement and less fault prone using GA. This cluster head change process has been continued till all nodes are exhausted within that cluster. By this process, we can do load balancing which improves network lifetime, avoid energy wastage for new cluster head election and reduce the probability of fault happened due to energy depletion. If an object is missed due to fault, target tracking can be recovered by alerting the neighboring clusters based on the velocity of the moving object. GA algorithm optimally uses the sensor nodes inside the cluster and also include the faulty nodes if repaired. Here the repair is possible since the fault is assumed to be transient in nature as discussed earlier.

Simulation:

To study the effectiveness of the proposed method, this is simulated and compared against the existing method using MATLAB software. For that, we assume the algorithm which would consider fixed cluster head, means the cluster head won't be changed dynamically as traditional algorithm and the one which would change the cluster head in the improved k-means approach with simulated annealing algorithm. For simplicity for this case, we assume the same track needs to be sensed and the same set of nodes is used for tracking purpose, if available. After iterating for different number of time periods, the energy consumption of nodes differ significantly for these 3 methods as shown in the graph below.

TABLE I: PARAMETERS SETTING

Simulation Parameters	**Value**
Population Size	90
Selection type	Proportional Selection
Cross Over rate	0.60
Cross Over type	One point
Mutation rate	0.0065
Generation Size	500

Performance Evaluation:

Here the performance is being evaluated with traditional clustering scheme with varying sensor nodes and can be deduced that improved k-means approach with simulated annealing algorithm gives better performance.

A. Energy Consumption: The energy consumption rate is reduced in Genetic algorithm based approach when compared to similar approach like traditional clustering, simulated annealing etc.

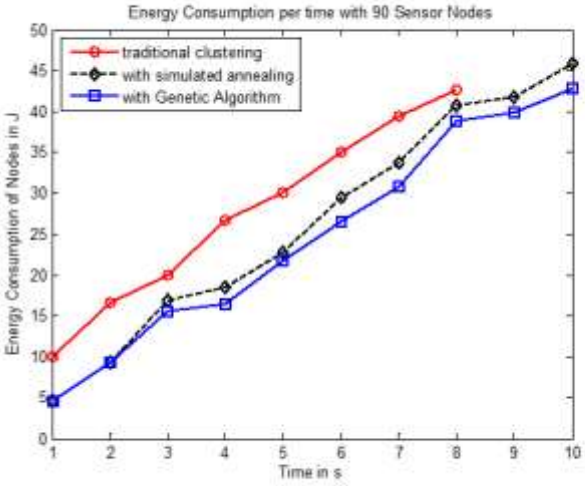


Fig.3. Energy consumption per time with 90 Sensor Nodes

B. Number of Alive Nodes:

The number of alive nodes decreases as the time increases and Genetic algorithm approach perform better when compared to other approaches.

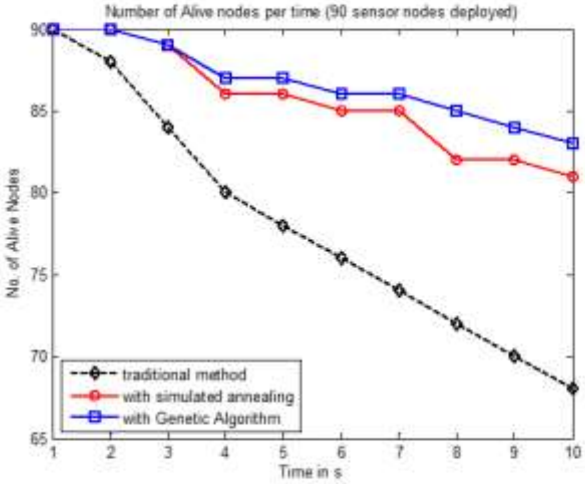


Fig.4. Number of Alive Nodes per time with 90 Sensor Nodes deployed

5. CONCLUSION

In this paper, the option of using Genetic Algorithm with improved k-means clustering algorithm is explored. Besides an idea to use Genetic algorithm to include faulty node that have repaired after transient fault is considered that could improve resource optimization. This would be evaluated on the basis of network lifetime and number of alive nodes. In the future, the method should be explored for other motion models with multiple targets tracking in wireless sensor networks should be studied.

6. REFERENCES:

[1] D. Estrin, R. Govindan, J. Heidemann, and S. Kumar. Next century challenges: scalable coordination in sensor networks. In MobiCom'99: Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking, pages 263–270, New York, NY, USA, 1999. ACM Press.

[2] Y. Xu, J.Winter, W.C Lee "Prediction-based Strategies for Energy Saving in Object Tracking Sensor Networks" In: Proceedings of the 2004 IEEE International Conference on Mobile Data Management (MDM'04) 0-7695-2070-7/04.

[3] S. M. Lee, H. Cha and R. Ha, "Energy-aware location error handling for object tracking applications in wireless sensor networks," Computer Communications, Vol. 30. No 7, pp.1443-1450, May 2007.

[4] Hai Liu, Amiya Nayak, and Ivan Stojmenovi´c,"Fault-Tolerant Algorithms/Protocols in Wireless Sensor Networks",Chapter 10, Springer-Verlag London Limited 2009

[5] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan. Energy-Efficient Communication Protocol for Wireless Micro-sensor Networks. In Proceedings of the Hawaii International Conference on System Science, Maui, Hawaii,2000.

[6] W. R. Heinzelman, A.P. Chandrakasan. An Application-Specific ProtocolArchitecture for Wireless Micro-sensor Network. IEEE Transactions on Wireless Communications, Vol. 1, No. 4, 2002.

[7] J. Tillett, R. Rao, F. Sahin, and T.M. Rao. Cluster-head Identification in Ad hoc Sensor Networks Using Particle Swarm Optimization. In Proceedings of the IEEE International Conference on Personal Wireless Communication, 2002.

[8]R. Ostrosky and Y. Rabani. Polynomial-Time Approxiamation Schemes for Geometric Min-Sum Median Clustering. Journal of the ACM, Vol. 49, No. 2, 2002, pp 139-156.

[9] Y. Xu, J.Winter, W.C Lee "Dual Prediction-based Reporting for Object Tracking Sensor Networks" In: Proceedings of the First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous'04) 0-7695-2208-4/04.

[10] O. Garcia, A. Quintero, and S. Pierre, "Profile-based energy minimisation strategy for Object Tracking Wireless Sensor Networks" Proc. of the IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob'06), pp 372 – 379, Montreal, June 2006.

[11] C. Gui, P. Mohapatra "Power conservation and quality of surveillance in target tracking sensor networks" The 10th Annual International Conference on Mobile Computing and Networking, Philadelphia, USA, September 2004, pp. 129–143.

[12] Rahim Pasha Khajei, Abolfazl Toroghi Haghighat, Jaber Karimpour. Fault-Tolerant Base Station Based Target Tracking in WSNs- 2011 Second International Conference on Intelligent Systems, Modelling and Simulation

[13] Venkatesh S, Dr.K.M.Mehata,"An Improved Fault Tolerant System Using Clustering for Multiple Object Tracking in Wireless Sensor Networks",Int.J.Computer Technology & Applications(IJCTA),Vol 4 (3),May-June 2013,pp.456-461

[14] S. Jerusha, K.Kulothungan and A. Kannan, "LOCATION AWARE CLUSTER BASED ROUTING IN WIRELESS SENSOR NETWORKS", International Journal of Computer & Communication Technology ISSN (PRINT): 0975 - 7449, Volume-3, Issue-5, 2012

[15]Shiyuan Jin, Ming Zhou, and Annie S. Wu "Sensor network optimization using a genetic algorithm", In the Proceedings of the 7th World Multiconference on Systemics, Cybernetics, and Informatics, Orlando, FL, July 2003

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 439 - 445, 2014, ISSN: 2319–8656

A Review on Classification Based Approaches for STE-Ganalysis Detection

Anjani Kumar Verma
SPM College, Department Of
Computer Science,
University Of Delhi
New Delhi, India

Abstract: This paper presents two scenarios of image steganalysis, in first scenario, an alternative feature set for steganalysis based on rate-distortion characteristics of images. Here features are based on two key observations: i) Data embedding typically increases the image entropy in order to encode the hidden messages; ii) Data embedding methods are limited to the set of small, imperceptible distortions. The proposed feature set is used as the basis of a steganalysis algorithm and its performance is investigated using different data hiding methods. In second scenario, a new blind approach of image Steganalysis based on contourlet transform and nonlinear support vector machine. Properties of Contourlet transform are used to extract features of images, the important aspect of this paper is that, it uses the minimum number of features in the transform domain and gives a better accuracy than many of the existing stegananlysis methods. The efficiency of the proposed method is demonstrated through experimental results. Also its performance is compared with the Contourlet based steganalyzer (WBS). Finally, the results show that the proposed method is very efficient in terms of its detection accuracy and computational cost.

Keywords: Steganography, Steganalysis, Contourlet transform, Structural similarity measure, Non linear support vector Machine, MAE, MSE, wMSE, Bayesian

1. INTRODUCTION

Steganography refers to techniques that establish a covert (subliminal) communications channel within regular, innocuous message traffic [1-3]. Steganography is the art and science of hiding secret messages by embedding them into digital media while steganalysis is the art and science of detecting the hidden messages. The goal of a high quality steganography is hiding information imperceptibly not only to human eyes but also to computer analysis.

The obvious purpose of steganalysis is to collect sufficient evidence about the presence of embedded message and to break the security of the carrier. Steganalysis can be seen as a pattern recognition problem also since based on whether an image contains hidden data or not, images can be classified into Stego or Cover image classes.

Steganalysis is broadly classified into two categories. One is meant for breaking a specific steganography. The other one is universal steganalysis, which can detect the existence of hidden message without knowing the details of steganography algorithms used. Universal steganalysis is also known as blind steganalysis and it is more applicable and practicable [4, 5] than the specific steganalysis. Based on the methods used, steganalysis techniques are broadly classified into two classes; signature based steganalysis and statistical based steganalysis. Specific signature based steganalysis are simple, give promising results when message is embedded sequentially, but hard to automatize and their reliability is highly questionable [6, 7]. The first blind steganalysis algorithm to detect embedded messages in images through a proper selection of image quality metrics and multivariate regression analysis was proposed by Avcibas et al. [8, 9]. In universal steganalysis, using statistical methods and identifying the difference of some statistical characteristic between the cover and stego image becomes a challenge. Due to the tremendous increase in steganography, there is a need for powerful blind steganalyzers which are capable of identifying stego images.

In this paper, the focuses are on image steganalysis problem and develop new algorithms based on rate-distortion concepts. In particular, it has been observe the effect of different steganographic methods on image rate-distortion characteristics and construct detectors to separate innocuous cover images from message bearing stego images. This paper proposes a new approach to blind steganalysis does not need any knowledge of the embedding mechanism. This approach utilizes contourlet transform to represent the images. A Gaussian distribution is used to model the contourlet subband coefficients and since skewness and kurtosis of a distribution could be analyzed using the first four moments, the first four normalized statistical moments are considered as the features along with the similarity measure among the medium frequency bands. The experimental results show the efficiency of our approach when analyzed with various steganography methods.

The rest of the paper is organized as below. Section 2 discusses the proposed methods. Section 3 gives Experimental evaluation of the proposed Methods or Steganalyzer with the actual results. At the end, section 5 concludes this paper.

2. PROPOSED METHODS

2.1 METHOD 1

It has been propose novel steganalysis algorithms based on the effect of data hididng process on image rate-distortion characteristics. In particular, we make the following assumptions/observations about the data embedding process:

1. *Data embedding typically increases the image entropy*: In order to encode the hidden messages, steganography methods modify parts of the image data. These modifications typically do not conform with the existing image statistics and therefore result in a net increase in image entropy.

2. *Data embedding methods are limited to the set of small, imperceptible distortions:* Typical steganography methods make only small modifications to ensure perceptual transparency. Perceptually significant parts of the image remain intact.

Stochastic embedding does not result in artifacts with known structures; therefore it is used to develop a generalized method to detect the changes in the rate- distortion characteristics. Since real rate-distortion points for signals with unknown probability distributions –such as natural images – cannot reliably calculated, the data rates achieved by lossy compression scheme. The flowchart of the detection process is seen in Figure. 1.

An image feature extraction phase is followed by a classifier that is trained on relevant data sets. As image features we use the distortion values at different rate points. Mean square error (MSE), mean absolute error (MAE) and weighted mean square error (wMSE) are used as distortion metrics. Here, Bayesian classifier preceded by a KL transform, which reduces the dimensionality of the feature vector.

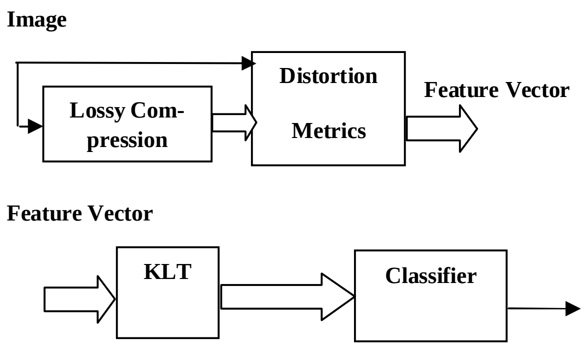


Figure 1. Flow chart describing detection of stochastic embedding

2.1.1 Classifier

Let us denoting w_i as different classes, where each corresponds to a different stego method. This is assuming that $1 \leq i \leq M$ that M such classes exist. This denotes the L dimensional feature vector by x.

$$p(x|w_i) = 1/(2\pi)^{1/2} |\sum_i|^{1/2} \exp(-1/2(x- \mu_i)^T \sum_i^{-1} (x- \mu_i)) \quad (1)$$

where bf μ_i = E[x]is the mean value of the w_i class and $\sum_i$ is the covariance matrix defined as

$$\sum\nolimits_{i =} E [(x-\mu_i) (x-\mu_i)^T] \quad (2)$$

and $|\sum_i|$ denotes the determinant of $\sum_i$ and E[.] denotes the expected value.
It is also define the discriminant function in the logarithmic form as

$$g_i(x) = \ln(p(x|w_i)P(w_i)) \quad (3)$$

$$= -1/2 (x- \mu_i)^T\sum\nolimits_i^{-1} (x- \mu_i)) + \ln(P(w_i)) - 1/2\ln(2\pi) - 1/2\ln(|\sum\nolimits_i |) \quad (4)$$

Assuming equiprobable classes and eliminating constant terms, Eqn.3 can be reduced to

$$g_i(x)= (x-\mu_i)^T\sum\nolimits_i^{-1}(x-\mu_i)+\ln(|\sum\nolimits_i|) \quad (5)$$

μ_i and $\sum_i$ are estimated from the training samples for each class during the training phase.

When the classifier has to operate on a limited number of training samples with relatively small number of classes, the high dimensionality of the problem adversely affects the classifier performance. In particular, the covariance matrix becomes nearly singular and classifications results become sensitive to acquisition noise. A method of reducing the dimensionality of the classification problem while keeping the discriminatory power of the feature vector is to project the feature vector onto a proper subspace.

Let us define the within class and between class scatter matrices, S_w and S_b as,

$$S_w = \sum\nolimits^M_{i=1} P_iE[(x-\mu_i)(x-\mu_i)^T] \quad (6)$$

$$S_b = \sum\nolimits^M_{i=1}P_i(x-\mu_0)(x-\mu_0)^T \quad (7)$$

where μ_0 is the global mean vector

$$\mu_0 = \sum\nolimits^M_{i=1}p_i\mu_i \quad (8)$$

This is further define the scattering matrix criterion J_3 as

$$J_3=\text{trace}\{ S-^1_w S_b \} \quad (9)$$

It can now define a linear projection from the L dimensional feature space to N dimensional sub-space.

$$\hat{A}=C^Tx \quad (10)$$

The optimal projection matix w.r.t. the scattering matrix criterion J_3 is the eigenvectors corresponding to the largest eigenvalues of the system $S-^1_w S_b$. As the individual scatter matrices S_i, the within class scatter matrix may also be ill conditioned. Therefore, in practice it has been used the pseudo-inverse of S_w in the calculations.

2.2 METHOD 2

The objective of the proposed scheme is to select the most relevant features using statistical characteristics of the subband coefficients, thus reduce the dimensionality of feature set and increase the accuracy of detection. In this paper, the first four normalized moments of high frequency, low frequency subband coefficients and structural similarity measure of medium frequency sub band coefficients are taken as the feature set. With these five features, a Non linear Support Vector Machine is trained for further classification. The block diagram of the proposed model is given in Figure 2. The following sub sections briefly explain contourlet transformation and how the feature set is extracted from images.

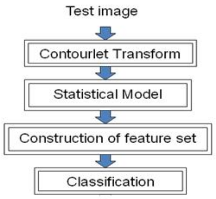


Figure 2. Block diagram for proposed scheme

2.2.1 Contourlet Transform

The Contourlet transform is a two-dimensional extension of the wavelet transform proposed by Do and Vetterli [11, 12] using multiscale and directional filter banks. The contourlet expansion is composed of basis images oriented at various directions in multiple scales with flexible aspect ratio that could effectively capture smooth contours of all images. The contourlet employs an efficient tree structured implementation, which is an iterated combination of Laplacian Pyramid (LP) [13] for capturing the point discontinuities, and the Directional Filter Bank (DFB) [14] to gather nearby basis functions and link point discontinuities into linear structures. Contourlet transform is more powerful than the wavelet transform in characterizing images rich of directional details and smooth contours [15, 16].

Let the image be a real-valued function I(t) defined on the integer valued Cartesian grid [2^l, 2^l]. The Discrete Contourlet Transform with scale j, direction k and level n of I(t) is defined as follows [17,19]:

$$\lambda_{j,k,n}(t) = \sum_{i=0}^{3} \sum_{m \in Z^2} d_k(m) \psi_{j,n}{}^{(i)}(t)$$

where $d_k(m)$ is the directional coefficient and

$$\psi^{(i)}{}_{j,n}(t) = \sum_{m \in Z^2} f_i(m) \phi_{j,n+m}(t)$$

where $\phi(t)$ is the scaling function and f_i is the spatial domain function.

Furthermore, the current existing steganalysis algorithms are limited to the domain of wavelet and DCT transforms. Therefore, identifying stego (constructed by embedding data into their contourlet coefficients) and cover image from the image data set is not easy by these steganalysis algorithms. This fact motivates us to develop efficient steganalysis algorithm in contourlet domain. In this paper, contourlet subband based features are used for steganalysis. *Sub-band Coefficient Modelling* The coefficients in the produced sub bands of contourlet transformed image are very appropriate to obtain the texture feature due to coarse to fine directional details of the image in these sub-bands. Besides, the distribution of the sub-bands coefficients is symmetric and unimodal with mean skewness approximately near to zero, though they have not exactly Gaussian distribution [18]. These special characteristics of subband coefficients make them suitable for modelling by Gaussian distribution with density function.

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty \prec x \prec \infty$$

where μ and σ are the mean and standard deviation of all the coefficients of sub-bands.

2.2.2 Feature Extraction

There are various methods in the literature to extract the relevant features of digital images based on different transforms or filtering techniques. Even though the accuracy of classifiers is based on the number of suitable features, higher the number of features slower will be the classification. So identifying a minimum number of features which can produce efficient classification is a challenge. In this paper, only 5 features have been used which is very less compared to the number of features used in the existing steganalysis. Contourlet transform is more sparser than wavelet as the majority of the coefficients have amplitudes close to zero. Also the moments of contourlet coefficients are more sensitive to the process of information hiding. The first four normalized moments of the high frequency and low frequency subband coefficients are more sensitive to the process of steganography. Since these moments could be a good measure for skewness and kurtosis due to information hiding, the first four normalized moments are extracted as features. Moments are computed as below:

$$m_k = \frac{E(X-\mu)^k}{\sigma^{2k}} \qquad \text{k=1,2,3 and 4.}$$

where X represents the coefficients of contourlet sub bands. Since these moments alone are not sufficient to detect the changes in the medium frequency sub-bands, another feature namely structural similarity measure (SSIM) is also included. For estimating SSIM, medium frequency band is split into two equal number of subband groups X and Y respectively. SSIM includes three parts: Luminance Comparison (LC), Contrast Comparison (CC) and Structural Comparison (SC) and they are defined as below [20, 21, 22]:

$$CC(x,y) = \frac{2\sigma_x\sigma_y}{\sigma^2{}_x + \sigma^2{}_y}$$

$$LC(x,y) = \frac{2\mu_x\mu_y}{\mu^2{}_x + \mu^2{}_y}$$

$$SC(x,y) = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

$$SSIM(X,Y) = [LC(X,Y)][CC(X,Y)][SC(X,Y)]$$

The similarity of the whole image (I) is

$$SSIM \quad (I) = \frac{\sum_{j=1}^{n} SSIM \quad j}{n}$$

where n is the number of middle frequency sub bands in the image. Feature set consists of the first four normalized moments m_k (k=1,2,3,4) and the similarity measure SSIM(I).

2.2.3 *Classification*

A three back propagation Neural Network (NN) is used as a classifier for identifying stego images as well as images [10]. The power of back propagation is that it enables us to compute an effective error for each hidden unit, and thus derive a learning rule for the input to hidden weights. Non linear Support Vector Machine (NSVM) classifier is used for effective classification of stego images and cover images in this work.

3. EXPERIMENTAL RESULTS

3.1 Method 1

The database collection is done through 108 images in Kodak PhotoCD format and spans a large variety of image subjects. In fact, the collection even includes some digitally manipulated images. It is assume that these images have not been modified by steganography software and hence represent the set of cover images.

All images are converted from their original photoCD format to RGB TIFF images at the base resolution of 768X512 pixels. As proposed methods operate on mono-chrome images, only green channel is used. Furthermore, the image is crop to remove black boundary regions (30 and 20 pixels).

3.1.1 *Detection of stochastic embedding*

The process can be modeled by noise addition, without loss of generality. Although, the method allows for alternative noise statistics, this uses a Gaussian noise in this experiment. It uses two different embedding strengths at σ^2 =3 and σ^2 =9, corresponding to PSNR of 41dB and 38dB, and embedding rates of 0.84bpp and 0.91bpp, respectively.

For each image, mean square error, mean absolute error, and weighted mean square error between the image and compressed version are computed. Compression is performed with JPEG2000 at 95, 90, 85, 80, 70, 60, and 50% of the lossless rates.

During training, feature vectors are processed to obtain an optimal projection onto a two dimensional feature space. Then a Bayesian classifier is trained on the reduced features using three classes (namely no embedding, low embedding, and high embedding). In the test phase, the projection matrix obtained in the training phase is used to reduce the feature vector dimensions. Afterwards, classification is performed using the previously learned parameters.

In the whole scenario, 9 cover images out of 54 are mis-labeled as stego-image, while 13 stego-images are mis-labeled as a cover image. Corresponding false alarm and miss rates are 16.7% and 12% respectively.

3.2 Method 2

The proposed steganalysis is implemented using MATLAB 7.6.0 with MATLAB scripts. The experiments are conducted on a personal computer with a 1 GB RAM and P-IV processor. For training we have used 12,200 images from Computer Vision image dataset and INRIA image dataset. It contains 5,500 cover images and 6,700 stego images which are generated by different embedding algorithms like LSB, F5, Contsteg, and YASS. Washington image dataset [22] is used for testing the proposed steganalytic method. 100 images are used to test the proposed scheme, with 60 cover images and 40 stego images.
In order to analyze the proposed method, four typical steganography methods are used. Table 1 gives a comparison of the average detection accuracy between NN classification and non linear support vector classification with same feature set. From this table, one can see that Non linear Support Vector Machine classifies stego images and cover images more accurately. Figure 3. Depicts the performance comparison of NN classifier and NSVM classifier in classifying stego images.

Table 1. Average correct detection rates for natural images and stego images

Steganography Methods	Classifier	Average correct detection rates						
		Embedding rates			Different image size			
		100%	50%	25%	512X512	256X256	128X128	64X64
LSB	NSVM	.945	.922	.904	.952	.976	.937	.902
	NN	.922	.912	.894	.975	.973	.895	.870
F5	NSVM	.950	.971	.977	.957	.951	.931	.894
	NN	.910	.969	.898	.985	.942	.973	.763
ConSteg	NSVM	.928	.917	.908	.905	.957	.903	.823
	NN	.921	.897	.878	.870	.856	.831	.723
YASS	NSVM	.966	.936	.914	.941	.987	.912	.853
	NN	.956	.906	.844	.901	.892	.879	.733

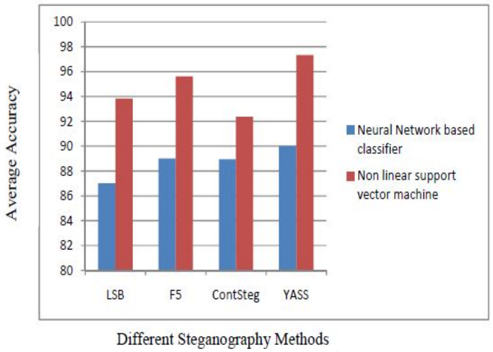


Figure 3. Performance comparison of Neural Network and Non linear Support vector machine based classifiers

The relevancy of the extracted features used in this steganalysis is evaluated using error estimation. Table 2 and Figure 4 display the sample Median Absolute Error (MAE) which exhibits a higher error than bias for all the embedding algorithms. So it is clear that, with this minimum dimensional feature set, proposed method can able to detect the stego image.

Table 2. Median absolute error and bias for the proposed method

Algorithm	MAE	Bias
LSB	$5.91X10^{-3}$	$-1.70X10^{-4}$
F5	$6.63\ X10^{-3}$	$-3.78X10^{-4}$
YASS	$4.19X10^{-3}$	$1.87X10^{-4}$
ContSteg	$3.25X10^{-3}$	$0.58X10^{-4}$

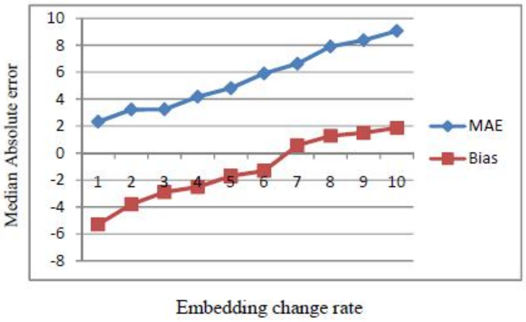


Figure 4. Median Absolute Error (MAE) and Bias of proposed steganalyzer, with respect to embedding rates

The proposed work is compared with the Contourlet-Based Steganalysis (CBS) [23] methods and the results show significant improvement and they are tabulated in Table 3. The Data set used in the proposed scheme for comparison is the Washington dataset which is used in ContSteg [24] and CBS [23].

Table 3. Accuracy of CBS and proposed steganalysis methods on detection of stego-image produced by ContSteg

Secret Data Size (bits)	**Steganalysis Method**	**Average Detection Accuracy (%)**
5000	CBS	59
	Proposed Method	77
10,000	CBS	63
	Proposed Method	89
15,000	CBS	68
	Proposed Method	93

The correct detection rate is improved in the proposed method compared to existing steganalysis schemes [23]. Especially proposed scheme is independent of file formats and image types. The new method based on statistical steganalysis utilizes fewer features than rest of the methods. Hence, it is fast and the computational cost of the new method in extracting the features and detecting the stego image are much less than that of the methods based on feature extraction.

4. CONCLUSIONS

In this paper, it has been proposed new steganalysis techniques based on rate-distortion arguments. These techniques are base on the observation that the steganographic algorithms invariably disturb the underlying statistics therefore change in rate-distortion characteristics of the signals. This is demonstrated the effectiveness of the proposed approach against the stochastic embedding algorithms with varying degrees of success.

On the other hand another approach has been proposed a steganalysis blind detection method based on contourlet transform and non linear support vector machine. This method extracts the statistical moments and structural similarity of the contourlet coefficients as the feature set. The performance of the proposed scheme is illustrated using various testing metrics. The average correct detection rate is improved, at the same time the dimension of the feature set and the average run time is reduced in this proposed scheme. Furthermore, the method proposed here is an universal blind scheme, which is independent of image type and file format.

5. ACKNOWLEDGMENTS

I would like to thanks the experts who are involved in this area for so long time, as well as their valuable resources and also like to indebtness towards my colleagues of their valuable help to write this paper.

6. REFERENCES

[1] R.J. A nderson and F.A. Petitcolas, “On the limits of steganography,” IEEE Journal of selected Areas in Communications 16, pp. 474-481, May 1998. Special issue on copyright & privacy protection.

[2] I.J. Cox, M.L. Miller, and J.A. Bloom, Digital Watermarking, Morgan Kaufmann Publishers, San Francisco, CA, USA, 2002.
S.Katzenbeisser and F.A.P. Petitcolas, eds., Information Hiding: techniques for steganography and digital watermarking, Artech House, Boston, MA, 2000.

[3] J. Fridrich and M. Goljan, “Digital image steganography using stochastic modulation,” in Proc. SPIE: Security and Watermarking of Multimedia Contents V, E.J. Delp and P.W. Wong , eds., E123, pp. 191-202, Jan. 2003.

[4] Fridrich J, Goljan M.(2002) “Practical: Steganalysis of digital images- state of the art. In:” Proceedings of SPIE, Security and Watermarking Multimedia Content IV.Vol. 4675. New York: SPIE, pp 1-13.

[5] McBride B T, Peterson G L, Gustafson S C.(2005) “A new blind method for detecting novel steganography”. Digit Invest, 2: 50-70.

[6] Johnson.N.F, Jajodia.S.:(1998) “Steganalysis: the investigation of hidden information”, In: Proc. IEEE Information Technology Conference, Syracuse, NY.

[7] Fridrich.J, Goljan.M.: Practical steganalysis of digital images state of the art, in: Proc. SPICE Photonics West, Electronic Imaging (2002), Security and watermarking of multimedia contents, San Jose, CA, vol. 4675, January 2002, pp 1-13.

[8] Avcibas I, Memon N D, Sankur B.:(2001) “Steganalysis of watermarking techniques using image quality metrics” In: Proceedings of SPIE, Security and Watermarking of Multimedia Content III, vol. 4314. New York: SPIE, 2001. 523-531.

[9] Avcibas I, Memon N, Sankur B. (2003):” Steganalysis using image quality metrics”. IEEE Trans Image Process, 12: 221- 229.

[10] V. Natarajan and R. Anitha,(2012) “Universal Steganalysis Using Contourlet Transform”, Advances in Intelligent and Soft Computing, Springer – Verlag, Volume 167/2012, 727-735.

[11] Do, M.N., Vetterli, M.: Contourlets (2002) “A directional multiresolution image representation”, Proc. of IEEE Int. Conf. on Image Process., Piscataway, NJ, pp. 357-360.

[12] Minh N.Do, Martin Vertterli. (2006) “The Contourlet Transform: An Efficient Directional Multiresolution Image Representation”, IEEE Transaction on Image Processing, vol.14 no.12,pp.2091-2106.

[13] Burt P.J and Adelson E.H.(1983) “The Laplacian pyramid as a compact image code”, IEEE Trans.Commun,vol 31,no.4, ppl 532-540.

[14] Bamberger R.H and Smith M.J.T.(1992) “A filter bank for the directional decomposition of images: theory and design”, IEEE Trans.Signal Process. Vol.40,n0.4,pp.882-893.

[15] `Yazdi, M., Mahyari, A.G. (2010) “ A new 2D fractal dimension estimation based on contourlet transform for texture segmentation”, The Arabian Journal for Science and Engineering, vol. 35, No. 13, pp.293-317.

[16] Ali Mosleh, Farzad Zargari, Reza Azizi. (2009) “Texture Image Retrieval Using Contourlet Transfrom”, International Symposium on Signal, Circuits and Systems.

[17] M.N.Do, and Vetterli.M. (2006) “Directional multiscale modeling of images using contourlet transform”, IEEE Transactions on Image Processing, Vol.15, no.6,pp.1610-1620.

[18] Chun Ling Yang, Fan Wang, Dongqin Xiao.(2009) “Contourlet Transform based Structural Similarity for image quality assessment”, Intelligent computing and intelligent systems.

[19] http://www.cs.washington.edu/research/imagedatabase.

[20] “Kodak PhotoCD images.” ftp://ftp.kodak.com/www/images/pcd.

[21] “HP Labs LOCO-I/JPEG-LS Home Page.” http://www.hpl.hp.com/loco/.

[22] Mathworks (MATLAB) http://www.mathworks.com

[23] Hedieh Sajedi, Mansour Jamzad.(2008)" A Steganalysis method based on contourlet transform coefficients", International Conference of Intelligent Information Hiding and Multimedia Signal Processing.

[24] Sajedi H., and Jamzad M. "ContSteg: Contourlet-Based Steganography Method", Wireless Sensor Network, Scientific Research Publishing (SRP) in California (US),1(3),163-170.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 434 - 438, 2014, ISSN: 2319–8656

A Fault tolerant system based on Genetic Algorithm for Target Tracking in Wireless Sensor Networks

Venkatesh S
Department of Computer Science and Engineering,
B.S.Abdur Rahman University,
Vandalur, Chennai, India.

K.M.Mehata
Department of Computer Science and Engineering,
B.S.Abdur Rahman University,
Vandalur, Chennai, India.

Abstract- In this paper, we explored the possibility of using Genetic Algorithm (GA) being used in Wireless Sensor Networks in general with specific emphasize on Fault tolerance. In Wireless sensor networks, usually sensor and sink nodes are separated by long communication distance and hence to optimize the energy, we are using clustering approach. Here we are employing improved K-means clustering algorithm to form the cluster and GA to find optimal use of sensor nodes and recover from fault as quickly as possible so that target detection won't be disrupted. This technique is simulated using Matlab software to check energy consumption and lifetime of the network. Based on the simulation results, we concluded that this model shows significant improvement in energy consumption rate and network lifetime than other method such as Traditional clustering or Simulated Annealing.

Keywords: Genetic Algorithm, Wireless sensor network, Fault tolerant, energy efficiency, Network lifetime, K-means clustering, Reliability

1. INTRODUCTION

Nowadays, there are numerous applications in which Wireless Sensor networks being used like Military and civilian as well. This includes target tracking, surveillance, and security management etc. Since a Wireless sensor is a small, lightweight, untethered, battery-powered device, it has limited energy, processing power, memory [15]. Hence, energy consumption is a critical issue in sensor networks. Since these devices are less expensive, It is possible to design and construct the deployment of large scale wireless sensor networks (WSN) with potentially thousands of nodes [1]. WSN and their applications have tremendous potential in both commercial and military environments due to their low cost and pervasive surveillance [2]. For critical environment with high degree of dependability is required, WSN should offer characteristics such as: reliability, availability and maintainability. Availability to a large extend depends on fault tolerance to keep the system working as expected. Availability on the service level means that the service delivered by a WSN (or part of it) is not affected by failures and faults in underlying components such as single nodes or node subsystems. In WSNs, the failure of such components is almost unavoidable. Most of the detection and recovery techniques therefore aim at reducing MTTR (the amount of time required for detecting and recovering from a failure) as much as possible. In these WSN deployments, it is common to have a node providing functionality to its neighbors. In a typical WSN, all data obtained by member sensors must be transmitted to a sink or data collector. More energy will be consumed during transmission, if the communication distance is longer. It is estimated that to transmit a *k*-bit message across a distance of *d*, the energy consumed can be represented as follows:

$$E(k,d)=E_{elec}*k + E_{amp}*k*d^2 \qquad (1)$$

where E_{elec} is the radio energy dissipation and E_{amp} is transmit amplifier energy dissipation. In this scenario, it would be ideal if there is a cluster head and it could aggregate sensor data before it is forwarded to a base station, thereby saving energy. [4] In the cellular networks and ad hoc networks energy requirements is not a constraint as base stations or batteries can be replaced as needed, but nodes in sensor networks have very limited energy and their batteries cannot usually be recharged or replaced due to hostile or hazardous environments. Hence energy saving is one of the important factor in WSN and hence important characteristic of sensor networks is the stringent power budget of wireless sensor nodes. There are two components of a sensor node viz sensing unit and wireless transceiver, usually directly interact with the environment, which is subject to variety of physical, chemical, and biological factors. On such cases, Nodes with stronger hardware capabilities can perform operations for other nodes that would either have to spend a significant amount of energy or would not be capable of performing these operations. These services, however, may fail due to various reasons, including radio interference, de-synchronization, battery exhaustion, or dislocation due to inhospitable conditions. Such failures are caused by software and hardware faults, environmental conditions, malicious behavior, or bad timing of a legitimate action. In general, the consequence of such an event is that a node becomes unreachable or violates certain conditions that are essential for providing a service. Besides, a failure caused by a trivial software bug can be propagated to become a massive failure of the sensor network. In other cases, some sensor nodes became faulty due to transient conditions like high heat due to hostile environment as well. For example, in multiple moving objects tracking which is an active research area in WSNs due to its practical use in a wide variety of applications [3], including military or environmental monitoring applications, may missed out the target being tracked if any one of the sensor nodes that monitoring it becomes faulty. In such scenarios, prediction or node failure in these methods must be handled quickly before target slips far away. For reporting on the other hand, optimization techniques are applicable on different clustering and data fusion methods. Since the number of nodes participate in the tracking process, clustering phase can be the critical phase of the tracking from the energy consumption point of view. After clustering phase is done, it is equally important to use the member nodes in the cluster optimally so that cluster lifetime should be extended which in turn extends the network life time.

The rest of the paper is organized as follows: section II discusses about the assumption and background , section III discusses related work that has been done in WSNs with Fault tolerant and GA perspective, section IV presents the challenges involved in monitoring multiple objects with transient fault in particular and the proposed method and Section V gives the conclusion and the future work.

2. BACKGROUND

A. Assumptions

Following assumptions are made about the sensors and the sensor network in the development of the proposed target tracking algorithm:

- A set of sensors are deployed in a square terrain. The nodes possesses the following properties
- The sensor network is static
- All nodes are assumed to have the capabilities of a cluster-head and the ability to adjust their transmission power based on transmission distance.
- Two nodes communicate with each other directly if they are within the transmission range
- The sensor nodes are assumed to be homogeneous i.e. they have the same processing power and initial energy.
- The sensor nodes are assumed to use different power levels to communicate within and across clusters.
- The sensor nodes are assumed to know their location and the limits S (Number of nodes in each cluster).
- For multiple object tracking, no 2 objects are in same cluster at same time.

3. RELATED WORK

In the paper by Heinzelman et al's paper[5] "Energy-Efficient Communication Protocol for Wireless Micro -sensor Networks" which describes a clustering-based protocol called LEACH. Here the performance of LEACH with direct communication and MTE is being compared. They use a pre-determined optimal number of clusters (5% of the total number of nodes) in their simulations. In the paper [6] Heinzelman et al determine that the optimal number of clusters for a 100-node network to be 3-5 by using a computation and communication energy model; however, determining the optimal number of cluster-heads depends on several factors such as sensor densities, the position of a sink, etc.

In the paper [7] by Tillett et al, the PSO (Particle Swarm Optimization) approach is proposed which would divide the sensor node field into groups of equal sized groups of nodes. PSO is an evolutionary programming technique that mimics the interaction of ants or termites to find a good solution. Although partitioning into equal sized clusters balances the energy consumption of cluster heads, this method is not applicable to some networks where nodes are not evenly distributed.

In the paper [8] by Ostrosky et al, address a somewhat different partitioning problem: Given n points in a large data set, partition this data set into k (k is known) disjoint clusters so as to minimize the total distance between all points and the cluster-heads to which they belong. The authors use a polynomial-time approximation scheme to solve the problem.

A small number of nodes are selected to become clusterheads. They are responsible for coordinating the nodes in their clusters, for instance by collecting data from them and forwarding it to the base station. In case that a cluster head fails, no messages of its cluster will be forwarded to the base station any longer. The cluster head can also intentionally or due to software bugs forward incorrect information. Depending on the application case, the impact of such a failure can vary from quality degradation of measurements to alarm messages not being delivered to a back-end system. While forwarding messages, nodes can aggregate data from multiple other nodes in order to reduce the amount of data sent to the base station. One common simple approach is to calculate the average of correlated measured values such as temperature, humidity and pressure, sending only one message to the back-end. If a node generates incorrect data, the data aggregation results can suffer deviations from the real value. Also, if a node responsible for generating the aggregated data is subject to a value failure, the base station will receive incorrect information of an entire region of the network.

This paper [9] investigates prediction-based approaches for performing energy efficient reporting in object tracking sensor networks. A dual prediction-based reporting mechanism (called DPR) has been proposed, in which both sensor nodes and the base station predict the future movements of the mobile objects. Transmissions of sensor readings are avoided as long as the predictions are consistent with the real object movements. DPR achieves energy efficiency by intelligently trading off multi-hop/long-range transmissions of sensor readings between sensor nodes and the base station with one-hop/short-range communications of object movement history among neighbor sensor nodes. The impact of several system parameters and moving behavior of tracked objects on DPR performance has been explored, and also two major components of DPR are studied: prediction models and location models through simulations. In this paper, Profile-Based Algorithm (PBA) [10] has been proposed that aims to use the information contained in the network and in the object itself to optimize energy consumption, thus extending lifetime. Here it utilizes the regularity in the object's behavior to reduce energy consumption. In this paper [11], In Target Tracking application, the sensor nodes collectively monitor and track the movement of an event or target object. The network operations have two states: the surveillance state during the absence of any event of interest, and the tracking state which is in response to any moving targets. Thus, the power saving operations, which is of critical importance for extending network lifetime, should be operative in two different modes as well. In this paper, we study the power saving operations in both states of network operations. During surveillance state, a set of novel metrics for quality of surveillance, which suggests that atleast p-sesnor nodes required to cover any location, is proposed specifically for detecting moving objects. In the tracking state, we propose a collaborative messaging scheme that wakes up and shuts down the sensor nodes with spatial and temporal preciseness. In this paper [12], a novel approach toward Base Station (BS) oriented clustering and tracking in WSNs is introduced. Proposed method overlooks ad-hoc ability of WSNs to earn energy efficiency and fault tolerance. BS is a powerful energy and computational resource, therefore, BS is burdened with major part of clustering and tracking operations. 3-D cubic antenna is used to enable our sensors to receive BS packets from long distance. Also, BS has a good knowledge of nodes energy level, as a result, BS rotates activated nodes and CH to avoid load balancing problem.

4. PROPOSED SOLUTION

Genetic Algorithm:

Chromosome:

All living organisms consist of cells. In each cell there is the same set of chromosomes. Chromosomes are strings of DNA and serves as a model for the whole organism. Usually a chromosome consist of genes which is basically blocks of DNA. Each gene encodes a particular protein. Basically it can be said, that each gene encodes a trait, for example color of eyes. Possible settings for a trait (e.g. blue, brown etc) are called alleles. Each gene has its own position in the chromosome. This position is called locus.

Complete set of genetic material (all chromosomes) is called genome. Particular set of genes in genome is called genotype. The genotype is with later development after birth base for the organism's phenotype, its physical and mental characteristics, such as eye color, intelligence etc.

Reproduction:

During reproduction, first occurs recombination (or crossover). Genes from parents form in some way the whole new chromosome. The new created offspring can then be mutated. Mutation means, that the elements of DNA are a bit changed. This changes are mainly caused by errors in copying genes from parents. The fitness of an organism is measured by success of the organism in its life.

Search Space:

If we are solving some problem, we are usually looking for some solution, which will be the best among others. The space of all feasible solutions (it means objects among those the desired solution is) is called search space (also state space). Each point in the search space represent one feasible solution. Each feasible solution can be "marked" by its value or fitness for the problem. We are looking for our solution, which is one point (or more) among feasible solutions - that is one point in the search space. The looking for a solution is then equal to a looking for some extreme (minimum or maximum) in the search space. The search space can be whole known by the time of solving a problem, but usually we know only a few points from it and we are generating other points as the process of finding solution continues.

The problem is that the search can be very complicated. One does not know where to look for the solution and where to start. There are many methods, how to find some suitable solution (ie. not necessarily the best solution), for example hill climbing, tabu search, simulated annealing and genetic algorithm. The solution found by this methods is often considered as a good solution, because it is not often possible to prove what is the real optimum.

Problem definition:

The clustering strategy limits the number of nodes in each cluster,S. The clustering aims to associate every node with one cluster. Here clustering has been done using improved K-means algorithm which takes into account the node distance from given cluster center, energy requirement and Fault prone.

The objective is to propose a fault tolerant approach in wireless sensor networks for target tracking application with Genetic algorithm for optimal resource utilization. The idea of target tracking is that, the sensor nodes are deployed randomly in a boundary and based on given cluster center, the improved K-means clustering algorithm select the cluster head. Now a node is faulty (transient fault) and the other nodes should take care of the functionality of this node and target should not slip away from the monitored region.

In the first step, for the given nodes and cluster center, the improved k-means clustering algorithm being used[13].

K means is an exclusive clustering algorithm and it is the one of the simplest unsupervised learning algorithms that solve the clustering problem [14]. Wireless Sensor Network has number of nodes, which are randomly scattered over the sensor network. The sensor nodes which are deployed in the sensor network, knows their location information. The coordinates (xi, yi) of each sensor node are used to estimate the distance between two sensor nodes. Based on minimum distance and highest energy, the sensor nodes are clustered by using improved K means clustering algorithm.

GA Operators:

Crossover and mutation provide exploration, compared with the exploitation provided by selection. The effectiveness of GA depends on the trade-off between exploitation and exploration.

Crossover: We use one-point crossover in this paper. The crossover operation takes place between two consecutive individuals with probability specified by *crossover rate*. These two individuals exchange portions that are separated by the crossover point The following is an example of crossover:

Indv1: 1 0 1 0 0 1 0 1

Indv2: 1 0 1 1 ↑ 1 1 1 0

Crossover point

After crossover, two offspring are created as below:

Child1: 1 0 1 0 1 1 1 0

Child2: 1 0 1 1 0 1 0 1

Mutation: As discussed earlier, the mutation operator is applied to each bit of an individual with a probability of *mutation rate*. When applied, a bit whose value is 0 is mutated into 1 and vice versa. An example of mutation is as follows.

Indv: 1 1 0 1 1 1 1

↓ ↓

Indv: 1 1 1 0 1 1 0

Selection: The selection process chooses the candidate individuals based on their fitnesses from the population in the current generation. Otherwords, if the chromosome is better, then the chances of getting selected is higher. Some of the selection methods being used are roulette wheel selection, Rank selection, steady-state selection etc. Proportional selection (or roulette wheel selection) is used in this algorithm. It is implemented by using a biased roulette wheel, where each individual is assigned a slot, the size of which is proportional to the fitness value. Those individuals with higher fitness values are more likely to be selected as the individuals of population in the next generation.

Fitness Evaluation:

The main criteria we need to consider is to find out nodes with most residual energy ($E_{n)}$ and least fault prone ($F_{n)}$ in a given cluster.

$F(x)= max(E_n) + min(F_n)$;

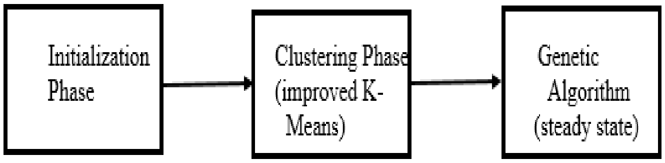

Initialization Phase
Clustering Phase (improved K-Means)
Genetic Algorithm (steady state)

Fig.1. Process Flow

In the first step, select a cluster center with their xi, yi coordinates. Then calculate the distance between each sensor node and the selected cluster center and also get the energy of each node. The node which is nearer to the cluster center with maximum energy becomes the cluster head and other associated nodes which are nearer to this cluster head than other cluster center becomes part of this cluster. This step is repeated for setting up of all initial clusters.

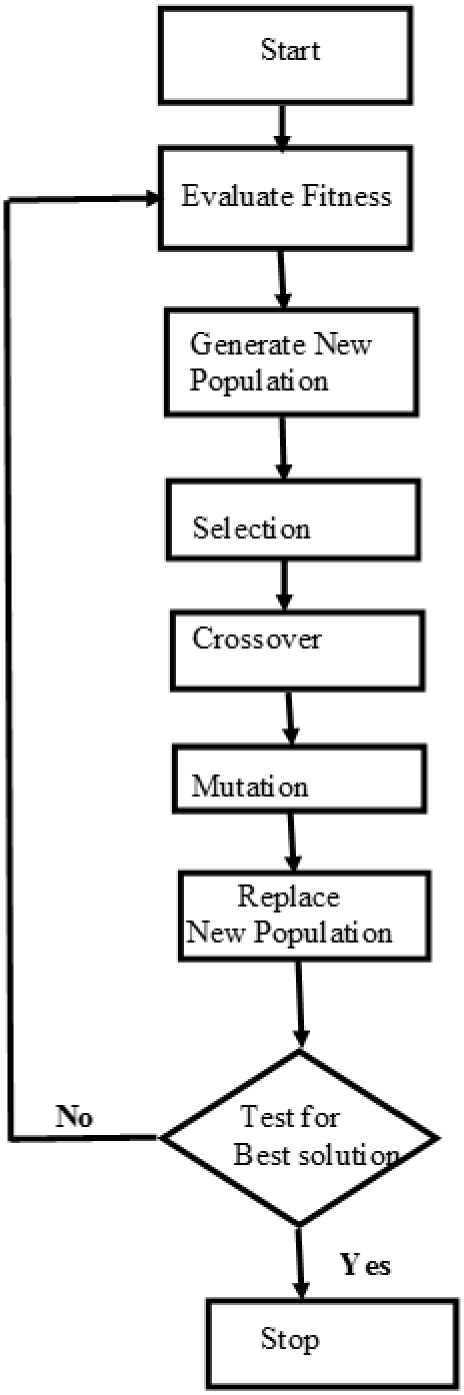


Fig.2. Flow Diagram

The distance between reference nodes is computed by using this formula,

Distance=√(x2-x1)^2+(y2-y1)^2 (2)

Where, (x1, y1) and (x2, y2) are the coordinates of the reference node.

Cluster head selection:

After the formation of cluster, re-compute the calculated distance of each cluster member with cluster head and based on maximum residual energy and least distance with less fault prone, the cluster member get ranking. This will be used when the cluster head gets down when their energy level threshold value becomes less than the fixed threshold value.

Object tracking with Fault Tolerance:

Once Object tracking started, If a cluster head gets down due to energy depletion, the next cluster head within that cluster being selected which satisfied energy requirement and less fault prone using GA. This cluster head change process has been continued till all nodes are exhausted within that cluster. By this process, we can do load balancing which improves network lifetime, avoid energy wastage for new cluster head election and reduce the probability of fault happened due to energy depletion. If an object is missed due to fault, target tracking can be recovered by alerting the neighboring clusters based on the velocity of the moving object. GA algorithm optimally uses the sensor nodes inside the cluster and also include the faulty nodes if repaired. Here the repair is possible since the fault is assumed to be transient in nature as discussed earlier.

Simulation:

To study the effectiveness of the proposed method, this is simulated and compared against the existing method using MATLAB software. For that, we assume the algorithm which would consider fixed cluster head, means the cluster head won't be changed dynamically as traditional algorithm and the one which would change the cluster head in the improved k-means approach with simulated annealing algorithm. For simplicity for this case, we assume the same track needs to be sensed and the same set of nodes is used for tracking purpose, if available. After iterating for different number of time periods, the energy consumption of nodes differ significantly for these 3 methods as shown in the graph below.

TABLE I: PARAMETERS SETTING

Simulation Parameters	**Value**
Population Size	90
Selection type	Proportional Selection
Cross Over rate	0.60
Cross Over type	One point
Mutation rate	0.0065
Generation Size	500

Performance Evaluation:

Here the performance is being evaluated with traditional clustering scheme with varying sensor nodes and can be deduced that improved k-means approach with simulated annealing algorithm gives better performance.

A. **Energy Consumption:** The energy consumption rate is reduced in Genetic algorithm based approach when compared to similar approach like traditional clustering, simulated annealing etc.

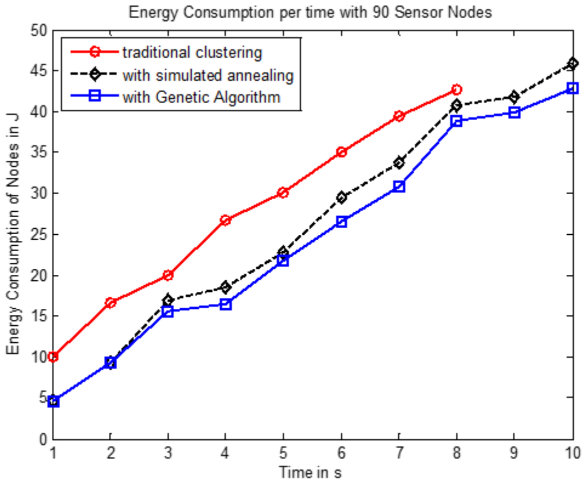


Fig.3. Energy consumption per time with 90 Sensor Nodes

B. **Number of Alive Nodes:**

The number of alive nodes decreases as the time increases and Genetic algorithm approach perform better when compared to other approaches.

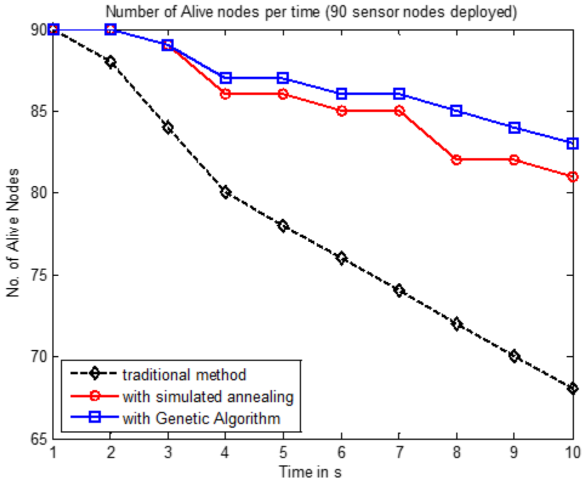


Fig.4. Number of Alive Nodes per time with 90 Sensor Nodes deployed

5. CONCLUSION

In this paper, the option of using Genetic Algorithm with improved k-means clustering algorithm is explored. Besides an idea to use Genetic algorithm to include faulty node that have repaired after transient fault is considered that could improve resource optimization. This would be evaluated on the basis of network lifetime and number of alive nodes. In the future, the method should be explored for other motion models with multiple targets tracking in wireless sensor networks should be studied.

6. REFERENCES:

[1] D. Estrin, R. Govindan, J. Heidemann, and S. Kumar. Next century challenges: scalable coordination in sensor networks. In MobiCom’99: Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking, pages 263–270, New York, NY, USA, 1999. ACM Press.

[2] Y. Xu, J.Winter, W.C Lee “Prediction-based Strategies for Energy Saving in Object Tracking Sensor Networks” In: Proceedings of the 2004 IEEE International Conference on Mobile Data Management (MDM’04) 0-7695-2070-7/04.

[3] S. M. Lee, H. Cha and R. Ha, “Energy-aware location error handling for object tracking applications in wireless sensor networks,” Computer Communications, Vol. 30. No 7, pp.1443-1450, May 2007.

[4] Hai Liu, Amiya Nayak, and Ivan Stojmenovi´c,"Fault-Tolerant Algorithms/Protocols in Wireless Sensor Networks",Chapter 10, Springer-Verlag London Limited 2009

[5] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan. Energy-Efficient Communication Protocol for Wireless Micro-sensor Networks. In Proceedings of the Hawaii International Conference on System Science, Maui, Hawaii,2000.

[6] W. R. Heinzelman, A.P. Chandrakasan. An Application-Specific Protocol Architecture for Wireless Micro-sensor Network. IEEE Transactions on Wireless Communications, Vol. 1, No. 4, 2002.

[7] J. Tillett, R. Rao, F. Sahin, and T.M. Rao. Cluster-head Identification in Ad hoc Sensor Networks Using Particle Swarm Optimization. In Proceedings of the IEEE International Conference on Personal Wireless Communication, 2002.

[8]R. Ostrosky and Y. Rabani. Polynomial-Time Approxiamation Schemes for Geometric Min-Sum Median Clustering. Journal of the ACM, Vol. 49, No. 2, 2002, pp 139-156.

[9] Y. Xu, J.Winter, W.C Lee “Dual Prediction-based Reporting for Object Tracking Sensor Networks” In: Proceedings of the First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous’04) 0-7695-2208-4/04.

[10] O. Garcia, A. Quintero, and S. Pierre, “Profile-based energy minimisation strategy for Object Tracking Wireless Sensor Networks” Proc. of the IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob’06), pp 372 – 379, Montreal, June 2006.

[11] C. Gui, P. Mohapatra “Power conservation and quality of surveillance in target tracking sensor networks” The 10th Annual International Conference on Mobile Computing and Networking, Philadelphia, USA, September 2004, pp. 129–143.

[12] Rahim Pasha Khajei, Abolfazl Toroghi Haghighat, Jaber Karimpour. Fault-Tolerant Base Station Based Target Tracking in WSNs- 2011 Second International Conference on Intelligent Systems, Modelling and Simulation

[13] Venkatesh S, Dr.K.M.Mehata,"An Improved Fault Tolerant System Using Clustering for Multiple Object Tracking in Wireless Sensor Networks",Int.J.Computer Technology & Applications(IJCTA),Vol 4 (3),May-June 2013,pp.456-461

[14] S. Jerusha, K.Kulothungan and A. Kannan, "LOCATION AWARE CLUSTER BASED ROUTING IN WIRELESS SENSOR NETWORKS”, International Journal of Computer & Communication Technology ISSN (PRINT): 0975 - 7449, Volume-3, Issue-5, 2012

[15]Shiyuan Jin, Ming Zhou, and Annie S. Wu "Sensor network optimization using a genetic algorithm", In the Proceedings of the 7th World Multiconference on Systemics, Cybernetics, and Informatics, Orlando, FL, July 2003

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 451 - 456, 2014, ISSN: 2319–8656

Classification and Searching in Java API Reference Documentation

Monali Metkar
K.K.W.I.E.E.R
Nashik,
University Of Pune, India

S. M. Kamalapur
K.K.W.I.E.E.R
Nashik,
University Of Pune, India

Abstract: Application Program Interface (API) allows programmers to use predefined functions instead of writing them from scratch. Description of API elements that is Methods, Classes, Constructors etc. is provided through API Reference Documentation. Hence API Reference Documentation acts as a guide to user or developer to use API's. Different types of Knowledge Types are generated by processing this API Reference Documentation. And this Knowledge Types will be used for Classification and Searching of Java API Reference Documentation.

Keywords: API, API Reference Documentation, Knowledge Types, Classification, Searching.

1. INTRODUCTION

An Application Programming Interface (API) is a set of commands, functions, and protocols. It also specifies the interaction between the software components. In most procedural languages, an API specifies a set of functions or routines that accomplish a specific task or are allowed to interact with a specific software component. For example, consider following Constructor in Java:

AbstractAction()

Whenever user or developer is referring to an API and has planned to use it for specific purpose API Reference documentation works as a guide. API Reference Documentation is an important part of programming with APIs and it complements the API by providing information about the API. So, it plays a crucial role in how developers learn and use an A PI, and developers will expect the information about API elements they should find therein By considering the above example, if new developer wishes to use "AbstractAction()" constructor in Java Program, he can refer to API Reference Documentation of Java and he will find the description of "AbstractAction()" constructor in Constructor Summary as:

Creates an Action.

In above example Java Documentation is considered and Java APIs are documented through Javadocs which is a set of web pages such that one for each package or type in the API.

To enhance the quality of API reference documentation and the efficiency with which the relevant information it contains can be accessed, it's necessary to first understand its contents by analyzing it. Therefore, to reason about the quality and value of Java API reference documentation, focus should be about what knowledge it contains. Because Knowledge refers to retrieve useful information from data and then use this knowledge for specific purpose. By analyzing the contents of Java API Reference Documentation, Knowledge is generated and this knowledge can be categorized further.

Previous work focused separately on Studies of Knowledge Categorization and of API reference Documentation and Knowledge retrieval was done based on Experience, Observations and Analysis.

So proposed system focuses on generation of Knowledge Types, classification of API Reference Document according to Knowledge Types and also on searching depending upon Knowledge Types.

Section 2 focuses on Literature Review. Section 3 gives Implementation Details with Block Diagram, Concept with Example and Algorithms are highlighted in Section 4. Data Set, Results obtained and Performance Measure are discussed in Section 5 of Results. The paper ends with concluding remarks.

2. LITERATURE REVIEW

The previous work mainly focused on the Knowledge Categorization and API Reference Documentation Separately.

2.1 API Reference Documentation

Study of documentation needs for a domain-specific API, using surveys and interviews of developers was done by Nykaza et al.[6] This study identified, among other requirements and the importance of an overview section in API documentation.

Jeong et al. [10] conducted a lab study with eight participants to assess the documentation of a specific service-oriented architecture. This study identified 18 guidelines they believe would lead to increased documentation quality for the system under study, including "explaining starting points" for using the API.

Robillard and DeLine [9] identified the obstacles faced by developers when trying to learn new APIs through surveys and interviews with Microsoft developers. The study showed that many obstacles were related to aspects of the documentation, but did not include the systematic analysis of API documentation content.

Similarly, Shi et al. [8] studied API documentation evolution. The authors apply data mining techniques over the source repository of five open-source APIs. Their study provides various quantitative measures of which parts of the API documentation are most frequently revised, and how often API documentation is changed consistently with the corresponding elements.

2.2 Knowledge Categorization based on Manual Methods

Researchers have applied Knowledge from one field to other field, they also studied which are the different questions raised in Software Project Development.

Mylopoulos et al.[5] discussed how knowledge representation techniques from the field of Artificial Intelligence can be applied to software engineering. The authors presented a categorization of different knowledge types, presumably derived from their experience.

Requirement and Design are the important stages in Software Project Development. Herbsleb and Kuwana[4] classified questions asked in design meetings to study the kinds of knowledge that may benefit from explicit capture at the requirements and design stages based on their general experience.

Hou et al.[2] studied 300 questions related to two specific Swing widgets (JButton and JTree) posted on the Swing forum. They then mapped the questions to the different design features of the widgets. Their classification focuses more on the target of the question and less on discovering the different types of knowledge provided to and sought by API users.

More recently, Ko et al.[1] observed 17 developers at Microsoft for a 90 minutes session each, studying their information needs as they perform their software engineering tasks. From the observation data the authors collected 334 specific information needs, which they abstracted into 21 general information needs.

Kirk et al.[3] investigated the knowledge problems faced by them and their students when trying to develop applications by extending the JHotDraw framework.

Similarly to Ko et al.'s study, Sillito et al.[7] produced a catalog of 44 types of questions developers ask during software evolution tasks. The questions in the catalog do not focus exclusively on API usage, but rather relate to software evolution and maintenance tasks.

So, researchers focused on how different stages of Software Project Development and tools required can be analyzed in different ways and they classified the Questions raised in different phases into different categories based on their Experience, Observations. Knowledge Types was not generated automatically.

Here, authors referred and studied API Reference Documentation in different ways. So, Separate study of Knowledge Categorization and API Reference Documentation was done previously.

The proposed work focuses on generation of Knowledge Types from Java API Reference Documentation, Classification and Searching in Java API Reference Documentation.

3. IMPLEMENTATION DETAILS

3.1 Block Diagram of the System

The following figure explains the Block Diagram of Proposed System:

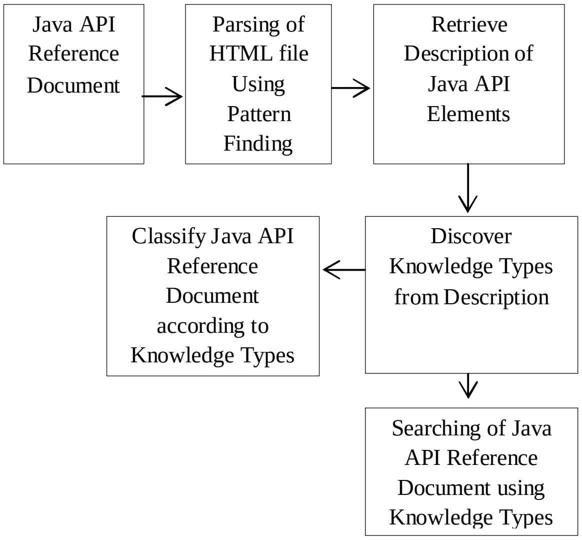

Java API Reference Document
Parsing of HTML file Using Pattern Finding
Retrieve Description of Java API Elements
Discover Knowledge Types from Description
Classify Java API Reference Document according to Knowledge Types
Searching of Java API Reference Document using Knowledge Types

Figure 1: Block Diagram of the System

The System focuses on Java API Reference Documents that is Javadocs. Input to the system is API Reference Document of Java which is HTML Page.
This Java API Reference Document is then parsed by finding Pattern for the Tags having Description.
After finding the Patterns, the required Text is retrieved from the HTML page.
In the next step, Description of API elements is analyzed and then Knowledge Types will be generated.

Following are the Knowledge Types that are to be generated:

1. Functionality and Behavior: This Knowledge Type describes functionality and features of API. And also specifies what happens when the API is used.
e.g.: protected boolean enabled
Specifies whether action is enabled; the default is true.

2. Directives: It is related to accessibility that is what users are allowed or not allowed to do with the API element. Directives

are clear contracts.
e.g.: public class AccessException extends RemoteException
An AccessException is thrown by certain methods of the java.rmi.Naming class (specifically bind, rebind, and unbind) and methods of the java.rmi.activation.ActivationSystem interface to indicate that the caller does not have permission to perform the action requested by the method call. If the method was invoked from a non-local host, then an AccessException is thrown.

3. Control-Flow: How the API (or the framework) manages the flow of control is described by this knowledge type. For example by stating what events cause a certain callback to be triggered?
e.g.: Set<String> getSupportedAnnotationTypes()
If the processor class is annotated with SupportedAnnotationTypes, return an unmodifiable set with the same set of strings as the annotation.

4. Code Examples: Code examples are provided for how to use and combine elements to implement certain functionality or design outcomes.
e.g.: public abstract class AbstractExecutorService extends Object implements ExecutorService
Provides default implementations of ExecutorService execution methods. This class implements the submit, invokeAny and invokeAll methods using a RunnableFuture returned by newTaskFor, which defaults to the FutureTask.
class provided in this package. For example, the implementation of submit(Runnable) creates an associated RunnableFuture that is executed and returned. Subclasses may override the newTaskFor methods to return RunnableFuture implementations other than FutureTask.
Extension example. Here is a sketch of a class that customizes ThreadPoolExecutor to use a CustomTask class instead of the default FutureTask:

```
 public class CustomThreadPoolExecutor extends
ThreadPoolExecutor {
  static class CustomTask<V> implements
RunnableFuture<V> {...}
  protected <V> RunnableFuture<V>
newTaskFor(Callable<V> c) {
      return new CustomTask<V>(c);
  }
  protected <V> RunnableFuture<V> newTaskFor(Runnable
r, V v) {
      return new CustomTask<V>(r, v);
  }
  // ... add constructors, etc.
 }
```

5. Environment: Aspects related to the environment in which the API is used is described in this type, but not the API directly, e.g., compatibility issues, differences between versions, or licensing information.
e.g: public abstract class AbstractElementVisitor7<R,P> extends AbstractElementVisitor6<R,P>
A skeletal visitor of program elements with default behavior appropriate for the RELEASE_7 source version.

6.External References: It includes any pointer to external documents, either in the form of hyperlinks, tagged ”see also” reference, or mentions of other documents (such as standards or manuals).
e.g: public interface DOMLocator
DOMLocator is an interface that describes a location (e.g. where an error occurred).
See also the Document Object Model (DOM) Level 3 Core Specification.

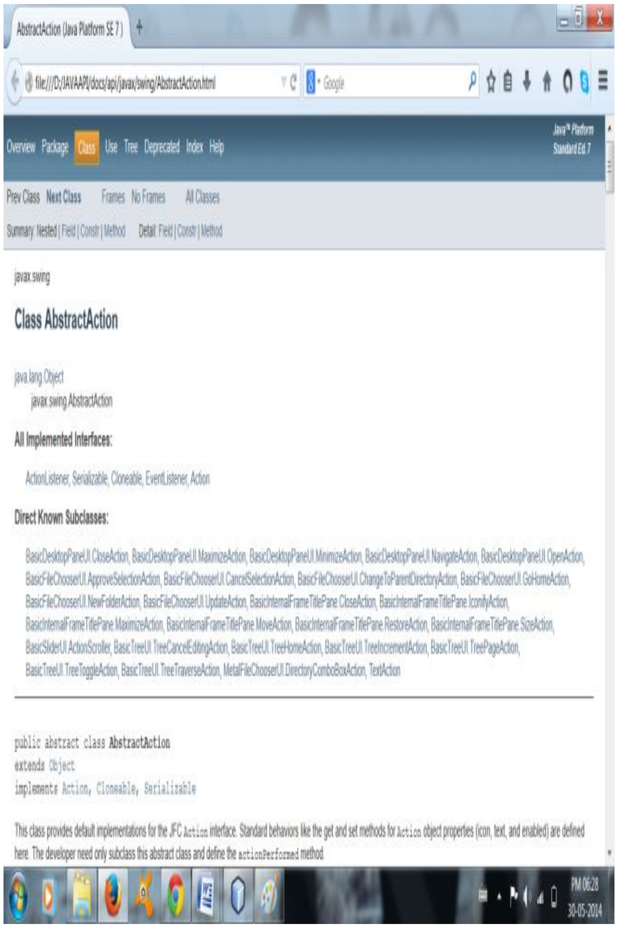
7. Non-information: A section of documentation containing any complete sentence or self-contained fragment of text that provides only uninformative boilerplate text.
e.g: DefinitionKindHelper()

After generating the Knowledge Types, the given Java API Reference Document is classified according to the Knowledge Types generated for class.

Also, searching of Document is done depending upon the Knowledge Types.

3.2 Concept in detail with example:

a) Consider following HTML file as Input: In this example, one of the class of Javadocs , named void AbstractAction is taken into consideration.


AbstractAction (Java Platform SE 7)
file:///D:/JAVAAPI/docs/api/javax/swing/AbstractAction.html
Google
Overview Package Class Use Tree Deprecated Index Help
Java™ Platform Standard Ed. 7
Prev Class Next Class Frames No Frames All Classes
Summary: Nested | Field | Constr | Method Detail: Field | Constr | Method
PM 06:28
30-05-2014

javax.swing

Class AbstractAction

java.lang.Object
    javax.swing.AbstractAction

All Implemented Interfaces:

ActionListener, Serializable, Cloneable, EventListener, Action

Direct Known Subclasses:

BasicDesktopPaneUI.CloseAction, BasicDesktopPaneUI.MaximizeAction, BasicDesktopPaneUI.MinimizeAction, BasicDesktopPaneUI.NavigateAction, BasicDesktopPaneUI.OpenAction, BasicFileChooserUI.ApproveSelectionAction, BasicFileChooserUI.CancelSelectionAction, BasicFileChooserUI.ChangeToParentDirectoryAction, BasicFileChooserUI.GoHomeAction, BasicFileChooserUI.NewFolderAction, BasicFileChooserUI.UpdateAction, BasicInternalFrameTitlePane.CloseAction, BasicInternalFrameTitlePane.IconifyAction, BasicInternalFrameTitlePane.MaximizeAction, BasicInternalFrameTitlePane.MoveAction, BasicInternalFrameTitlePane.RestoreAction, BasicInternalFrameTitlePane.SizeAction, BasicSliderUI.ActionScroller, BasicTreeUI.TreeCancelEditingAction, BasicTreeUI.TreeHomeAction, BasicTreeUI.TreeIncrementAction, BasicTreeUI.TreePageAction, BasicTreeUI.TreeToggleAction, BasicTreeUI.TreeTraverseAction, MetalFileChooserUI.DirectoryComboBoxAction, TextAction

```
public abstract class AbstractAction
extends Object
implements Action, Cloneable, Serializable
```

This class provides default implementations for the JFC `Action` interface. Standard behaviors like the get and set methods for `Action` object properties (icon, text, and enabled) are defined here. The developer need only subclass this abstract class and define the `actionPerformed` method.

Figure 2: Example of the API Reference Document

b) Parsing of HTML document is done using following technique:
In this technique, initially all HTML Tags are fetched from

the HTML Page.
After fetching all the HTML Tags, the Tags having the required descriptions are observed.
And then the required Text is retrieved from the all HTML Tags.

For Example: To get the Description of Class , all HTML Tags are observed.
And then Pattern is detected as: Description of the Class is present under div tag having identity as <div class="block">.
But here, there will be multiple div tags in one HTML page with same class=”block”.

So again, pattern is observed in all HTML pages of Java API Reference Documents as: Description of the Class is always present in the First tag having class=”block”.
And then Text is retrieved from this tag.
So after this First div tag, multiple div tags with class= “block” may be present.

c) After separating the tags having the description, next step is to generate Knowledge Types for the given API Reference Document. Here description of one API elements of the API Reference Document may fall under more than one Knowledge Types.

To generate Knowledge Types, identity of each Knowledge Type is observed.
For Example:
For generating the Functionality and Behavior Knowledge Type, Description of API Elements is considered as it is. Because Functionality and Behavior Knowledge Type describes functionality and features of API. And also specifies what happens when the API is used.

For above class, Following Knowledge Types are generated:
For Description of class: External References, Environment, Functionality and Behavior.
For Constructor of class: Functionality and Behavior.
For Fields, Methods of class: Control Flow, Functionality and Behavior.

d) After generating the Knowledge Types, depending upon the Knowledge Types generated for class the given Java API Reference Document is classified into respective Knowledge Type category.

So here for above example of AbstractAction , this document will be classified in External References, Environment ,Functionality and Behavior.

e) Knowledge Types generated will be used for searching.
For above example of AbstractAction , user can search for “Code Example of AbstractAction” for getting Example of AbstractAction.

3.3 Algorithms

3.3.1 Parsing of HTML Files to fetch Description:

a) Initially, one of the Javadocs pages that is Java API Reference Document which is to be processed is taken as input.
b) Source code of the Javadocs is HTML tags and hence the actual input to the first step is HTML and JavaScript tags.
c) After taking HTML tags as input, the next step is to parse the HTML tags to fetch the tags having description.
d) So, to fetch the description of API element from the current page using Pattern Finding.

3.3.2 For Generation of Knowledge Types for the Description of API elements:

a) After fetching the description of API elements in second step, next step is to process this description.
b) To process the description of the API elements, the patterns of the Descriptions are observed, that is whether the descriptions are having some common words in them or they are starting with same words or having some common format.
c) So, after finding some common patterns in the descriptions, the Knowledge Types are generated.
That is description will be classified to the appropriate Knowledge Type. For Example, Description of all API elements will have common Knowledge Type as Functionality and Behaviour.

3.3.3 For classification of Java API Reference Document:

a) Here the Knowledge Types generated for Class are observed first.
b) Depending on the Knowledge Types generated for Class , the document will be classified into the respective Knowledge Types.

3.3.4 Searching using Knowledge Types

a) After classification of Documentation in above step, the searching will be performed.
b) Here depending upon query given by user, the query will be parsed and the searching of the query will be done depending upon the Knowledge Types.

4. RESULTS

4.1 Data Set

The Data Set for the system are set of API Reference Documents. The jdk-7u51-apidocs.zip file contains the set of API Reference Documents for Java. The above said file can be obtained by using following link:
http://www.oracle.com/technetwork/java/javase/documentation/java-se-7-doc-download-435117.html

4.2 Results

4.2.1 Results of Classification:

Consider the API Reference Documentation in Figure 2 for Class AbstractAction.

For API Reference Document in Figure 2, the document will be classified as follows:

Table 1: Results obtained for AbstractAction and for other Classes

Sr. No.	Class Name	Knowledge Types generated Of Class	Classification of Document Into Following Knowledge Types
1	AbstractAction (Java Platform SE 7)	External References, Environment, Functionaliy and Behavior	External References ,Environment ,Functionaliy and Behavior
2	AbstractAnnotationValueVisitor6 (Java Platform SE 7)	Environment, Functionality and Behavior	Environment, Functionality and Behavior
3	BoxLayout (Java Platform SE 7)	External References, Environment, Control Flow, Functionality and Behavior	External References, Environment, Control Flow, Functionality and Behavior
4	ButtonGroup (Java Platform SE 7)	External ReferencesEnvironmentFunctionality and Behavior	External ReferencesEnvironmentFunctionality and Behavior
5	CellRenderer Pane (Java Platform SE 7)	External ReferencesCode ExampleEnvironmentFunctionality and Behavior	External ReferencesCode ExampleEnvironmentFunctionality and Behavior
6	DefaultButton Model (Java Platform SE 7)	External ReferencesEnvironmentFunctionality and Behavior	External ReferencesEnvironmentFunctionality and Behavior
7	DefaultListCellRenderer (Java Platform SE 7)	External ReferencesEnvironmentControl FlowFunctionality and Behavior	External ReferencesEnvironmentControl FlowFunctionality and Behavior
8	DefaultRowSorter (Java Platform SE 7)	External ReferencesControl FlowFunctionality and Behavior	External ReferencesControl FlowFunctionality and Behavior

4.2.2 Results of Searching:

When user wish to search query like “Code Examples of AbstractAction” the results will show that respective document.

So depending upon user requirement the Searching is done.

4.2.3 Performance Measure:

Following table shows Performance Measures:

Table 2: Performance Measure for different documents tested by system

Java API Reference Document Tested	Knowledge Types Expected	Knowledge Types Generated	Precision
AbstractAction (Java Platform SE 7)	4	4	1.00
AbstractAnnotationValueVisitor6 (Java Platform SE 7)	2	2	1.00
BoxLayout (Java Platform SE 7)	4	4	1.00
ButtonGroup (Java Platform SE 7)	4	3	0.75
CellRendererPane (Java Platform SE 7)	6	5	0.83
DefaultButtonModel (Java Platform SE 7)	3	3	1.00
DefaultListCellRenderer (Java Platform SE 7)	5	4	0.80
DefaultRowSorter (Java Platform SE 7)	3	3	1.00
GroupLayout (Java Platform SE 7)	3	3	1.00
ImageIcon (Java Platform SE 7)	4	3	0.75

5. CONCLUSION AND FUTURE WORK

API's are used as interface for using predefine functions, packages, classes etc. Developers read API reference documentation to learn how to use the API and answer specific questions they have during development tasks. Thus API Reference Documentation provides guide to user for referring to API. API Reference Documentation contains description of API elements; this description will be analyzed for generating Knowledge. This system focuses on classification of description of API elements into different Knowledge Types for Java API Reference Documentation. After generating Knowledge Types, classification of the Java API Reference Document is done according to Knowledge Types. Java API Reference Document then can be searched using the Knowledge Types. Other Types of Documentation like MSDN, Documentation of Python can be considered for parsing and processing further.

6. REFERENCES

[1] A. J. Ko, R. DeLine, and G. Venolia, "Information needs in collocated software development teams," in Proceedings of the 29th International Conference on Software Engineering, 2007, pp. 344–353.

[2] D. Hou, K. Wong, and J. H. Hoover, "What can programmer questions tell us about frameworks?" in Proceedings of the 13th International Workshop on Program Comprehension, 2005, pp. 87–96.

[3] D. Kirk, M. Roper, and M. Wood, "Identifying and addressing problems in object-oriented framework reuse," Empirical Software Engineering, vol. 12, pp. 243–274, June 2007.

[4] J. D. Herbsleb and E. Kuwana, "Preserving knowledge in design projects: what designers need to know," in Proceedings of the Joint INTERACT '93 and CHI '93 Conferences on Human Factors in Computing Systems, 1993, pp. 7–14.

[5] J. Mylopoulos, A. Borgida, and E. Yu, "Representing software engineering knowledge," Automated Software Engineering, vol. 4, no. 3, pp. 291–317, 1997.

[6] J. Nykaza, R. Messinger, F. Boehme, C. L. Norman, M. Mace,and M. Gordon, "What programmers really want: Results of a needs assessment for SDK documentation," in Proceedings of the 20th Annual ACM SIGDOC International Conference on Computer Documentation, 2002, pp. 133–141.

[7] J. Sillito, G. C. Murphy, and K. D. Volder, "Asking and answering questions during a programming change task," IEEE Transactions on Software Engineering, vol. 34, no. 4, pp. 434–451, July-August 2008.

[8] L. Shi, H. Zhong, T. Xie, and M. Li, "An empirical study on evolution of API documentation," in Proceedings of the Conference on Fundamental Approaches to Software Engineering, 2011, pp. 416–431.

[9] M. P. Robillard and R. DeLine, "A field study of API learning obstacles," Empirical Software Engineering, vol. 16, no. 6, pp. 703–732, 2011.

[10] S. Y. Jeong, Y. Xie, J. Beaton, B. A. Myers, J. Stylos, R. Ehret, J. Karstens, A. Efeoglu, and D. K. Busse, "Improving documentation for eSOA APIs through user studies," in Proc. 2nd Int'l Symp. on End-User Development, ser. LNCS, vol. 5435. Springer,2009, pp. 86–105.

[11] Walid Maalej and Martin P. Robillard , Patterns of Knowledge in API Reference Documentation, IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 39, NO. X, XXXXXXX 2013.

A Dependent Set Based Approach for Large Graph Analysis

Shital Deshmukh
University of Pune,
KKWIEER,
Nashik, India.

S. M. Kamalapur
University of Pune,
KKWIEER,
Nashik, India

Abstract: Now a day's social or computer networks produced graphs of thousands of nodes & millions of edges. Such Large graphs are used to store and represent information. As it is a complex data structure it requires extra processing. Partitioning or clustering methods are used to decompose a large graph. In this paper dependent set based graph partitioning approach is proposed which decomposes a large graph into sub graphs. It creates uniform partitions with very few edge cuts. It also prevents the loss of information. The work also focuses on an approach that handles dynamic updation in a large graph and represents a large graph in abstract form.

Keywords: Clustering, Graph Partitioning, Large Graph, Sub Graph.

1. INTRODUCTION

Large graph is one which consists of hundreds to thousands of nodes and millions of edges. Web graphs, social networks, recommendation systems are some examples of large graph. As it is a complex data structure such graphs require excessive processing, more memory for storage and knowledge of a pattern of the graph. It is very difficult to comment on exact size and pattern of a large graph as it changes with time. In large graph analysis the first step is to divide the input graph into number of small parts called as sub graph as whole graph cannot fit into memory for processing at given time and second step is graph summarization which finds the strong connected component i.e a node which is connected to maximum nodes in the sub graph. All such components are then used to maintain connection between different sub graphs by using hierarchical representation. For the first step many serial and parallel graph partitioning methods like spectral bisection, multilevel partitioning, and incremental partition are proposed so far.

Graph partitioning problem complexity is NP complete. For any graph partitioning method to be the best or efficient it must answer following questions:

1. What is the threshold value of partition for given graph?
2. How the connection between sub graphs is maintained?

Some algorithms fail to answer both the questions. For example spectral bisection method produces excellent partitions but connection between sub graphs is difficult to maintain as it is matrix based approach and partitions are stored in matrix form, Multilevel partitioning method is a K –way partitioning method which does not provide threshold value for number of partitions to be produced. The proposed method focuses on both the aspects i.e threshold value and connection between sub graphs.

For the second step, CEPS summarization method is commonly used which uses random walk with restart concept to find connected component/vertex of a graph but it's a matrix based approach so it is not scalable for large graph. The proposed graph partitioning method calculates this connected component/vertex while producing partitions of a large graph. One more issue in large graph analysis is graph size changes with time because information in social network, web graphs which best explains large graphs changes with time. So, dynamic updation like addition or deletion of information in produced sub graphs should also be handled.

So proposed system focuses on implementation of two methods one is graph partitioning and other for dynamic updations in large graph.

Section 2 focuses on literature review, section 3 explains block diagram, algorithms of the proposed approaches, data sets for the proposed method and results are briefed in section 4, and section 5 concludes the paper.

2. LITERATURE REVIEW

This chapter covers related work done on large graph analysis i.e different graph partitioning methods.

2.1 Graph Partitioning Methods

The graph partition problem is , Let graph G = (V, E), with V vertices and E edges, it is possible to form sub graphs or partitions of G into smaller components with some properties also called as k-way partitioning which divides the vertex set into k smaller components or sub graphs. A good partition is one in which the number of edge cuts are less and uniform graph partition is one which divides graph into equal size sub graphs.

Spectral bisection partitioning [8] method is a matrix based approach in which for a given a graph with adjacency matrix A, where A_{ij} gives an edge between node i and j, and Degree matrix D, is a diagonal matrix, in which each diagonal entry of a row i, d_{ii}, represents the degree of node i. The Laplacian of matrix L is defined as $L = D – A$, then a partition for graph $G = (V, E)$ is defined as a partition of set V into disjoint sets U, and W, such that cost of cut $(U, W)/ (|U|\cdot|W|)$ is minimum. The second smallest eigenvalue (λ) of L gives a lower bound of the optimal cost (c) of partition where $c \geq \lambda/n$.

The eigenvector (*V*) corresponding to *λ*, which is called as Fiedler vector, bisects the graph into only two sub graphs based on the sign of the corresponding vector entry. To do the division into a larger number of sub graphs is usually achieved by repeated bisection, but this does not always give satisfactory results which is a drawback of the method also minimum cut partitioning fails when the number of sub graphs to be formed, or partition sizes are unknown.

Multilevel partitioning method is analogous to multigrid method to solve numerical problems. Karypis and Kumar has proposed k-way graph partitioning known as METIS [4] which is based on multilevel partitioning in which the proposed method reduces the size of the graph by collapsing vertices and edges, partitions the graph into smaller graph, and then uncoarsen it to construct a partition for the original graph. The drawback is the graph partitions are stored in adjacency matrix, as it uses static data structure to store partitions node or edge addition or deletion in sub graphs (partitions) at run time is not possible.

To execute several scientific and engineering applications parallel, requires the partitioning of data or among processors to balance computational load on each node with minimum communication. To achieve this parallel graph partitioning there are many algorithms like geometric, structural, spectral & refinement algorithms are proposed. One of such method is parallel incremental graph partitioning [2] in which recursive spectral bisection-based method is used for the partitioning of the graph which needs to be updated as the graph changes over time i.e a small number of nodes or edges may be added or deleted at any given instant. The drawback of the method is initial partition is to be calculated using linear programming based bisection method.

The Proposed approach focuses on uniform partitions creation with no loss of information.

3. IMPLEMENTATION DETAILS

3.1 Block Diagram of the System

The following figure explains the Block Diagram of Proposed System:

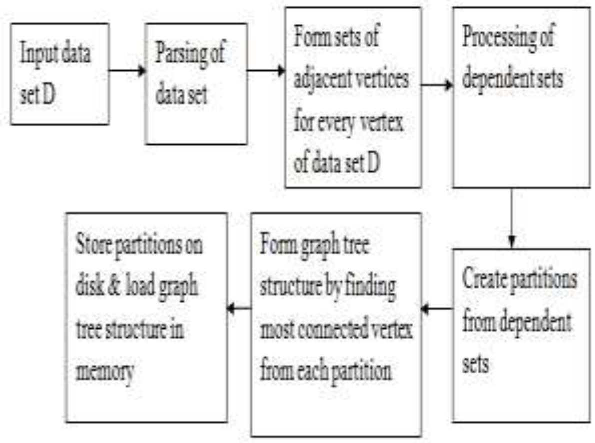

Input data set D
Parsing of data set
Form sets of adjacent vertices for every vertex of data set D
Processing of dependent sets
Create partitions from dependent sets
Form graph tree structure by finding most connected vertex from each partition
Store partitions on disk & load graph tree structure in memory

Fig 1. Block Diagram of Proposed Method

Here the Input to the system is a data set D consisting of Set of connected vertices. The proposed system will directly form sub graphs i.e partitions of input data which is different from previous work.

3.2 Proposed Approach with Example

Dependent Set: For a given vertex dependent set is set of all vertices connected to it.

Consider the following graph G:

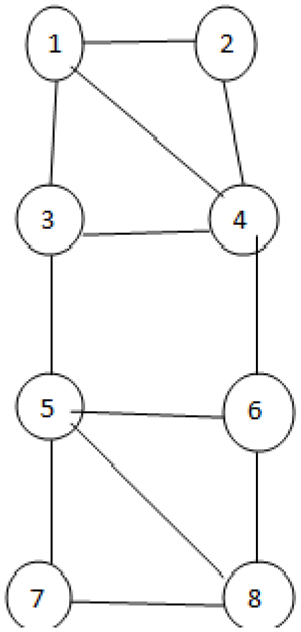


Fig 2. Undirected Graph

The dependent sets are:
Vertex 1 = {2, 3, 4}
Vertex 2 = {1, 4}
Likewise calculate all dependent sets
After calculating all dependent sets generate uniform partitions of given graph by processing dependent sets.

3.3 Dependent Set Based Graph Partitioning Algorithm

The proposed graph partitioning method consists of following steps:

1. Read the input data set D.
2. Parse the data set i.e arrange it in one order (Ascending / Descending).
3. Calculate the sets of adjacent vertices for every vertex from input data set. These sets are called as dependent sets.
4. Calculate the size of each dependent set, process and analyze the sets to calculate threshold value of number of partitions
5. Calculate the partitions for sets by considering largest set first till all the vertices of data set does not get covered in any of the partition.
6. Store these partitions and dependent sets on the disk.
7. Form a hierarchical representation of all partitions by taking most connected vertex of every partition.
8. Store this representation that is tree on the disk.

3.4 Algorithm for Dynamic Large Graph Analysis

This section explains the working of proposed algorithm to perform operations on large graph dynamically. The steps of the algorithm are:

1. Traverse the graph – Tree from Super Graph i.e root node till the Leaf Super Node of required partition.
2. Select the Leaf Super Node, its corresponding partition will be loaded in memory and shown on the system.
3. Now add or delete any node or edge in the partition.
4. Once the updation is done the store the partition again on the disk and update the corresponding leaf node information.

4. Result

4.1 Data Set

To analyse the performance of the proposed methods following data sets are used

1. DBLP Data Set: It is a database of Computer Science publications which represents an authorship graph in which every graph node represents an author and the edge represents co-author relationship.

2. Social Networks: Twitter
http://www.socialcomputing.asu.edu/dataset/Twitter

4.2 Expected Results

Following table shows the expected results of proposed graph partitioning method on given data:

Table 1. Expected Results

No. of Nodes	**No. of Edges**	**No. of Partitions**
8	11	2
10	17	2
12	22	3
25	100	4
50	300	5

5. CONCLUSION

The paper concludes that, the main issue in large graph analysis is to decompose it into sub graph. The existing graph portioning methods requires excessive processing and some are not scalable for large graph. The proposed graph partitioning method addresses the issue of limited main memory by storing the partitions on the disk. All existing approaches work on static large graph the method proposed here also addresses dynamic updation in large graph.

6. REFERENCES

[1] C. Faloutsos, K.S. McCurley, and A. Tomkins, “Fast Discovery of Connection Subgraphs,” Proc. ACM 10th Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD), pp. 118-127, 2004.

[2] Chao-Wei Ou and Sanjay Ranka “Parallel Incremental Graph Partitioning”IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 8, NO. 8, AUGUST 1997

[3] C.R. Palmer and C. Faloutsos, “Electricity Based External Similarity of Categorical Attributes,” Proc. Seventh Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 486-500, 2003

[4] G. Karypis and V. Kumar, “Multilevel Graph Partitioning Schemes,” Proc. IEEE/ACM Conf. Parallel Processing, pp. 113-122, 1995.

[5] G. Kasneci, S. Elbassuoni, and G. Weikum, “Ming: Mining Informative Entity Relationship Subgraphs,” Proc. 18th ACM Conf. Information and Knowledge Management (IKM), pp. 1653- 1656, 2009

[6] H. Tong and C. Faloutsos, “Center-Piece Subgraphs: Problem Definition and Fast Solutions,” Proc. 12th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 404-413, 2006.

[7] J.F. Rodrigues Jr., H. Tong, A.J.M. Traina, C. Faloutsos, and J.Leskovec, “Large Graph Analysis in Gmine System” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013

[8] Stephen T. Barnard and Horst D. Simon.A fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems. In Proceedings of the sixth SIAM conference on Parallel Processing for Scientific Computing, pages 711–718, 1993.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 460 - 463, 2014

Study of Different Multi-instance Learning kNN Algorithms

Rina S. Jain,
Department of Computer Engineering,
K.K.W.I.E.E.R., Nashik,
University of Pune, India

Abstract: Because of it is applicability in various field, multi-instance learning or multi-instance problem becoming more popular in machine learning research field. Different from supervised learning, multi-instance learning related to the problem of classifying an unknown bag into positive or negative label such that labels of instances of bags are ambiguous. This paper uses and study three different k-nearest neighbor algorithm namely Bayesian -kNN, citation -kNN and Bayesian Citation -kNN algorithm for solving multi-instance problem. Similarity between two bags is measured using Hausdroff distance. To overcome the problem of false positive instances constructive covering algorithm used. Also the problem definition, learning algorithm and experimental data sets related to multi-instance learning framework are briefly reviewed in this paper.

Keywords: Bayesian kNN, citation kNN, constructive covering algorithm, Machine learning, Multi-instance problem

1. INTRODUCTION

According to ambiguity in training data, machine learning is roughly categorized into three frameworks-Supervised, Unsupervised and Reinforcement learning. Unlike to supervised learning where all training instances are with known labels, in multi-instance learning the labels of the training instances are unknown; different to unsupervised learning where all training instances are without known labels, in multi-instance learning the labels of the training bags are known and different from reinforcement learning where the labels of the training instances are delayed, in multi-instance learning there is no delay. Figure 1 shows ambiguity spectrum for learning framework. The prediction tasks for MI problems are more difficult than those for single-instance learning problems because the real cause of the bag label is ambiguous. It has been shown that learning algorithms ignoring the characteristics of multi-instance problems, such as popular decision trees and neural networks, could not work well in this scenario [1].

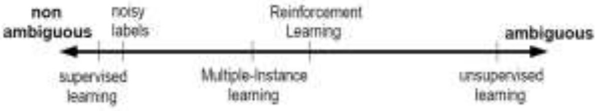
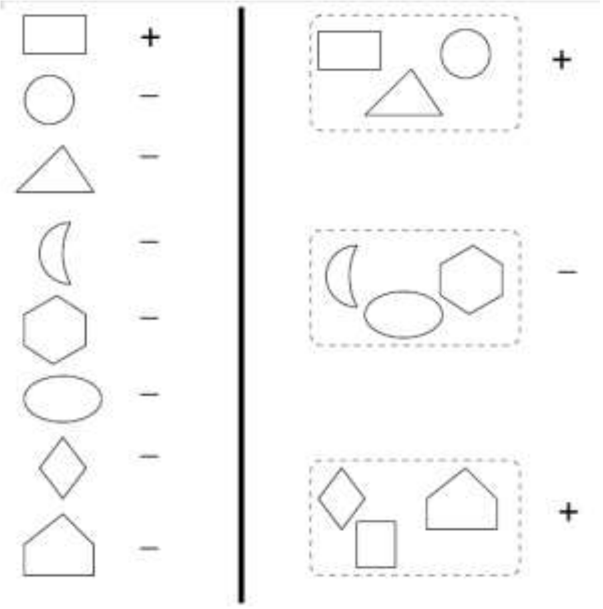


Figure. 1 Ambiguity spectrum (by O. Maron et al. [3])

In multiple-instance learning, the input consists of labeled examples (called bags) consisting of multisets of instances, each described by an attribute vector, training set comprises of such bags and task is to predict the labels of unobserved bags. Figure 2 (by S Ray et al.) illustrates the relationships between standard supervised learning and multiple-instance learning to get clear idea about them. Consider the example "whether figure is rectangle or contain at least rectangle", Label can be positive (+) or negative (-), as for two-class classification problem. (a) In supervised learning, each example (geometric figure) is labeled. A possible concept that explains the example labels shown is "the figure is a rectangle". (b) In MI learning, bags of examples are labeled. A possible concept that explains the bag labels shown is "the bag contains at least one figure that is a rectangle."



a) Supervised Learning b) Multi-Instance Learning

Figure. 2 Example depicting Relationship between supervised and multi- instance learning

So, in MIL, if bag comprise of at least one positive instance which represent the output then bag is labeled as positive. If bag consists of all negative instances then it labeled as negative.

MIL has received considerable amount of attention due to both its theoretical interest and type of representation fits for a number of real-world learning scenarios e.g. drug activity prediction[1]**,**text categorization[7], image classification[18],object detection in images[14],content based image classification[15],visual tracking [16], computer security[17] ,web mining[10],spam filtering[9] etc.

Rest of paper is organized as follows. Section 2 presents survey of literature along with pros and cons of some the existing methods. MIL algorithm with the main contribution of this paper is described in section 3. Section 4 reports the data sets and results. Finally, section 5 summarizes this paper and raises issues for future work.

2. RELATED WORK

The multi-instance concept was first formally introduced by Dietterich et al. [1]. It was originally inspired by a drug activity prediction problem. In this task, a molecule can have several conformations (i.e. shapes) with different properties that result in the molecule being of "musk" or "non-musk" type. However, it is unidentified which particular conformation is the cause of a molecule being of the "musk" type. The conventional single-instance setting (Supervised learning) cannot represent this application problem properly as one molecule may have several other conformations. Therefore, Dietterich et al. [1] proposed the multi-instance setting, in which each sample is represented by a collection of single instances instead of a single instance and also made an asymmetric assumption regarding the process that decides class labels of bag based on instances in the bag. Many algorithms use that as standard assumption. DD algorithm is projected by Maron et al. [2] that includes a concept point that describes a portion of instance space that is dense w.r.t. instances from positive bags. A few years later, DD algorithm stretched further by adding it with the EM (Expectation-Maximization) algorithm, resulting in EM-DD algorithm proposed by Zhang et al. [5]. DD and its extension EMDD have been used on various MIL problems such as stock selection, drug activity prediction, natural scene classification and image retrieval.

In 2002, to solve MIL problems, Andrews et al. [7] advises two methods to exploit the standard Support Vector Machine. The aim was to point out the maximum-margin multiple-instance separating hyper plane in which at least one positive instance from all positive bags was placed on the other side of hyper plane and all instances in each negative bag were located on other side. In 2006, Zhang et al. [8] recommended RBF-MIP algorithm, which is derived from the well-known Radial Basis Function (RBF).

Wang et al. [4] suggested a lazy learning approach using kNN algorithm that in turn uses Hausdorff distance for measuring the distance between set of point. Two variants of this method, Bayesian KNN and Citation KNN were proposed in [2]. Deselaers and Ferrari [13] uses conditional random field where bag treated as nodes and instances treated as states of node. Babenko et al. [12] proposed bag as manifolds in the instance space. Recently, Jiang et al. [19] suggested improved version of lazy learning kNN algorithm as Bayesian Citation-kNN (BCkNN) algorithm.

The experimental results also show that different MI algorithms are appropriate for different MI problems and that no single MI algorithm was well-suited to every MI domain that was tested. In order to improve the classification accuracy and lessen the complexity of algorithm, proposed system suggests and compare the constructive covering algorithm with different kNN algorithm to create a set of covers to exclude the false positive instances.

3. MULTI- INSTANCE LEARNING

Multi-instance learning, as well-defined by Dietterich et al. [1], is a variation on the standard supervised machine learning scenario. In MI learning, every example consists of a multiset (bag) of instances. Each bag has a class label, but the instances themselves are not directly labeled. The learning problem is to form a model based on given example bags that can precisely predict the class labels of future bags. An example will help to illuminate the concept. Chevaleyre et al. [21] refer to this example as the simple jailer problem. Visualize that there is a locked door, and has N keychains, each containing a bunch of keys. If a keychain (i.e. bag) contains a key (i.e. instance) that can unlock the door, that keychain is considered to be useful. The learning problem is to build a model that can predict whether a given keychain is useful or not.

3.1 Definition of MIL

Let X be input space and Y= {0, 1} be the class label, binary output space. $F:2^X \rightarrow Y$ is a learning function for traditional supervised learning, form a set of training instances $\{(x_1,y_1)(x_2,y_2),..,(x_m,y_m)\}$Where x_i ϵ X is one instance and y_i ϵ Y is label associated with x_i.

In MIL, $\{(B_1,y_1)(B_2,y_2),..(B_m,y_m)\}$ is a training set of m labeled bags. Given a dataset D, instances in bag Bi defined as $\{x_1,x_2,..x_n\}$. Let d be dimension of x. Now, D^+ and D^- denotes all instances of positive and negative bags resp. where $D^+ =\{x_i^+ | i=1,2,..,p\}$, $D^- =\{x_j^+ | j=1,2,..,n\}$ and $D=D^+ U D^-$. In this each instance belongs to one specific bag. So, B_i U B_j=ф.

Block diagram of Multi-instance learning is explained below.

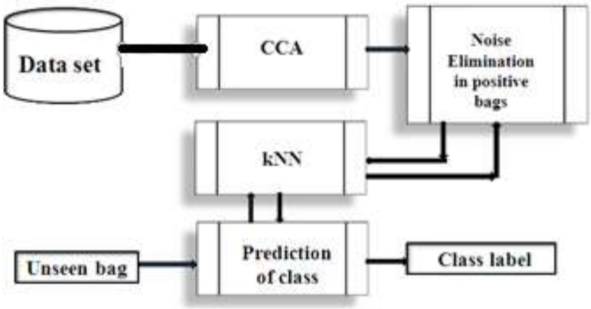


Figure. 2 Block diagram of MIL system

3.2 Constructive Covering Algorithm (CCA)

Supervised learning algorithm for McCulloch-Pitts neural model named as Constructive Cover Algorithm (CCA) proposed by Zhang and Zhang [6].This algorithm act as backbone of the system as it reorganize the structure of bags and uses the cover set as the new structure of bag. This algorithm is transformed by Zhao et al. [16] so that it can be useful in multi-instance concept. This algorithm finds and eliminate the false positive instances. The main idea of CCA is mapping all instances in the data set to a d-dimensional sphere S^d at first. Cover set is the final output and cover is nothing but a sphere with an instance as center and r as radius.

First data converted using T(x) = (x, $\sqrt{R^2 - ||x||^2}$, R≥max {||x|| | x ϵ D}) such that x is random instance and R is the greater or equal to maximum value of all instances. Transformation T: $D \rightarrow S^d$, where S^d is a d-dimensional sphere of the d+1 dimensional space, $\sqrt{R^2 - ||x||^2}$ is the additional value of x. After that, sequence of positive covers that only consists of instances from the positive bags and sequence of negative covers that only consists of instances of negative bags are constructed. To generate covers, first of all, an instance x_i ϵ D selected arbitrarily. Consider, X be the set of instances has the same label as x_i and -X the set of instances having opposite label from x_i. Then distance d_1 and d_2 computed such that

d_1= max {<x_i, x_j> | x_i ϵ X, x_j ϵ -X },

d_2= min {<x_i, x_k> | x_i, x_k є X}
Here x_j is the closest instance from x_i which belongs to set of X, whereas x_k is furthest instance from x_i which belongs to set of X. d2 must be smaller than d1 and where <x_1,x_2> signify the inner product between instances x_1 and x_2. Note that smaller the distance bigger the inner product. Next, radius r of sphere neighbor is calculated as r = (d_1+d_2)/2. The result of CCA is a series of covers, each of which contain samples belonging to the same class.

3.3 K Nearest Neighbor Algorithm (kNN)

kNN is widely used learning algorithm and well known for its relatively simple execution and decent results. The main idea of kNN algorithm is to find a set of k objects in the training data that are close to the test pattern, and base the assignment of a label on the predominance of a particular class in this neighbor. kNN is a lazy learning technique based on voting and distances of the k nearest neighbors. Given training set D and a test pattern x, kNN computes the similarity (distance) between x and the nearest k neighbors.The label of x is assigned by voting from the majority of neighbors. But rather than Euclidean distance, Hausdorff distance is used to measure similarity between two covers. Wang et al. [5] presented two types of Hausdorff distance that are - maximal Hausdorff distance (maxHD) and minimal Hausdorff distance (minHD).Given two sets of instances X={x_1,x_2,…x_n} and Z={z_1,z_2,…,z_n}, the maxHD defined as -
maxHD(X,Z)=max{h(X,Z),h(Z,X)}
where $h(X,Z)=\max_{x \epsilon X} \min_{z \epsilon Z} || x\text{-}z||$
The minHD is defined as -
minHD(X,Z)=min|| x-z ||
where || x-z || is the Euclidean distance between instance x and y.
Proposed work uses and compares accuracy and computation time of three kNN algorithm of this system. They are-
I. Citation kNN[5]:- C-kNN algorithm not only takes into account the k neighbors (references) of bag b but also the bags that count b as a neighbor (citers).Where number of citers c set to k+2 ,same as in [5] .
II. Bayesian kNN [5]:- Bayesian approach provides probabilistic approach that calculates explicit probability of hypotheses. For each hypothesis y that the class of b can take, the posterior probability of y is p(y| {y_1,y_2,..y_k }).According to the Bayes theorem, the maximally probable hypothesis is:
$\arg\max_{c} p(y| \{y_1,y_2,..y_k \})=\arg\max_{c} p(\{y_1,y_2,..y_k \}|y)p(c).$
where y_i is either positive or negative,

III Bayesian-Citation-kNN (BCKNN) [19]:- It is combined approach of Bayesian and distance weighting where firstly, Bayesian approach is applied to its k references and then distance weighting approach is applied to its c citers.

3.4 Noise Elimination in Positive Bags

For eliminating false positive instances kNN algorithm is utilized on cover obtained by CCA. For each $PCover_i$, its nearest neighbor calculated and checks if majority of its neighbors are belongs to set NCover, then it added in Ncover and deleted from NCover. The distance between two covers is calculated using Hausdroff distance (HD). MI data set transformed into positive cover set (PCover) and negative cover set (NCover). Fair amount of noises in positive bags are excluded.

3.5 Predication of Class label of test bags

In this method, a $PCover_i$, and $NCover_j$ is treated as the new structure of bag. Large numbers of noises in the positive bags are excluded during above procedures, it's now quite convenient to predict the labels of test bags using kNN algorithm at bag-level. It estimates the resemblance between each test bag and its nearest neighbors, if there are more negative covers around a test bag, then bag labeled as negative otherwise positive.

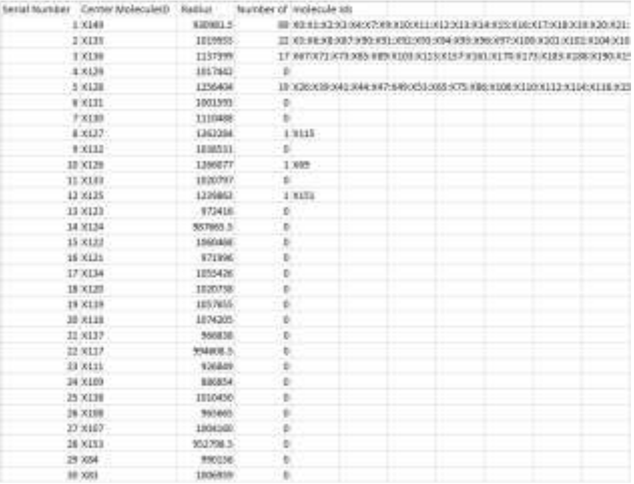
4. RESULTS AND DISCUSSION

The Multiple-instance classification learning algorithm eliminates the false positive instances at cover-level and labels the unknown bags at the bag-level. Two real-world benchmark data sets – Musk data sets (http://archive.ics.uci.edu/ml/datasets/Musk+%28Version+2%29) i.e. Musk1 and Musk 2 are used for experiments. Dataset contains different feature vectors of molecules and their class label. In this case, if molecule binds to target protein (putative receptor in human nose), then it smells like a musk. For determining whether molecule and target protein bind, shape of molecule is an important factor. However molecules are flexible and exhibit a wide range of shapes. Each molecule is represented by a bag and the bag's label is positive i.e. musky if molecule binds well to target protein. A bag made up of instances, where each instance represents one formation i.e. shape that molecule can take. After learning, it returns a concept which tells constraints on the shape of molecule that would bind to the target protein.

TABEL I
SUMMARY OF THE TWO MUSK DATA SETS

Data set	Total bags	Number of Positive bags	Number of Negative bags	Avg. instances per bag
Musk 1	92	47	45	5.17
Musk 2	102	39	63	64.69

Characteristics of datasets described in table 1. Each conformation represented by feature i.e. ray representation described in [1]. Musk 2 contains molecule that have more possible conformations i.e. instances than Musk1 is the main difference between two datasets



Serial Number	Center MoleculeID	Radius	Number of	molecule ids
1	X149	920901.5	[illegible]	[illegible]
2	X133	1019955	22	[illegible]
3	X136	1157399	17	[illegible]
4	X126	1017442	0	
5	X128	1256404	19	[illegible]
6	X131	1001593	0	
7	X130	1110488	0	
8	X127	1262284	1	X115
9	X132	1038531	0	
10	X129	1266077	1	X89
11	X103	1020797	0	
12	X125	1226862	1	X131
13	X123	972416	0	
14	X124	957865.5	0	
15	X122	1060468	0	
16	X121	971996	0	
17	X134	1055426	0	
18	X120	1020758	0	
19	X119	1057655	0	
20	X118	1074205	0	
21	X137	966838	0	
22	X117	994808.5	0	
23	X111	926849	0	
24	X109	886854	0	
25	X138	1010450	0	
26	X108	993465	0	
27	X107	1004340	0	
28	X133	952798.5	0	
29	X84	990156	0	
30	X83	1006859	0	

Figure 3 Snapshot of output file of CCA

Figure 3 shows output file of CCA which depicts some cover contains many instances while some none.

5. CONCLUSION

The multi-instance problem is the extension of supervised problem, arises in real world tasks where the samples are ambiguous, single example may has many alternative feature vectors that represent it and yet only one of those feature vectors may be responsible for the observed classification of object. CCA is used to break through and restructure the original bags into covers so that noises in the bags can be excluded by using various kNN algorithms. Then, covers as a whole, determines the labels of the unknown bags. So this is a cover-level multi-instance kNN learning algorithm, differ from previous bag or instance-level algorithm.

Detection of different fields, where this algorithm can be appropriate and suitable is one direction of future work. In addition, whether Multi-instance learning problem can be transformed into a supervised problem is additional direction of future work.

6. REFERENCES

[1] T. G. Dietterich, R. H. Lathrop, and T. Lozano-P erez, Solving the multiple Instance problem with axis-parallel rectangles, Artificial Intelligence, vol. 89,no. 1, p31-71, 1997.

[2] O. Maron and T. Lozano-P erez, A framework for multiple instance learning, Advances in Neural Information Processing Systems, vol. 10, pp. 570-576, 1998.

[3] O. Maron, Learning from ambiguity, Ph.D. dissertation, Massachusetts Institute Technology, USA, 1998.

[4] L. Zhang and B. Zhang, A geometrical representation of mcculloch-pitts neural model and its applications, IEEE Transactions on Neural Networks , vol. 10, no. 4, pp. 925- 929, 1999

[5] J. Wang and J. D. Zucker, Solving the multiple-instance problem: A lazy learning approach, in Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, pp. 1119-1125, 2000.

[6] Q. Zhang and S. A. Goldman, EM-DD: An improved multiple-instance learning technique, Advances in Neural Information Processing Systems, vol. 14, no. 2022, p1073-1080, 2001.

[7] S. Andrews, I. Tsochantaridis, and T. Hofmann, Support vector machines for multiple-instance learning, Advances in Neural Information Processing Systems, vol. 15, pp. 561-568, 2002.

[8] M. L. Zhang and Z. H. Zhou, Adapting RBF neural networks to multi-instance learning, Neural Processing Letters, vol. 23, no. 1, pp. 1-26, 2006.

[9] Z. Jorgensen, Y. Zhou, and M. Inge, A multiple instance learning strategy for combating good word attacks on spam filters, The Journal of Machine Learning Research, vol. 9, no. 6, pp. 1115-1146, 2008.

[10] B. B. Ni, Z. Song, and S. C. Yan, Web image mining towards universal age estimator, in Proceedings of the 17th ACM International Conference on MultimediaInt, ACM,pp. 85-94, 2009.

[11] T. Deselaers and V. Ferrari, A conditional random field for multiple-instance learning, in Proceedings of the 27th International Conference on Machine Learning, Morga Kaufmann Publishers Inc,pp. 1119-1125,2010.

[12] B. Babenko, N. Verma, P. Dollar, and S. J. Belongie, Multiple instance learning with manifold bags, in Proceedings of the 28th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, pp. 81-88, 2011.

[13] D. Zhang, Y. Liu, L. Si, J. Zhang, and R. D. Lawrence, Multiple instance learning on structured data, Advances in Neural Information Processing Systems, vol. 24, pp.145-153, 2011.

[14] Z. Q. Qi, Y. T. Xu, L. S. Wang, and Y. Song, Online multiple instance boosting for object detection, Neurocomputing, vol. 74, no. 10, pp. 1769-1775, 2011.

[15] Z. Y. Fu, A. Robles-Kelly, and J. Zhou, MILIS:Multiple instance learning with instance selection,IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 5, pp. 958-977, 2011.

[16] Y. Xie, Y. Y. Qu, C. H. Li, and W. S Zhang, Online multiple instance gradient feature selection for robust visual tracking, Pattern Recognition Letters, vol. 33, no. 9, pp. 1075-1082, 2012.

[17] M. Bellare, T. Ristenpart, and S. Tessaro, Multi-instance security and its application to password-based Cryptography, Advances in Cryptology-CRYPTO 2012, Springer, pp. 312-329, 2012.

[18] D. T. Nguyen, C. D. Nguyen, R. Hargraves, L. A. Kurgan,and K. J. Cios, mi-DS: Multiple-instance learning algorithm, IEEE Transactions on Systems, Man, and Cybernetics Society. Part B, Cybernetics, vol. 43, no. 1,pp. 143-154, Feb. 2013.

[19] L. X. Jiang, Z. H. Cai, D. H. Wang, and H. Zhang, Bayesian citation-KNN with distance weighting, International Journal of Machine Learning and Cybernetics, pp. 1-7,2013.

[20] Shu Zhao, Chen Rui, and Yanping Zhang, MICkNN : Multi-Instance Covering kNN Algorithm, TSINGHUA SCIENCE AND TECHNOLOGY,vo;. 18,no. 4, pp.360-368,2013.

[21] Chevaleyre, Y. & Zucker, J.-D. 2001. Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. Application to the mutagenesis problem. Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, Springer, pp.204-214.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 464 - 467, 2014, ISSN: 2319–8656

Classification with No Direct Discrimination

Deepali P. Jagtap
Department of Computer
Engineering,
KKWIEER, Nashik,
University of Pune, India

Abstract: In many automated applications, large amount of data is collected every day and it is used to learn classifier as well as to make automated decisions. If that training data is biased towards or against certain entity, race, nationality, gender then mining model may leads to discrimination. This paper elaborate direct discrimination prevention method. The DRP algorithm modifies the original data set to prevent direct discrimination. Direct discrimination takes place when decisions are made based on discriminatory attributes those are specified by the user. The performance of this system is evaluated using measures MC, GC, DDPP, DPDM etc. Different discrimination measures can be used to discover the discrimination.

Keywords: Classifier, Discrimination, Discrimination measures, DRP algorithm, Mining model.

1. INTRODUCTION

The Latin word 'Discriminare' is origin of the word Discrimination, its meaning is 'Distinguish between'. Discrimination is treating people unfairly based on their belonging to particular group. It restrict members of certain group from opporunities that are available to others [1]. Discrimination can also be observed in data mining. In data mining, large amount of data is collected and is used for training classifier. Classification model act as support to decision making process and the basis of scoring system. This makes business decision maker's work more easier [1,11]. If that historical data itself is biased towards or against certain entity such as gender, nationality, race, age group etc. Then resulting mining model may show discrimination. The use of automated decision making system gives sense of fair decision as it does not follow any personal preference but in actual results may be discriminatory [9]. Publishing such data leads to discriminatory mining results. The simple solution for discrimination prevention would consist neglecting or removing discriminating attributes, but even after removing those attribute, the discrimination may persist as many other nondiscriminatory attributes may strongly co-related with discriminatory attributes. The publicly available data may reveal co relation between them. Also removal of sensitive attribute results into loss of quality of original data [11].

Direct discrimination is observed when decisions are made based on the input data containing protected groups, whereas Indirect discrimination occurs when decisions are made based on nondiscriminatory input data but it is strongly or indirectly co-related with discriminatory one. For example, If discrimination occurs against foreign worker and even after removing that attribute from data set, one cannot guarantee that discrimination has been prevented completely as publicly available data may reveal nationality of that individual, hence shows indirect discrimination.

There are various laws against discrimination, but those are reactive not proactive. The use of technology and new mining algorithm helps to make them proactive. Along with mining algorithm, some algorithm and methods from privacy preservation such as data sanitization helps to prevent discrimination where original data is modified or support and confidence of certain attributes is changed to make them discrimination free. This system is useful in various applications such as credit/insurance scoring, lending, personnel selection and wage discrimination, job hiring, crime detection, activities concerning public accommodation, education, health care and many more. Benefit and services must be made available to everyone in a nondiscriminatory manner.

Rest of the paper is organized as follows. Section one provides introduction, survey of literature along with pros and cons of some of the existing methods are discussed in section two. Section three highlights basic terminology associated with this topic and section four describes algorithm as well as block diagram for discrimination prevention. Section five contains results and discussion about data set and finally last section presents conclusion along with the future scope of system.

2. LITERATURE SURVEY

Various studies have been reported the discrimination prevention in the field of data mining. Pedreschi noticed the discriminatory decisions in data mining based on classification rule and discriminatory measure. The work in this area can be traced back from year 2008. S. Ruggieri, Pedreschi and F. Turini [14] have implemented the DCUBE. It is oracle based tool to explore discrimination hidden in data. Discrimination prevention can be done in three ways based on when and in which phase data or algorithm is to be changed. Three ways for Discrimination prevention are: Preprocessing method, Inprocessing method and Postprocessing method. Discrimination can be of 3 types: Direct, Indirect or combination of both, based on presence of discriminatory attributes and other attributes that are strongly related with discriminatory one. Dino Pedreschi, Salvatore Ruggieri, Franco Turini[3] has introduced a model used in Decision Support System for the analysis and reasoning of discrimination that helps DSS owners and control authorities in the process of discrimination analysis [3].

Discrimination Prevention by Preprocessing Method

In preprocessing method, the original data set is modified in such a manner that it will not result in discriminatory classification rule. In this method any data mining algorithm can be applied to get mining model. Kamiran and Calder[4] proposed a method based on “data massaging” where class label of some of the records in the dataset is changed but as this method is intrusive, concept of "Preferential sampling" was introduced where distribution of objects in a given dataset is changed to make it non-discriminatory[4]. It is based on the idea that, "Data objects that are close to the decision boundary are more vulnerable to be victim of discrimination." This method uses Ranking function and there is no need to change the class labels. This method first divides data into 4 groups that are DP, DN, PP, PN, where first letters D and P indicate Deprived and Privileged class respectively and second letters P, N indicates positive and negative class label. The ranker function then sorts data in ascending order with respect to positive class label. Later it changes sample size in respective group to make that data biased free. Sara Hajian and Josep Domingo-Ferrer[9] proposed another preprocessing method to remove direct and indirect discrimination from original dataset. It employees 'elift' as discrimination measure to prevent discrimination in crime and intrusion detection system[10].

Preprocessing method is useful in applications where data mining is to be performed by third party and data needs to publish for public usage [9].

Discrimination Prevention by Inprocessing Method

Faisal Kamiran, Toon Calders and Mykola Pechenizkiy [5] introduced algorithm based on inprocessing method using decision Tree where instead of modifying original dataset data mining algorithm is modified. This approach consists of two techniques for the decision tree construction process, first is Dependency-Aware Tree Construction and another is Leaf Relabeling. The first technique focuses on splitting criterion for tree construction to build a discrimination-aware decision tree. In order to do so, it first calculates the information gain with respect to class & sensitive attribute represented by IGC and IGS respectively. There are three alternative criteria for determining the best split that uses different mathematical operation: (i) IGC-IGS (ii) IGC/IGS (iii) IGC+IGS. The second approach consists of processing of decision tree with discrimination-aware pruning and it relabel the tree leaves [5]. This methods requires special purpose data mining algorithms.

Discrimination Prevention by Postprocessing Method

Sara Hajian, Anna Monreale, Dino Pedreschi ,Josep Domingo Ferrer[12] proposed algorithm based on postprocessing method that derive frequent classification rule and modifies the final mining model using α-Protective k-Anonymous pattern sanitization to remove discrimination from Mined Model. Thus in postprocessing method, resultant mining model is modified instead of modifying original data or mining algorithm. The disadvantage of this method is, it doesn't allow original data to be published for public usage, and also the task of data mining should be performed by data holder only. Toon Calders and Sicco Verwer[15] proposed approach where the Naive Bayes classifier is modified to perform classification that is independent with respect to a given sensitive attribute. There are three approaches in order to make the Naive Bayes classifier discrimination-free: (i) modifying the probability of the decision being positive where the probability distribution of the sensitive attribute is modified. This method has disadvantage of either always increasing or always decreasing the number of positive labels assigned by the classifier, depending on how frequently the sensitive attribute is present in dataset, (ii) training one model for every sensitive attribute value and balancing them. This is done by splitting the dataset into two separate sets and the model is learned using only the tuples from the dataset that have a favored sensitive value, (iii) adding a latent variable to the Bayesian model. This method models the actual class labels using a latent variable[15].

3. THEORETICAL FOUNDATION

The basic terms in data mining are described in short as below:

3.1 Discrimination Measures

a. elift

Pedreschi [2] introduced ‘*elift*’ called extended lift as one of the discrimination measure. For a given classification rule, Extended lift can be calculated as below. Elift provides gain in confidence due to presence of discriminatory item [1].

$$elift\ (A, B \rightarrow C) = \frac{\text{Confidence}\ (A, B \rightarrow C)}{\text{Confidence}\ (B \rightarrow C)}$$

b. slift

The selection lift i.e. ‘*slift*’ for a classification rule of the form (A, B → C) is given as,

$$slift\ (A, B \rightarrow C) = \frac{\text{Confidence}\ (A, B \rightarrow C)}{\text{Confidence}\ (\neg A, B \rightarrow C)}$$

c. glift

Pedreschi [2] introduced ‘*glift*’ to strengthen the notion of α-protection. For a given classification rule glift is computed as,

$$glift(\beta,\gamma) = \begin{array}{ll} \beta/\gamma & \text{if } \beta \geq \gamma \\ (1 - \beta) / (1- \gamma) & \text{otherwise} \end{array}$$

3.2 Direct discrimination

Direct discrimination consists of rules or procedures that explicitly mention disadvantaged or minority groups based on sensitive discriminatory attributes [9]. For example, the rule r: (Foreign_worker = Yes, City = Nasik → Hire = No) shows direct discrimination as it contains discriminatory attribute Foreign_worker = yes.

3.3 Indirect discrimination

Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or un-intentionally could generate discriminatory decisions [9], for example, the rule r: Pin_code = 422006, City = Nashik → Hire = No shows indirect discrimination, as attribute Pin_code corresponds to area with mostly people belonging to particular religion.

3.4 PD Rule

A classification rule is said to be Potentially Discriminatory rule if it contains discriminatory item in premise of a given rule.

3.5 PND Rule

A classification rule is said to be Potentially Non-discriminatory rule if it doesn't contains any discriminatory item in premise of a given rule [1, 9].

4. DISSERTATION WORK

4.1 Algorithm and Process flow

The diagram showing overall process flow of discrimination prevention is shown in fig 1. The system takes original dataset containing discriminatory items as an input.

4.1.1 Input Data Preprocessing

The original dataset contains numerical values so it should be preprocessed i.e. discretization is performed on some attributes.

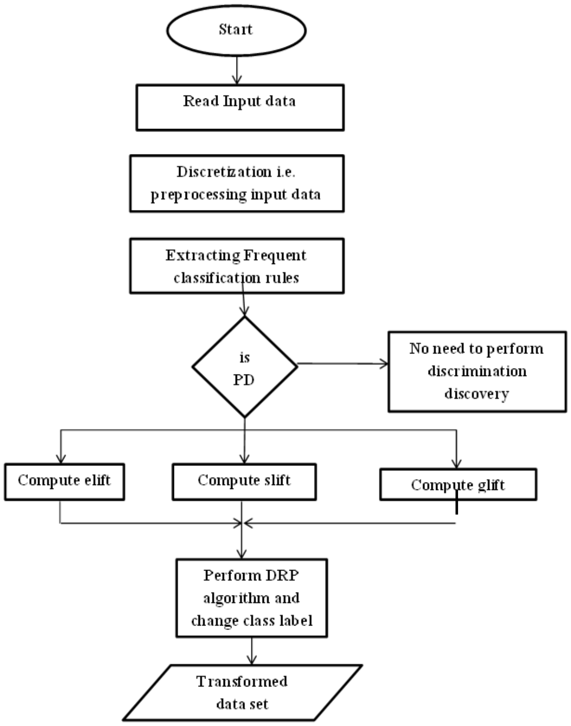

Start
Read Input data
Discretization i.e. preprocessing input data
Extracting Frequent classification rules
is PD
No need to perform discrimination discovery
Compute elift
Compute slift
Compute glift
Perform DRP algorithm and change class label
Transformed data set

Fig. 1. Flow chart for Discrimination Prevention

4.1.2 Frequent Classification Rule extraction

Later on using Apriori algorithm frequent item sets are generated. In Apriori algorithm candidate set generation and pruning steps are performed. The resultant frequent item sets are used to generate frequent classification rules.

4.1.3 Discrimination Discovery Process

The frequent classification rules are then categorized into Potentially Discriminatory and Potentially Nondiscriminatory groups in discrimination discovery. For discrimination discovery each classification rule is examined and is placed into either PD or PND group based on presence of discriminatory items in premise of the rule. If the rule found to be PD then for every PD rule elift, glift and slift is calculated. If that calculated value is greater than threshold value (α) then that rule is considered as α-discriminatory.

4.1.4 Data Transformation using DRP Algorithm

Data transformation is carried out in the next step as α-discriminatory rules need to be treated further to remove discrimination where class label of some of the records is perturbed to prevent discrimination. As a result of above process finally the transformed dataset is obtained as an output.

Data transformation is second step in discrimination prevention where the data is actually modified to make it biased free. In this step modifications are done using the definition of elift/glift/slift i.e. equality constraint on rule are enforced to satisfy the definition of corresponding discrimination prevention measure.

Direct Rule Protection (DRP) algorithm is used here that converts α-discriminatory rules into α-protective rule using the definition of elift. It can be done in following way:

let r': α-discriminatory rule, condition enforced on r' is:

$= \text{elift}(r') < \alpha$

$= \frac{\text{Confidence } (A,B \rightarrow C)}{\text{Confidence } (B \rightarrow C)}$

= confidence(r': A, B→C)/confidence(B→C) <α

= confidence(r': A, B→C/α <confidence (B→ C)

Here one needs to increase confidence (B→C), so change the class item from ¬C to C for all records in original DB that supports the rule of the form(¬A,B → ¬C). In this way this method changes the class label of class item in some records[9]. Similar method for slift as well as glift can be carried out.

4.2 Performance measures

To measure the success of the method in removing all evidence of Direct Discrimination and to measure quality of the modified data, following measures are used:

4.2.1 Direct discrimination prevention degree (DDPD)

The DDPD counts the percentage of α-discriminatory rules that are no longer α-discriminatory in the transformed data set.

4.2.2 Direct discrimination protection preservation (DDPP)

This measure counts the percentage of the α-protective rules in the original data set that remain α-protective in the transformed data set.

4.2.3 Misses cost (MC)

This measure helps to find the percentage of rules that are extractable from the original data set but cannot be extracted from the transformed data set. This is considered as side effect of the transformation process.

4.2.4 Ghost cost (GC)

This ghost cost quantifies the percentage of the rules that are extractable from the transformed data set but were not extractable from the original data set.

This MC and GC are the measures that are used in the context of privacy preservation. As similar approach of data sanitization is used in some methods for discrimination prevention, the same measures that are MC and GC can be applied to find out the information loss [16].

5. RESULTS AND DISCUSSION

German Credit Data set

This data set consists of 1000 records as well as 20 attributes. Out of those 20 attributes 7 are numerical and remaining 13 are categorical attributes. The class attributes indicates good or bad class for given bank account holder. Here the attribute foreign worker = Yes, Personal status = Female but not single and age = old are considered as discriminatory items where age > 50 is considered as old.

The Table I show the partial results computed on German credit dataset containing total number of classification rules generated and number of α-Discriminatory rules and the number of lines modified in original data set.

Table 1. German Credit dataset: Columns show the partial results for No. of α-Discriminatory rules, No. of lines modified.

Total No. of Classification rules	No. of α-Discriminatory rules	No. of Lines modified
9067	45	49

6. CONCLUSION

It is very important to remove discrimination, which can be observed in data mining, from original data. The removal of discriminatory attributes does not solve the problem. In order to prevent such discrimination, Discrimination Prevention by preprocessing technique is advantageous over the other two methods. The approach mentioned in this paper works in two steps: first is the discrimination discovery where α-Discriminatory rules are extracted and then in second step data transformation is performed in which the original data is transformed to prevent direct discrimination. This second step follows similar approach of Data Sanitization that is used in privacy preservation context. Many such algorithms uses 'elift' as a measure of discrimination, but instead of that one may use slift, glift as a measure of discrimination. The performance measure metrics i.e. DDPD, DDPP, MC, GC analyses data to check quality of transformed data as well as presence of direct discrimination. The less number of classification rules will be extracted from transformed data set as compare to original data set. The use of different discrimination measures such as slift, glift results into varying number of discriminatory rules and it have varying impact on original data.

In the future, one may explore how rule hiding in privacy preservation or other privacy preserving algorithms helps to prevent discrimination.

7. ACKNOWLEDGMENTS

With deep sense of gratitude I thank to my guide **Prof. Dr. S. S. Sane**, Head of Department of Computer Engineering, (KKWIEER, Nashik, University of Pune, India) for guiding me and his constant support and valuable suggestions that have helped me in the successful completion of this paper.

8. REFERENCES

[1] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining," Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD'08), pp. 560-568, 2008. (Cited by 56)

[2] D. Pedreschi, S. Ruggieri, and F. Turini, “Measuring Discrimination in Socially-Sensitive Decision Records,” Proc. Ninth SIAM Data Mining Conf. (SDM ’09),pp. 581-592, 2009.

[3] D. Pedreschi, S. Ruggieri, and F. Turini, “Integrating Induction and Deduction for Finding Evidence of Discrimination,”Proc. 12thACM Int’l Conf. Artificial Intelligence and Law (ICAIL ’09), pp. 157-166, 2009.

[4] F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling,"Proc. 19th Machine Learning Conf. Belgium and TheNetherlands, 2010.

[5] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning,"Proc. IEEE Int'l Conf. Data Mining (ICDM '10), pp. 869-874, 2010.

[6] M.Kantarcioglu, J. Jin and C. Clifton. When do data mining results violate privacy? In KDD 2004, pp. 599-604. ACM, 2004.

[7] P. N. Tan, M. Steinbach and V. Kumar, “Introduction to Data Mining”.Addison-Wesley, 2006.

[8] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases,"Proc. 20th Int'l Conf. Very Large Data Bases, pp. 487-499, 1994.

[9] S. Hajian and J. Domingo, "A Methodology for Direct and Indirect Discrimination prevention in data mining." IEEE transaction on knowledge and data engineeing, VOL. 25, NO. 7, pp. 1445-1459, JULY 2013. (Cited by 12)

[10] S. Hajian, J. Domingo-Ferrer, and A. Martnez Balleste, “Discrimination Prevention in Data Mining for Intrusion and Crime Detection,"Proc. IEEE Symp. Computational Intelligence in Cyber Security (CICS '11), pp. 47-54, 2011.

[11] S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, “Rule Protection for Indirect Discrimination Prevention in Data Mining,”Proc. Eighth Int’l Conf. Modeling Decisions for Artificial, Intelligence (MDAI ’11),pp. 211-222, 2011,

[12] S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer and F. Giannotti. “Injecting discrimination and privacy awareness into pattern discovery,” In 2012 IEEE 12th International Conference on Data Mining Workshops, pp. 360-369. IEEE Computer Society, 2012.

[13] S. Ruggieri, D. Pedreschi and F. Turini. “Data mining for discrimination discovery,” ACM Transactions on Knowledge Discovery from Data (TKDD), 4(2), Article 9, 2010.

[14] S. Ruggieri, D. Pedreschi, and F. Turini, "DCUBE: Discrimination Discovery in Databases,"Proc. ACM Int’l Conf. Management of Data (SIGMOD’10), pp. 1127-1130, 2010.

[15] T. Calders and S. Verwer. “Three naive Bayes approaches for discrimination-free classification,” Data Mining and Knowledge Discovery, 21(2):277-292, 2010. (Cited by 44)

[16] V. Verykios and A. Gkoulalas Divanis, "A Survey of Association Rule Hiding Methods for Privacy," Privacy-Preserving Data Mining: Models and Algorithms,C.C. Aggarwal and P.S. Yu, eds.,Springer, 2008.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 468 - 472, 2014, ISSN: 2319–8656

Joint Sentiment-Topic Detection from Text Document

Gauri Nivrutti Tuplondhe
Department of Computer Engineering,
KKWIEER, University of Pune,
Nashik- 422003, India

Abstract: Automated tools are used to detect subjective information like attitudes, opinions and feelings. Such process is called as sentiment analysis. The Joint Sentiment-Detection (JST) model is the probabilistic model which is extension of Latent Dirichlet Allocation (LDA) model that detects sentiment and topic simultaneously from text. Supervised approaches to sentiment classification often fail to produce satisfactory results when applied to other domains while the JST model is weakly supervised in nature where supervision only comes from domain independent sentiment lexicon. Thus, makes JST model portable to other domains. The proposed system incorporates a small amount of domain independent prior knowledge which is sentiment lexicon to further improve the sentiment classification accuracy. It also carry out experiments and evaluates the model performance on different datasets.

Keywords: Joint sentiment-topic (JST) model, Latent Dirichlet Allocation (LDA) , semi-supervised approach, sentiment analysis.

1. INTRODUCTION

Companies and consumers have the greater impact of opinion reach resources like online reviews and social networks compared to traditional media. The demand of gleaning insights into such vast amount of user-generated data, work on developing new algorithms for automated sentiment analysis has bloomed in the past few years.

Sentiment classification is the major task of sentiment analysis. A large portion of work concentrates on classifying a sentiment-bearing document according to its sentiment polarity, i.e. either positive or negative as a binary classification like [1], [2], [3], [9]. Most of this work rely on labeled corpora where documents are labeled as positive, negative prior to the training. In real world applications such labeled corpora may not be easily available. Also, sentiment classification models trained in one domain might not work well when moving to another domain. Furthermore, topic/feature detection and sentiment classification are mostly performed separately. But sentiments are context dependent, so that sentiment expressions can be quite different for different topics or domains. For instance, when appearing under different topics within movie review data, the adjective “complicated” may have negative sentiment orientation as “complicated role” in one topic, and positive orientation as “complicated plot” in another topic. This suggests that modeling sentiment and topic simultaneously may help find better feature representations for sentiment classification. Therefore, these problems motivated the need of using weakly supervised or unsupervised approaches for domain-independent sentiment classification.

Sentiment and topic of sentiment are simultaneously detected from text at document level by Joint Sentiment-Topic (JST) which is weakly supervised in nature. A mechanism is introduced to incorporate prior information about the sentiment lexicons into model learning by modifying the Dirichlet priors of the topic-word distributions. . This model extends the topic model latent dirichlet allocation (LDA) [6] by adding sentiment layer. It is different from other sentiment-topic model in that: 1) It is weakly supervised. 2) It can detect topics and sentiment simultaneously. Unlike supervised approaches to sentiment classification, which often fail to produce satisfactory performance when applied to other domains, the weakly-supervised nature of JST makes it highly portable to other domains, as will be verified by the experimental results on datasets from different domains.

2. RELATED WORK

2.1 Sentiment Classification

Standard machine learning techniques such as support vector machines (SVMs) and Naive Bayes (NB) classifiers are used for sentiment classification approaches. These approaches are corpus-based, in which a domain-specific classifier is trained with labeled training data. The work in [3] employed machine learning techniques including SVMs, NB and Maximum Entropy to determine whether the sentiment expressed in a movie review was “thumbs up’’ or “thumbs down”. In subsequent work [4], they further improved sentiment classification accuracy on the movie review dataset using a cascaded approach. The work [2], [3], [4] only focus on sentiment classification in one domain while the work in [5] addresses the issue of cross-domain sentiment classification. Four strategies have been explored for customizing sentiment classifiers to new domains [5] like small number of labeled examples can be used as training set or it can combine labeled data with large amount of unlabeled data from target domain. All the above work has some similar limitations: 1) the mixture of topics is ignored while doing sentiment classification, 2) They consider supervised learning approach by using labeled corpora for training which is not suitable for cross-domain work.

2.2 Sentiment-Topic Models

The work related to jointly determine sentiment and topic simultaneously from text is relatively sparse. Most closely related to our work is [7], [8], [9].]. Topic-sentiment model (TSM) [7] models the mixture of topics and sentiments simultaneously from web-blogs. TSM is based on the

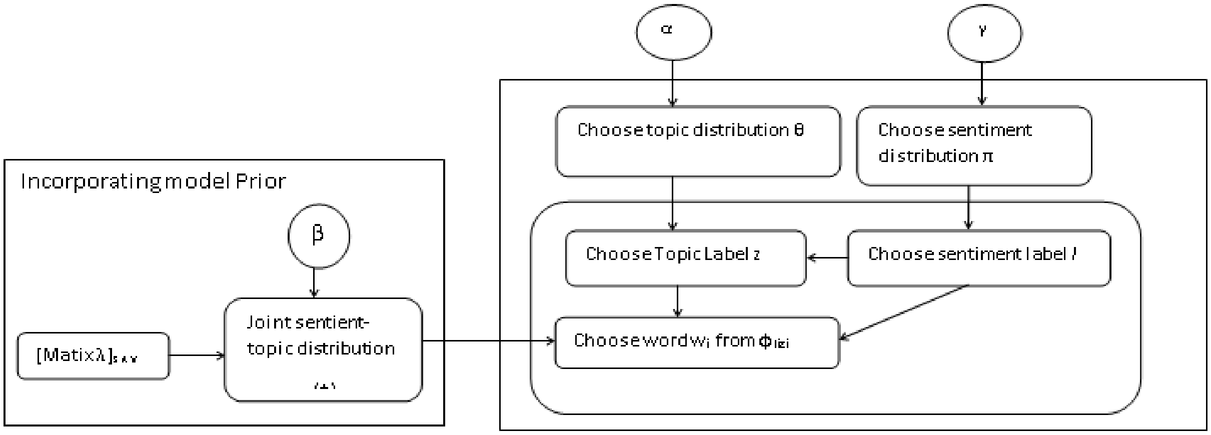

α
γ
Choose topic distribution θ
Choose sentiment distribution π
Incorporating model Prior
β
Choose Topic Label z
Choose sentiment label l
Joint sentient-topic distribution
[Matixλ]$_{S \times V}$
Choose word w_i from $\phi_{l_i z_i}$

Figure 1. Block Diagram

probabilistic latent semantic indexing (pLSI). It finds the latent topics in a Weblog collection, sentiments and the subtopics in the results of query. If the word is common English then it samples a word from background component model. Else, a word is sampled from a topical model or sentiment model. Thus, the word generation for sentiment is independent of topic. While in JST, a word is drawn from the joint distribution of sentiment and topic label. To obtain the sentiment coverage, TSM performs postprocessing. JST gives the document sentiment by using probability distribution of sentiment label for a given document.

The Multi-Grain Latent Dirichlet Allocation (MG-LDA) [8] is more appropriate to build topics in which a customer provide a rating for each aspect that is customer will annotate every sentence and phrase in a review as being relevant to some aspect. Each word is generated from either a global topic or a local topic. The model uses a topic model in that it assigns words to a set of induced topics, each of which may represent one particular aspect. The limitation of MG-LDA is that it does not considers the associations between sentiments and topics.

The MG-LDA model is extended to Multi-Aspect Sentiment [MAS] model [9]. The model extracts the ratable aspects of an object and cluster them into coherent topics. Then model uses various techniques to classify and aggregate sentiment over each of these aspects. Thus limitation of MG-LDA is overcome by MAS. It differs from JST in that it is a supervised model because it requires that every aspect should be rated which may not be possible in real world applications. While JST is a weakly supervised model which only requires minimum prior information.

3. METHODOLOGY

3.1 Joint Sentiment-Topic Model

JST model is the extension of existing LDA framework which has three hierarchical layers, where topics are associated with documents, and words are associated with topics. JST [10] introduces fourth layer to the LDA model called sentiment layer in order to consider sentiment of the document. Hence, JST becomes four-layer model, where sentiment labels are associated with documents, under which topics are associated with sentiment labels and words are associated with both sentiment labels and topics. The graphical model of JST is given in figure 1.

Consider a corpus with a collection of D documents denoted by C = $\{d_1, d_2, d_{3\ldots}, d_D\}$, each document in the corpus is a sequence of N_d words denoted by d = $(w_1, w2, \ldots. w_{nd})$, and each word in the document is an item from a vocabulary index with V distinct terms denoted by{1,2….V}. S be the number of distinct sentiment labels, and T be the total number of topics. The procedure for generating a word w_i in document d under JST can be given as: 1) Choose a sentiment label l from the per-document sentiment distribution π_d. 2) Choose a topic from the topic distribution $\theta_{d,l}$, where $\theta_{d,l}$ is conditioned on the sampled sentiment label l. Each document is associated with S topic distributions, each of which corresponds to a sentiment label l with the same number of topics. Thus, JST model can predict the sentiment associated with the extracted topics. 3) Draw a word from the per-corpus word distribution conditioned on both topic and sentiment label.

The graphical model of JST approach as shown in figure 1 can be defined as follows:

1) For every l (sentiment label) ϵ $\{1\ldots.,S\}$

- For every topic j ϵ $\{1\ldots.,T\}$, draw φ_{lj} ~ Dir $(\lambda_l \text{ X } \beta^{T}{}_{lj})$.

2) For every document d, choose a distribution π_d ~

Dir(γ).

3) For every l ϵ $\{1\ldots.S\}$ under document d, choose

a distribution $\theta_{d,l}$ ~ Dir (α).

4) For every word w_i in document d

- choose l_i ~ Mult (π_d),
- choose z_i ~ Mult $(\theta_{d,ll})$,

- choose a word w_i from φ_{lizi} which is a multinomial distribution over words conditioned on both sentiment label l_i and topic z_i.

The hyperparameters α and β in JST is the number of times topic *j* associated with sentiment label *l* is sampled from a document and the number of times words sampled from topic *j* are associated with sentiment label *l*, respectively. The hyperparameter γ is number of times sentiment label *l* sampled from a document before any word from the corpus is observed. π is per-document sentiment distribution, θ is per-document sentiment label specific topic distribution, and φ is per corpus joint sentiment-topic word distribution .

3.2 Incorporating Model Priors

JST model is the extension of LDA model in which additional dependency link of φ on the matrix λ of size S X V is used to encode word prior sentiment information into the JST model. A transformation matrix λ modifies the Dirichlet priors β of size S X T X V , so that the word prior sentiment polarity can be captured. The process of incorporating prior knowledge into the JST model is as follows: first, λ is initialized with all the elements equal to 1. For every sentiment label l ϵ {1…..S} and every word w ϵ {1….. V } in the corpus vocabulary, if word w is also available in the sentiment lexicons used, the element λ_{lw} is updated as follows:

$$\lambda lw = \begin{cases} 1, & if\ S(w) = 1 \\ 0, & otherwise. \end{cases}$$

where S(w) is the function which returns the prior sentiment label of w found in a sentiment lexicon (neutral, positive, or negative). Suppose, a word 'Bad' have polarity negative which is from vocabulary with index i. The corresponding row vector of λ is given by [1, 0, 0] which corresponds to negative, positive, neutral prior polarity. Now, for each topic j ϵ {1,…,T}, multiply λ_{li} with β_{lji}. Here, the value of β_{lnegji} is retained only and β_{lposji} and β_{lneuji} becomes 0.

3.3 Model Inference

To obtain the distributions of π, ɵ, and γ, first estimate the posterior distribution over *z* and *l*, i.e., the assignment of word tokens to topics and sentiment labels for a corpus. The sampling distribution for a word given remaining topics and sentiment labels is given by, $P(z_t{=}j,\ l_t{=}k|\ \alpha,\beta,\gamma)$. All words in the collection except for the word at location *'t'* in document D are given by z^{-t} and l^{-t} which are vectors of assignment of topics and sentiment labels.

The joint probability of the words, topics and sentiment label assignments can be given by

P(w, z, l)=P(w|z, l) P(z, l) = P(w|z, l)P(z|l) P(l) (1)

To estimate the posterior distribution by sampling the variables z_t and l_t , the process of Gibbs sampling is used. Let, the superscript -t denote a quantity that excludes word from t^{th} position. By marginalizing out random variables φ, θ and π , the conditional posterior for variables of interest z_t and l_t is given as

$$P(z(t) = j, l(t) = k|w, z(-t), l(-t), \alpha, \beta, \gamma) \ \alpha$$

$$\left(\frac{N(k,j,wt) - t + \beta}{N(k,j) - t + V\beta} \cdot \frac{N(d,k,j) - t + \alpha(k,j)}{N(d,k) - t + \sum j\, \alpha(k,j)} \cdot \frac{N(d,k) - t + \gamma}{N(d) - t + S\gamma}\right)$$

Samples obtained from the Gibbs sampling are used to approximate the per-corpus sentiment-topic word distribution which can be given as:

$$\varphi(k,j,i) = \frac{N(k,j,wt) + \beta}{N(k,j) + V\beta}$$

The approximate per-document topic distribution specific to the sentiment label can be given as:

$$\theta(d,k,j) = \frac{N(d,k,j) + \alpha(k,j)}{N(d,k) + \sum j\, \alpha(k,j)}$$

And the approximate per-document sentiment distribution can be given as

$$\pi(d,k) = \frac{N(d,k) + \gamma}{N(d) + S\gamma}$$

3.4 Algorithm

Algorithm : Procedure of Gibbs sampling for JST model.
Input: corpus, α, β, γ
Output : sentiment and topic label assignment for all word tokens in the corpus.

```
Initialize S X T X V matrix Φ , D X S X T matrix Θ,
  D X S matrix Π .
for i = 1 to maximum Gibbs sampling iterations do
  for all documents d = [1, D] do
    for all terms t = [1, Nd] do
      Exclude term t associated with topic label z
      and sentiment label l from variables Nd , Nd,k , Nd,k,j
      Nk,j and Nk,j,I;
      Sample a new sentiment-topic pair l̃ and z̃
                using above equation 2;
      Update variables Nd , Nd,k , Nd,k,j, Nk,j and Nk,j,i
      using the new sentiment label l̃ and topic label z̃ ;
    end for
  end for
  for every 25 iterations do
    Using Maximum Likelihood Estimation
    Update hyperparameter α;
  end for
¬¬¬ for every 100 iterations do
    Update matrices Θ, Φ, and Π with new
    Sampling results;
  end for
end for
```

3.5 Hyperparameter Setting and Classifying Document Sentiment

In the JST model implementation, set the symmetric prior β = 0:01, the symmetric prior γ = (0:05 X L) / S, where L is the average document length, S the is total number of sentiment labels. The asymmetric prior α is learned directly from data using maximum-likelihood estimation [11] and updated every 25 iterations during the Gibbs sampling procedure.

3.6 Classifying Document Sentiment

The document sentiment is classified as the probability of a sentiment label given a document P(l|d). Experiments only consider the probability of positive and negative labels for a given document, while the neutral label probability is ignored. A document d is classified as a positive if the probability of a positive sentiment label P(lpos|d) is greater than its probability of negative sentiment label P(lneg|d), and vice versa.

4. RESULTS AND DISCUSSION

4.1 Datasets

Two easily available data sets, movie review (MR) data set (http://www.cs.cornell.edu/people/pabo/movie-review-data) and Multi-domain sentiment (MDS) data set http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html are used in the experiments. The MR data set contains 1,000 positive and 1,000 negative movie reviews with average of 30 sentences each document. MDS data set is crawled from Amazon.com which includes reviews of four different products. Both data sets are first preprocessed in which punctuation, non-alphabet characters, numbers and stop words are removed. Two subjectivity lexicons, appraisal lexicon (http://lingcog.iit.edu/arc/appraisal_lexicon_2007b.tar.gz) and MPQA lexicon (http://www.cs.pitt.edu/mpqa/) are combined and incorporated as model prior information. Stemming is performed on both data sets and both lexicons in the preprocessing. The two lexicons used in work are fully domain independent and do not bear any supervised information related to the MR and MDS data set.

4.2 Performance Analysis

4.2.1 *Sentiment Classification Results versus Different Number of Topics*

As JST models sentiment and topic mixtures simultaneously, it is therefore worth exploring how the sentiment classification and topic extraction tasks affect/benefit each other and in addition, the model behave with different topic number settings on different data sets when prior information is incorporated.

Modeling sentiment and topics simultaneously help to improve sentiment classification. For the cases where a single topic performs the best , it is observed that the drop in sentiment classification accuracy by additionally modeling mixtures of topics is only marginal, but it is able to extract sentiment-oriented topics in addition to document-level sentiment detection.

4.2.3 *Topic Extraction*

Manually examining the data reveals that the terms that seem not convey sentiments under the topic in fact appear in the context of expressing positive sentiments.

5. CONCLUSION

JST model detects sentiment and topic simultaneously from a text at document level in a weakly supervised fashion. Only sentiment prior knowledge is incorporated which is independent of the domain. For general domain sentiment classification, by incorporating a small amount of domain independent prior knowledge, JST model achieves either better or comparable performance compared to existing semi-supervised approaches without using labeled documents. Thus, JST is flexible in the sentiment classification task. Weakly supervised nature of JST makes it highly portable to other domains. Moreover, the topics and topic sentiments detected by JST are indeed coherent and informative.

In future, incremental learning of the JST parameters can be done when facing with new data. Also, the modification of the JST model can be achieved by incorporating other supervised information into JST model learning, such as some known topic knowledge for certain product reviews or document labels derived automatically from the user supplied review ratings.

6. REFERENCES

[1] C. Lin, Yulan He, R. Everson “Weakly Supervised Joint Sentiment-Topic Detection from Text” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012.

[2] P.D. Turney, “Thumbs Up Or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews,” Proc. Assoc. for Computational Linguistics (ACL ’01), pp. 417-424, 2001.

[3] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs Up?: Sentiment Classification Using Machine Learning Techniques,” Proc. ACL Conf. Empirical Methods in Natural Language Processing (EMNLP) pp. 79-86, 2002.

[4] B. Pang and L. Lee, “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts,” Proc. 42th Ann. Meeting on Assoc. for Computational Linguistics (ACL), pp. 271-278, 2004.

[5] A. Aue and M. Gamon, “Customizing Sentiment Classifiers to New Domains: A Case Study,” Proc. Recent Advances in Natural Language Processing (RANLP), 2005.

[6] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation,” J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[7] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, “Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs,” Proc. 16th Int’l Conf. World Wide Web (WWW), pp. 171-180, 2007.

[8] I. Titov and R. McDonald, “Modeling Online Reviews with Multi-Grain Topic Models,” Proc. 17th Int’l Conf. World Wide Web, pp. 111-120, 2008.

[9] I. Titov and R. McDonald, “A Joint Model of Text and Aspect Ratings for Sentiment Summarization,” Proc. Assoc. Computational Linguistics—Human Language Technology (ACL-HLT), pp. 308-316 2008.

[10] C. Lin and Y. He, “Joint Sentiment/Topic Model for Sentiment Analysis,” Proc. 18th ACM Conf. Information and Knowledge Management (CIKM), pp. 375-384, 2009.

[11] T. Minka, “Estimating a Dirichlet Distribution,” technical report, MIT, 2003.

[12] S. Li and C. Zong, “Multi-Domain Sentiment Classification,” Proc. Assoc. Computational Linguistics—Human Language Technology (ACL-HLT), pp. 257-260, 2008.

[13] T. Hofmann, “Probabilistic Latent Semantic Indexing,” Proc. 22nd Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 50-57, 1999.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 473 - 478, 2014, ISSN: 2319–8656

Flood Prediction Model using Artificial Neural Network

Abhijit Paul
Assam University
Silchar, Assam, India

Prodipto Das
Assam University
Silchar, Assam, India

Abstract: This paper presents a Flood Prediction Model (FPM) to predict flood in rivers using Artificial Neural Network (ANN) approach. This model predicts river water level from rainfall and present river water level data. Though numbers of factors are responsible for changes in water level, only two of them are considered. Flood prediction problem is a non-linear problem and to solve this nonlinear problem, ANN approach is used. Multi Linear Perceptron (MLP) based ANN's Feed Forward (FF) and Back Propagation (BP) algorithm is used to predict flood. Statistical analysis shows that data fit well in the model. We present our simulation results for the predicted water level compared to the actual water level. Results show that our model successfully predicts the flood water level 24 hours ahead of time.

Keywords: FPM; ANN; MLP; FF; BP

1. INTRODUCTION

Flood occurs when river bursts its banks and the water spills on top of the floodplain. Flooding tends to be caused by heavy rainfall, when absorption of water is low and overflows are not controllable by river channels. The faster the rainwater reaches the river channel, the more likely it is to flood. Floods can cause damage to lives and property and possessions as well as disruption to communications. There is no mechanism to avoid flood but only a prediction can secure the life of inhabitants and also can reduce damages. To predict possible flood, most of the factors such as amount of rainfall, present river water level, degree of ground saturation, degree of permeable soil etc. need to be determined. If a forecast is issued after the prediction, then a flood warning can be communicated to warn the public about the possible extent of the flood, and to give people time to move out of the area. If forecasts can be made with long lag time between the storm and peak discharge, damages can be reduced in great scale.

The effective implementation of flood monitoring and forecasting system is non-trivial, since it requires the reliability coupled with the availability of related information. Over the years, flooding has been studied under various considerations and methodologies such as wireless sensors network, embedded system with a middleware, internet-based real-time data acquisition, and flood modeling and forecasting [1-3]. In addition to sensor technologies, space and satellite data technologies have been used to improve the accuracy [4] [5]. These papers provide great insights into the development of flood forecasting and modeling using data from satellite, image processing, and GIS. Different Mathematical and Statistical models are also used for flood forecasting [6]. Mathematical models are based on physical consideration and statistical models are based on analysis. To build flood monitoring and warning system, one of the widely used infrastructures is ad-hoc wireless sensor network [7]. In ad-hoc network, set of remote wireless are deployed in monitoring area to monitor water condition data, and these data are broadcasted in the form of web, sms or email technology to build a real time flood monitoring and forecasting system.

The wide variety of available forecasting techniques used by the hydrologists today, include physically based rainfall-runoff modeling techniques, data-driven techniques, and combination of the both, with forecasts ranging from short-term to long-term [8] [9]. Although hydrologists have used many models to predict flooding, the problem remains. Some of the models find difficulties with dynamic changes inside the watersheds. Some models are too difficult to implement and need to have robust optimization tools and some models require an understanding of the physical processes inside the basin. These problems have lead to exploration of a more data driven approach.

Therefore, to improve the accuracy of flood models and to deal with some of the above limitations, in recent years, several hydrological studies have used new techniques such as ANN, fuzzy logic and neuro-fuzzy to make flood predictions [10] [11]. These techniques are capable of dealing with uncertainties in the inputs and can extract information from incomplete or contradictory datasets. These new methods are frequently developed for hydrological and flood modeling only with rainfall and runoff as input and output, usually without taking into consideration of other flood causative factors.

Difficulty in river flood prediction is river water level fluctuation in highly nonlinear way. To solve this nonlinear problem, MLP based ANN approach is used. The input and output parameters used in this model are based on real-time data obtained from Flood Forecasting and Warning Centre, Bangladesh Water Development Board [12]. FPM is tested by the statistical fit functions- SST, SSE, MSE, RMSE, MAPE and R2 [13]. Then we present our simulation results of FPM for the predicted water level compared to the actual water level.

2. STUDY AREA

The river Manu originates from Sakhan range, Tripura, India and flows northerly via Kailashahar, Tripura, India to Bangladesh [14]. It joins the Kushiyara River at Manumukh in Maulvi Bazar district of Bangladesh. Basin area of Manu is 1979 sq. km. and its annual flow is 170034 in thousand m3 [15]. Its highest water level is 20.42 m and danger level is 18.0 m [12]. (Data collected online at http://www.ffwc.gov.bd/).

3. DATA SET

Daily water levels and rainfall data are collected online at Manu RB gauging station of river Manu from the Bangladesh Water Development Board (BWDB) website [12]. Approximate 200000 data at Manu RB gauging stations of river Manu under Meghna basin are utilized for training and testing.

4. FLOOD SIMULATION

4.1 ANN

ANNs are mathematical models of human perception that can be trained for performing a particular task based on available empirical data [16] [17]. When the relationships between input data and output data are unknown, they can make a powerful tool for modeling. The theory and mathematical basis of ANNs are explained in detail by many researchers. The model is based on a Feed Forward Multilayer Perceptron (FFMLP). As shown in Fig-1, an FFMLP includes a number of neurons or nodes that work in parallel to transform the input data into output categories. Typically, an FFMLP consists of three layers namely input, hidden layers and output. Each layer, depending on the specific application in a network, has some neurons. Each neuron is connected to other neurons in the next consecutive layer by direct links. These links have a weight that represents the strength of outgoing signal.

The input layer receives the data from different sources. Hence, the number of neurons in the input layer depends on the number of input data sources. The data are processed in hidden and output layers actively. The number of hidden layers and number of neurons in each layer are often defined by trial and error [18]. The number of neurons in output layers is fixed according to the application. Each hidden neuron responds to the weighted inputs it receives from the connected neurons from the preceding input layer. Once the combined effect on each hidden neuron is determined, the activation at this neuron is determined via a transfer function. Many differentiable nonlinear functions are available as a transfer function. Since the sigmoid function enables a network to map any nonlinear process, most networks of practical interest make use of it [19].

A typical Feed Forward ANN is shown in Fig-1.

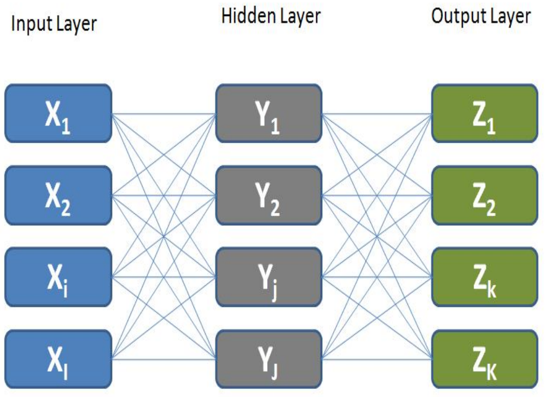


Figure. 1 Three Layer Feed Forward ANN

The first layer, called the input layer, consists of I nodes and connects with the input variables. This layer performs no computation but is used to distribute I inputs into the network. The last layer connects to the output variables and is called the output layer. This layer consists of K output nodes. The one or more layers of processing units located between the input and output layers have no direct connections to the outside world and are called hidden layers. An MLP may have a number of hidden layers. However, we have considered only two hidden layers. This layer consists of J hidden nodes.

In general, all connections are 'feed forward'; that is, they allow information transfer only from an earlier layer to the next consecutive layers. Nodes within a layer are not interconnected, and nodes in nonadjacent layers are not connected.

Each hidden node j receives I incoming signals (x_i), from every node i in the previous layer (for example from Input layer to Hidden layer). Associated with each incoming signal (x_i), there is a weight (w_{ij}) connected between layer I and J. The effective incoming signal (net_j) to node j is the weighted sum of all the incoming signals as follows:

$$net_j = \sum_{i=1}^{I} w_{ij} x_i \qquad (1)$$

This effective incoming signal (net_j) passes through a nonlinear activation function to produce the outgoing signal, called the activation or activity level, (y_j) of the node.

The delta learning rule or back propagation algorithm is used for learning the network. The purpose of this algorithm is to adjust the weights w_{ij} and w_{jk} which connects Input layers to Hidden layer and Hidden layers to Output layers to assure a minimization of the error function (E_k).

$$E_k = \frac{1}{2}(z_k - t_k)^2 \qquad (2)$$

where z_k, is the output from output layer and t_k is the target value.

The outgoing signal (z_k) is a function of the activation as follows:

$$z_k = f(net_k) \qquad (3)$$

The effective incoming signal net_k, comprising weighted sum signals from the hidden layer (y_j) is calculated as follows:

$$net_k = \sum_{j=1}^{J} w_{jk} y_j \qquad (4)$$

The outgoing signal (y_j) is the result of the incoming signal net_j, being passed through the activation function, as follows:

$$y_j = f(net_j) \qquad (5)$$

The weight adjustment of both w_{ij} and w_{jk} are based on the gradient descent search, which changes the weights in the direction in which the error surface goes down most steeply, as follows:

From input layer to hidden layer:

$$\Delta w_{ij} = -\eta \frac{\partial E_k}{\partial w_{ij}} \qquad (6)$$

where η is a learning parameter.

From hidden layer to output layer:

$$\Delta w_{jk} = -\eta \frac{\partial E_k}{\partial w_{jk}} \qquad (7)$$

Finally, the updated weights are

$$w_{ij} = w_{ij} + \Delta w_{jk} \qquad (8)$$

$$w_{jk} = w_{jk} + \Delta w_{jk} \qquad (9)$$

There are various activation functions employed in ANNs, the most commonly used ones being the following the unipolar binary function or sigmoid function (S). The sigmoid function is defined as

$$S(net_j) = \frac{1}{1 + e^{-\lambda net_j}} \qquad (10)$$

where $\lambda > 0$. λ is proportional to the neuron gain determining the steepness of the function.

The term λnet_j can vary on the range -∞ to +∞, but $S(net_j)$ is bounded between 0 and 1.

4.2 ANN Architecture

The ANN architecture refers to the number of layers and connection weights. It also defines the flow of information in the ANN. In ANN, design of suitable structure is the most important and also the most difficult part. There are no strict rules to define the number of hidden layers and neurons in the literature.

In this research, three-interconnection ANN architecture comprises an input layer, two hidden layers, and an output layer is used. The input layer contains two neurons (one for rainfall and another for water level) each representing a causative factor that contributes to the occurrence of the flood in the catchment. The output layer contains a single neuron representing river water level after 24 hours. The hidden layers and their number of neurons are used to define the complex relationship between the input and output variables.

4.3 Data Normalization

Neural network training can be made more efficient by performing certain pre-processing steps on the network inputs and targets [20]. Network input processing functions transforms inputs into better form for the network use. The normalization process for the raw inputs has great effect on preparing the data to be suitable for the training. Without this normalization, training the neural networks would have been very slow. There are several types of data normalization.

To normalize input and output dataset, the requirements are-

(i) Dataset minimum and maximum value.

(ii) Normalized scale minimum and maximum value.

$$X_{nor} = a + \frac{(X - A)(b - a)}{(B - A)} \qquad (11)$$

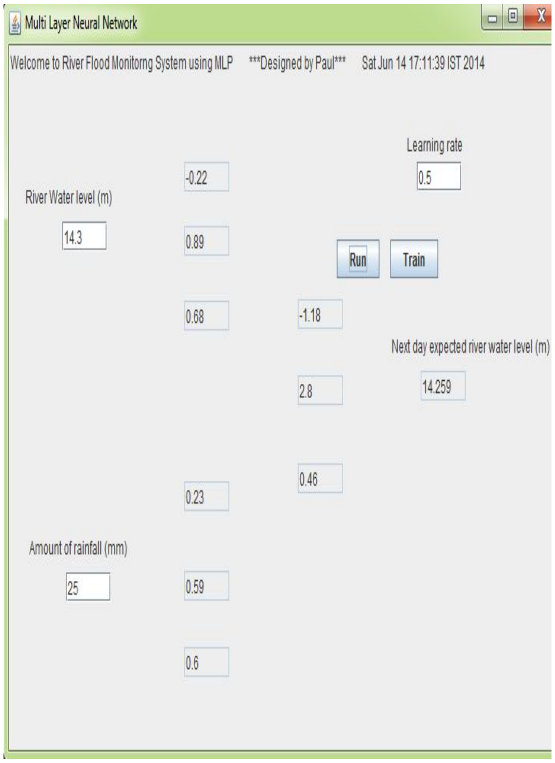
where X_{nor} is the normalized value of X, A is minimum value in the dataset, B is maximum value in the dataset, a is minimum value in the normalized scale and b is maximum value in the normalized scale.

4.4 Training and Testing the network

The aim of training process is to decrease the error between the ANN output and the real data by changing the weight values based on a BP algorithm.

A successful ANN model can predict target data from a given set of input data. Once the minimal error is achieved and training is completed, the FF algorithm is applied by ANN to generate a classification of the whole data set.

To train the ANN, a 2-N–N-1 format is used in this study, where N represents number of nodes in the hidden layer. By varying the number of neurons in both hidden layers, the neural networks are run several times to identify the most appropriate neural network architecture based on training and testing accuracies. Most appropriate ANN is selected based on minimum mean square error.



Welcome to River Flood Monitorng System using MLP ***Designed by Paul*** Sat Jun 14 17:11:39 IST 2014

Learning rate: 0.5

River Water level (m): 14.3

Amount of rainfall (mm): 25

-0.22

0.89

0.68

0.23

0.59

0.6

-1.18

2.8

0.46

Next day expected river water level (m): 14.259

Figure. 2 Snapshot of simulation result

The number of neurons in the second and third layers is checked from 1 to 10 for each of the layer. For each ANN configuration the training procedure is repeated starting from independent initial conditions and ultimately ensuring selection of the best performing network. The decreasing trend in the minimum mean square error in the training and validation sets is used to decide the optimal learning. The training is stopped when the minimum mean square error was achieved. Here 2-6-3-1 format gives us optimal result compare to other format. Therefore this 2-6-3-1 format is used for our experiment.

Here one input layer with two nodes- river water level and amount of rainfall is taken. Two hidden layers with six and three nodes are taken. One output layer with one node is taken. Initially all the weights are supplied random value. Using back propagation algorithm all the weights are updated so that it gives minimum error. After using back propagation algorithm, the updated weights are shown in fig-2. Initially the network is trained by past database. Once training is complete, the network can predict the value for its two inputs. For 14.3 m river water level and 25 mm rainfall, the network predict next day river water level as 14.259 which is shown in fig-2. Here learning rate 0.5 is used.

An important result in testing these data was that the ANN was able to identify all values same as training stage. This result yields a R2 value of 1 which is acceptable result and it shows a high level of prediction. The simulated and ANN predicted river flow, and the regression plot are shown in figs. 4 and 5, respectively.

After ANN training process is completed, different datasets are used to extend, and to determine the model accuracy. Using new data, the network performance was evaluated. These data had the same properties as the training data but they have not been used during the training of the model.

4.5 Algorithm

Step-1: Declare two inputs (river water level as i1 and rainfall as i2), nine weights (w[9]), one learning rate (l), one output (next day water level as o).

Step-2: Initially w[0] to w[8] by random values lies between 0-1.
Initialize l=0.5.

Step-3: Calculate dataset for i1 from 0-21. So A=0, B=21.
Calculate normalized scale for i2 from 0.1-1. So a=0.1, b=1.
Calculate $i1_{nor}$ according to the formula-2.
Also calculate dataset for i2 from 0-300. So A=0, B=300.
Calculate normalized scale for i2 from 0.1-1. So a=0.1, b=1.
Calculate $i2_{nor}$ according to the formula-2.

Step-4: Train the network with the two inputs i1, i2 and measured output o from the database.
Call the function train(i1,i2,o) with different values of i1, i2 and o. Training continues until it reaches to the threshold value 0.05.

Step-5: Now o can be calculated for any given value of i1 and i2.

Step-6: The calculated output o is in normalized form. This normalized value is converted into original value according to the formula-2, which is out predicted output.

5. MODEL PERFORMANCE ASSESMENTS

Variations between the predicted and observed values are shown using gnu plot graph. Data are predicted for 24 hours, 48 hours and 72 hours lead time.

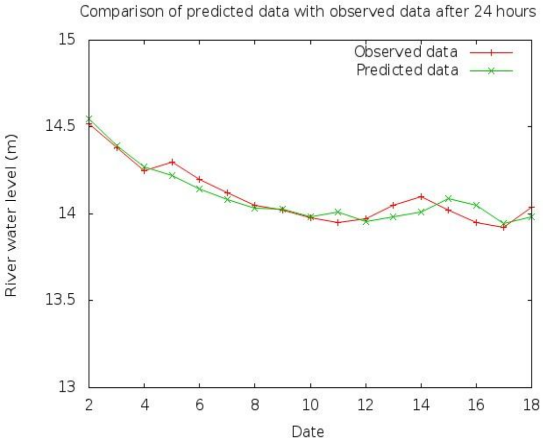


Figure. 3 Comparison of data for 24 hours lead time

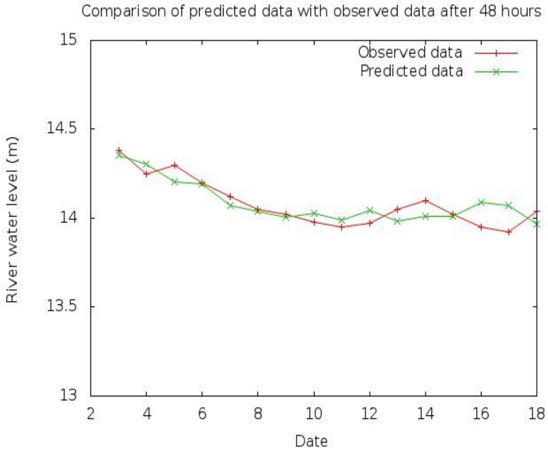


Figure. 4 Comparison of data for 48 hours lead time

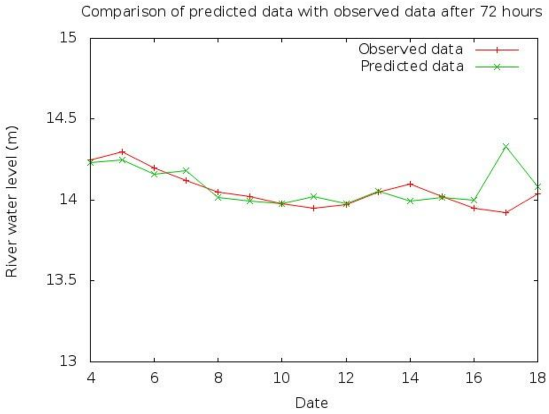


Figure. 5 Comparison of data for 72 hours lead time

The model accuracy assessment is described in terms of the error of forecasting or the variation between the observed and predicted values. In the literature, there are many performance assessment methods for measuring the accuracy and each one has advantages and limitations. Here for comparing the model, statistical measures are taken. Goodness-of-fit statistical functions measure how well data fit into the model. In this study, the most widely used methods namely coefficient of determination (R^2), total sum of squares (SST), sum square of error (SSE), mean sum of error (MSE), root mean square error (RMSE), mean absolute percentage of error (MAPE) are used to check the performance of the model. Each method is estimated from the ANN predicted values and the observed values.

Statistical formulae are given below-

Total sum of square,

$$\mathrm{SST} = \sum_{i-1}^{n} (y_i - \bar{y})^2 \qquad (12)$$

Sum of square of error,

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (13)$$

Mean sum of error,

$$\mathrm{MSE} = \frac{1}{n}\ x\ \mathrm{SSE} \qquad (14)$$

Root mean square of error,

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \qquad (15)$$

Mean absolute percentage of error,

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n} |(y_i - \hat{y}_i)/y_i| \times 100 \qquad (16)$$

Coefficient of determination R square,

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \qquad (17)$$

TABLE 1.
STATISTICAL MEASURES

	For 24 hours lead time	For 48 hours lead time	For 72 hours lead time
SST	0.45695294	0.275775	0.180573
SSE	0.046602	0.083753	0.199997
MSE	0.002741	0.005235	0.013333
RMSE	0.052357	0.07235	0.115469
MAPE	0.935756	0.763017	0.62728
R^2	0.898016	0.6963	-0.10757

The results showed that the model has less SSE, MSE, and RMSE. Overall, the errors are negligible. The higher value (close to 1) of MAPE and R^2 seems the model has excellent agreement with the real data. The results show that data for 24 hours lead time has better result compare to others. Therefore our model is used to predict only water level after 24 hours which means next day river water level prediction.

6. CONCLUSION

The focus of this paper is to apply optimized ANN for next day river water level forecasting by determination of suitable input parameters and designing the best network architecture. The study reported in this article has led to the conclusion that MLP type network, consistently performed better compared to other network. Among the water level prediction after 24 hours, 48 hours and 72 hours; prediction after 24 hours performs well. Therefore our ANN model with MLP is used only for predicting next day (24 hours) water level.

7. REFERENCES

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “A survey on sensor networks,” IEEE Communications Magazine, vol. 40 (8), pp. 102–114, 2002.

[2] V. Seal, A. Raha, S. Maity, S. K. Mitra, A. Mukherjee, and M. K. Naskar, “A Simple Flood Forecasting Scheme using Wireless Sensor Networks,” *International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC)*, vol.3, no.1, pp. 45-60, 2012.

[3] V. Sehgal and C. Chatterjee, “Auto Updating Wavelet Based MLR Models for Monsoonal River Discharge Forecasting,” *International Journal of Civil Engineering Research*, vol. 5, no. 4, pp. 401-406, 2014.

[4] M. B. Kia, S. Pirasteh, B. Pradhan, A. R. Mahmud, W. N. A. Sulaiman, and A. Moradi, “An artificial neural network model for flood simulation using GIS:Johor River Basin, Malaysia,” *Environ Earth Sci*. Available: doi:10.1007/s12665-011-1504-z.

[5] S. Fang, L. Xu, H. Pei, Y. Liu, Z. Liu, Y. Zhu, J. Yan, and H. Zhang, "An Integrated Approach to Snowmelt Flood Forecasting in Water Resource Management," *Industrial Informatics, IEEE Transactions on*, vol. 10, no. 1, pp. 548-558, 2014.

[6] Y. Chakhchoukh, P. Panciatici, and L. Mili, "Electric Load Forecasting Based on Statistical Robust Methods," *Power Systems, IEEE Transactions on*, vol. 26, no. 3, pp. 982-991, 2011.

[7] Z.J. Haas, T. Small, "A new networking model for biological applications of ad hoc sensor networks," *Networking, IEEE/ACM Transactions on*, vol. 14, no. 1, pp. 27-40, 2006.

[8] F. Hossain, E.N. Anagnostou, and T. Dinku, "Sensitivity analyses of satellite rainfall retrieval and sampling error on flood prediction uncertainty," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 1, pp. 130-139, 2004.

[9] Y. Zhang, Y. Hong, X. Wang, J.J. Gourley, J. Gao, H.J. Vergara, and B. Yong, "Assimilation of Passive Microwave Streamflow Signals for Improving Flood Forecasting: A First Study in Cubango River Basin, Africa," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of* , vol. 6, no. 6, pp. 2375-2390, 2013.

[10] F. J. Chang, J. M. Liang, and Y. C. Chen, "Flood forecasting using radial basis function neural networks," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* , vol. 31, no. 4, pp. 530-535, Nov 2001.

[11] E. Todini, "Using a Desk-Top Computer for an On-Line Flood Warning System," *IBM Journal of Research and Development*, vol.22, no.5, pp.464-471, 1978.

[12] (2014) Flood Forecasting & Warning Centre, Bangladesh Water Development Board (BWDB) website. [Online]. Available http://www.ffwc.gov.bd/

[13] S. Shakya, H. Yuan, X. Chen, and L. Song, "Application of radial basis Function Neural Network for fishery forecasting," *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on* , vol. 3, no., pp. 287-291, 10-12 June 2011.

[14] M Deb, D. Das, and M. Uddin, "Evaluation of Meandering Characteristics Using RS & GIS of Manu River," *Journal of Water Resource and Protection*, vol. 4, pp. 163-171, 2012.

[15] (2002) State of Environment Report, Tripura State Pollution Control Board. [Online]. Available http://envfor.nic.in/sites/default/files/State%20of%20Environment%20Report%20-%20Tripura%202002.pdf

[16] A.R. Gainguly, "A hybrid approach to improving rainfall forecasts," *Computing in Science & Engineering*, vol. 4, no. 4, pp. 14-21, 2002.

[17] L. C. Chang, P. A. Chen, F. J. Chang, "Reinforced Two-Step-Ahead Weight Adjustment Technique for Online Training of Recurrent Neural Networks," Neural Networks and Learning Systems, IEEE Transactions on , vol. 23, no. 8, pp. 1269-1278, 2012.

[18] Atkinson PM, Tatnall ARL, "Neural networks in remote sensing," *International Journal of Remote Sensing*, vol. 18, pp. 699–709, 1997.

[19] Bishop CM, "Neural networks and their application," *Review of Scientific Instruments*, vol. 65, no. 6, pp. 1803–1830, 1994.

[20] (2014) R. K. Biswas and A. W. Jayawardena, "Water Level Prediction By Artificial Neural Network In The Surma-Kushiyara River System of Bangladesh". [Online]. Available http://www.icharm.pwri.go.jp/training/master/pubilication/pdf/2009/1.synopsis_mee08177_robin.pdf

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 479 - 483, 2014, ISSN: 2319–8656

A Survey of Image Steganography

Sandeep Kaur
Guru Nanak de Engg. Collage
Ludhiana, India

Arounjot Kaur
Guru Nanak de Engg. Collage
Ludhiana, India

Kulwinder Singh
Guru Nanak de Engg. Collage
Ludhiana, India

Abstract: This paper presents a general overview of the steganography. Steganography is the art of hiding the very presence of communication by embedding secret messages into innocuous looking cover documents, such as digital images. Detection of steganography, estimation of message length, and its extraction belong to the field of steganalysis. Steganalysis has recently received a great deal of attention both from law enforcement and the media. In this paper review the what data types are supported, what methods and information security professionals indetecting the use of steganography, after detection has occurred, can the embedded message be reliably extracted, can the embedded data be separated from the carrier revealing the original file, and finally, what are some methods to defeat the use of steganography even if it cannot be reliably detected.

Keywords: steganography, Image steganography, cryptography, stego image and stego key.

1. INTRODUCTION

Steganography comes from the Greek words Steganós (Covered) and Graptos (Writing) [1]. In the past, people used hidden tattoos or invisible ink to convey steganographic content. Today, computer and network technologies provide easy-to-use communication channels for steganography[2].

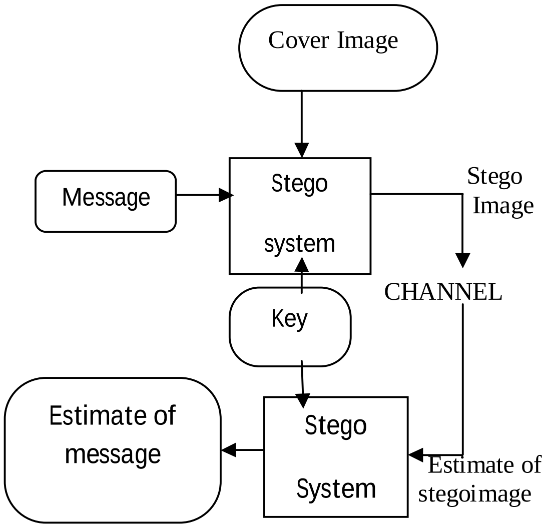


Figure 1. Overview of steganographic system

Steganography is a technique of information security that hides secret information within a normal carrier media, such as digital image, audio, video, etc. An unauthorized attempt to detect and extract the hidden secret information from stego is known as steganalysis [3]. The embedding process creates a stego medium by replacing these redundant bits with data from the hidden message. Modern steganography goal is to keep its mere presence undetectable. classical steganographic system's security relies on the encoding system's secrecy. An example of this type of system is a Roman general who shaved a slave's head and tattooed a message on it. After the hair grew back, the slave was sent to deliver the now-hidden message.Although such a system might work for a time, once it is known, it is simple enough to shave the heads of all the people passing by to check for hidden messages—ultimately, such a steganographic system fails. Modern steganography attempts to be detectable only if secret information is known—namely, a secret keys [4]. A block diagram of a generic blind image steganographic system is depicted in Fig. 1. A message is embedded in a digital image by the stegosystem encoder, which uses a key or password. The resulting stegoimage is transmitted over a channel to the receiver, where it is processed by the stegosystem decoder using the same key[5].

2. CATEGORIES OF IMAGE STEGANOGRAPHY

Almost all digital file formats can be used for steganography, but the formats that are more suitable are those with a high degree of redundancy.

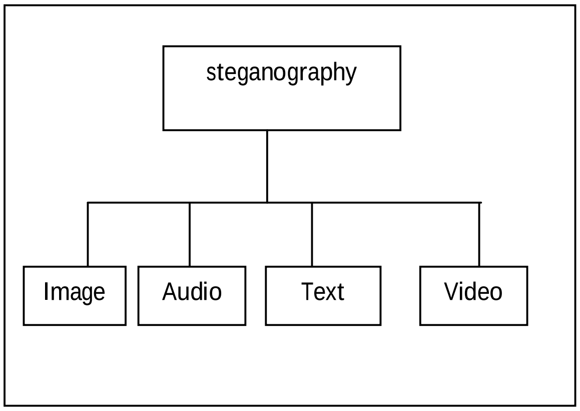


Figure 2. Categories of steganography

Redundancy can be defined as the bits of an object that provide accuracy far greater than necessary for the object's use and display [6].

2.1 Image steganography

To hide information, straight message insertion may encode every bit of information in the image or selectively embed the message in ―noisy‖ areas that draw less attention—those areas where there is a great deal of natural color variation [7]. The message may also be scattered randomly throughout the image. A number of ways exist to hide information in digital media. Common approaches include

2.1.1 Least significant bit insertion
2.1.2 Masking and filtering
2.1.3 Redundant Pattern Encodings
2.1.4 Encrypt and Scatter
2.1.5 Algorithms and transformations

2.2 Audio steganography

Audio Steganography is a method of hiding the message in the audio file of any formats.EAS provides an easy way of implementation of mechanisms .When compared with audio steganography. Apart from the encoding and decoding in

Audio steganography[10]. EAS contain extra layers of encryption and decryption. The four layers in EAS are:

2.2.1 Encoding
2.2.2 Decoding
2.2.3 Encryption
2.2.4 Decryption

2.3 Text steganography

Since everyone can read, encoding text in neutral sentences is doubtfully effective. But taking the first letter of each word of the previous sentence, you will see that it is possible and not very difficult. Hiding information in plain text can be done in many different ways[8][9] Many techniques involve the modification of the layout of a text, rules like using every n-th character or the altering of the amount of white space after lines or between words.

2.4 Video steganogrphy

Video files are generally a collection of images and sounds, so most of the presented techniques on images and audio can be applied to video files too .The great advantages of video are the large amount of data that can be hidden inside and the fact that it is a moving stream of images and sounds.Therefore, any small but otherwise noticeable distortions might go by unobserved by humans because of the continuous flow of information.

3. STEGANOGRAPHIC TECHNIQUES

There are quite a lot of approaches in classifying steganographic techniques. These approaches can be classified in accordance with the type of covers used with secret communications. Another possibility is done via sorting such approaches depending on the type of cover modification already applied in the process of embedding. Steganographic techniques that modify image files for hiding information include the following[11]:

- Spatial domain;
- Transform domain;
- Spread spectrum;
- Statistical methods; and
- Distortion techniques

3.1 Steganography in the spatial domain

In spatial domain methods a Steganographer modifies the secret data and the cover medium in the spatial domain, which is the encoding at the level of the LSBs. This method has the largest impact compared to the simplicity [12]. Spatial steganography mainly includes LSB (Least Significant Bit) steganography Least significant bit (LSB) insertion is a common, simple approach to embedding information in a cover image . The least significant bit (in other words, the 8th bit) of some or all of the bytes inside an image is changed to a bit of the secret message.

Pixel: (10101111 11101001 10101000)

(10100111 01011000 11101001)

(11011000 10000111 01011001)

Secret message: 01000001

Result: (10101110 11101001 10101000)

(10100110 01011000 11101000)

(11011000 10000111 01011001)

3.2 Steganography in the frequency domain

New algorithms keep emerging prompted by the performance of their ancestors (Spatial domain methods), by the rapid development of information technology and by the need for an enhanced security system. The discovery of the LSB embedding mechanism is actually a big achievement. DCT is used extensively in Video and image (i.e., JPEG) lossy compression. Most of the techniques here use a JPEG image as a vehicle to embed their data.

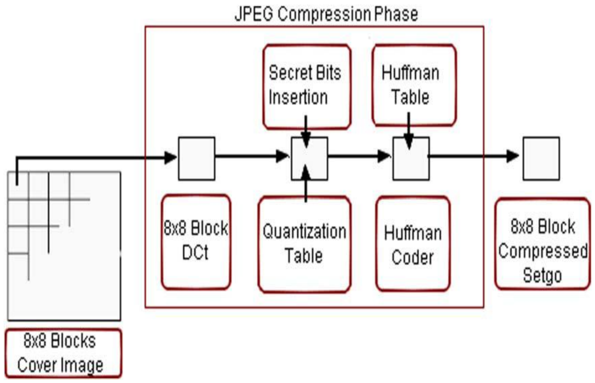


Figure 3. Data Flow Diagram showing a general process of embedding in the frequency domain.

JPEG compression uses DCT to transform successive sub-image blocks (8x8 pixels) into64 DCT coefficients.

3.3 Transform domain technique

We have seen that LSB modification techniques are easy ways to embed information but they are highly vulnerable to even small cover modifications. It has been noted early in the development of steganographic systems that embedding information in the frequency domain of a signal can be much more robust than embedding rules operating in the time domain. Transform domain methods hide messages in significant areas of the cover image which makes them more robust to attacks, such as compression, cropping, and some image processing, than the LSB approach. However, while they are more robust to various kinds of signal processing, they remain imperceptible to the human sensory system. Many transform domain variations exist. One method is to use the discrete cosine transformation (DCT) [13][14]. This method is used, but similar transforms are for example the Discrete Fourier Transform (DFT). These mathematical transforms convert the pixels in such a way as to give the effect of “spreading” the location of the pixel values over part of the image. The DCT transforms a signal from an image representation into a frequency representation, by grouping the pixels into 8 × 8 pixel blocks and transforming the pixel blocks into 64 DCT. DCT is used in steganography as- Image is broken into 8×8 blocks of pixels. Working from left to right, top to bottom, the DCT is applied to each block[15].

3.4 Spread spectrum

Spread spectrum communication describes the process of spreading the bandwidth of a narrowband signal across a wide band of frequencies. This can be accomplished by modulating the narrowband waveform with a wideband waveform, such as white noise. After spreading, the energy of the narrowband signal in any one frequency band is low and therefore difficult to detect[16].

3.5 Statistical methods

Statistical Methods also known as model-based techniques, these techniques tend to modulate or modify the statistical properties of an image in addition to preserving them in the embedding process. This modification is typically small, and it is thereby able to take advantage of the human weakness in detecting luminance variation [17]. Statistical steganographic techniques exploit the existence of a “1-bit”, where nearly a bit of data is embedded in a digital carrier. This process is done by simply modifying the cover image to make a sort of significant change in the statistical characteristics if a “1” is transmitted, otherwise it is left unchanged. To send multiple bits, an image is broken into sub-images, each corresponding to a single bit of the message [18].

3.6 Distortion techniques

Distortion techniques store information by signal distortion and measure the deviation from the original cover in the decoding step[19]. In contrast to substitution systems, distortion techniques require the knowledge of the original cover in the decoding process. Alice applies a sequence of modifications to a cover in order to get a stego-object; she chooses this sequence of modifications in such a way that it corresponds to a specific secret message she wants to transmit. Bob measures the differences to the original cover in order to reconstruct the sequence of modifications applied by Alice, which corresponds to the secret message[20].

4. DIFFERENCE BETWEEN CRYPTOGRAPHY AND STEGANOGRAPHY

Basically, the purpose of cryptography and steganography is to provide secret communication. However, steganography is not the same as cryptography. Cryptography scrambles a message by using certain cryptographic algorithms for converting the secret data into unintelligible form. On the other hand, Steganography hides the message so that it cannot be seen. Cryptography offers the ability of transmitting information between persons in a way that prevents a third party from reading it. Cryptography can also provide authentication for verifying the identity of someone or something. In contrast, steganography does not alter the structure of the secret message, but hides it inside a cover-image so it cannot be seen. Staganography and cryptography differences are briefly summarized following in Table I.

TABLE I. DIFFERENCE BETWEEN CRYPTOGRAPHY AND STEGANOGRAPHY

CRYPTOGRAPHY	STEGANOGRAPHY
Known message passing	Unknown message passing
Common technology	Little known technology
Technology still being developed for certain Formats	Most of algorithm known by all
Cryptography alter the structure of the secret message	Steganography does not alter the structure of the secret message

5. METHODS OF STEGANALYSIS

Steganalysis is "the process of detecting steganography by looking at variances between bit patterns and unusually large file sizes”. It is the art of discovering and rendering useless covert messages [21]. The goal of steganalysis is to identify suspected information streams, determine whether or not they have hidden messages encoded into them, and, if possible, recover the hidden information, unlike cryptanalysis, where it is evident that intercepted encrypted data contains a message. The process of steganalysis is depicted in Fig. 4.

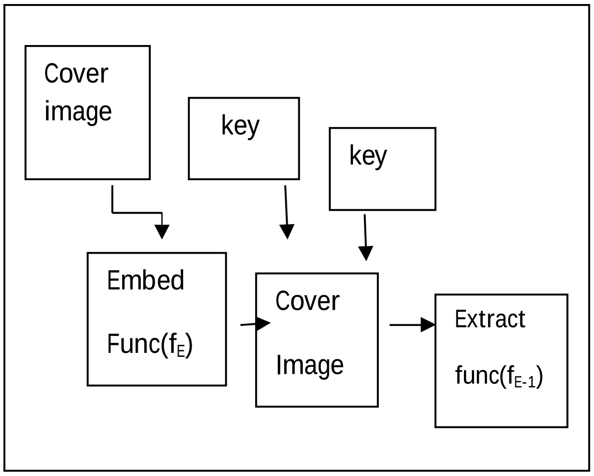


Figure 4. Process of steganalysis

It is the art of discovering and rendering useless covert messages [21]. The goal of steganalysis is to identify suspected information streams, determine whether or not they have hidden messages encoded into them, and, if possible, recover the hidden information, unlike cryptanalysis, where it is evident that intercepted encrypted data contains a message. The process of steganalysis is depicted in Fig. 4.

6. VISUAL DETECTION

Most steganographic programs embed message bits either sequentially or in some pseudo-random fashion. In most programs, the message bits are chosen non-adaptively independently of the image content. If the image contains connected areas of uniform color or areas with the color saturated at either 0 or 255, we can look for suspicious artifacts using simple visual inspection after preprocessing the stego-image. Even though the artifacts cannot be readily seen,we can plot one bit-plane (for example, the LSB plane) and inspect just the bit-plane itself [22].This attack is especially applicable to palette images for LSB embedding in indices to the palette.

7. STATISTICAL DETECTION

Statistical attack that can be applied to any steganographic technique in which a fixed set of Pairs of Values (PoVs) are flipped into each other to embed message bits[23]. These methods use first or higher order statistics of the image to reveal tiny alterations in the statistical behavior caused by steganographic embedding and hence can successfully detect even small amounts of embedding with very high accuracy.

8. CONCLUSION

The meaning of Steganography is hiding information and the related technologies.The purpose of this paper is to present a survey of various approaches for image steaganography based on their various types and techniques.

9. ACKNOWLEDGMENTS

I thanks to a great many people who helped and supported me during writing of this paper. My deepest thanks to Arunjot kaur Brar Assistant Professor of department of Information Technology Guru Nanak dev Engineering college ludhiana Punjab. Who guided and supported me in every phase of writing this paper. I am grateful to my parents who are inspirational in their understanding patience and constant encouragement.

10. REFERENCES

[1] Dr. Ekta Walia , Payal Jain , Navdeep ‘An Analysis of LSB & DCT based Steganography’ Vol. 10 Issue 1 (Ver 1.0), April 2010.

[2] Niels Provos and Peter Honeyman, University of Michigan, ‘Hide and

[3] Seek: An Introduction to Steganography’ published by the ieee computer society, 2003 IEEE.

[4] Hardik Patel*, Preeti Dave, ‘Steganography Technique Based on DCT Coefficients’ International Journal of Engineering Research and Applications Vol. 2, Issue 1,Jan-Feb 2012, pp.713-717

[5] Niels Provos and Peter Honeyman, University of Michigan, ‘Hide and Seek: An Introduction to Steganography’ published by the ieee computer society, 2003 IEEE.

[6] Lisa M. Marvel, Member, IEEE, Charles G. Boncelet, Jr., Member,

IEEE, and Charles T. Retter, Member, IEEE, ‘ Spread Spectrum Image Steganography’,IEEE TRANSACTIONS ON IMAGE PROCESSING.

[7] T. Morkel , J.H.P. Eloff, M.S. Olivier, ‘an overview of image steganography’, Information and Computer Security Architecture (ICSA) Research Group Department of Computer Science University of Pretoria, 0002, Pretoria, South Africa

[8] Alain C. Brainos II. ― A Study Of Steganography And The Art Of Hiding Information‖, IEEE Trans. Inf. Forens. Secur. 2006

[9] Robert Krenn. ― Steganography and steganalysis,‖ Computer, vol. 31, no. 2, Feb. 1998,pp. 26-34.

[10] Udit Budhiaa, Deepa Kundura. ―Digital video steganalysis exploiting collusion sensitivity‖, IEEE Tans. On Image Processing,vol.15, No.8, August 2006, pp. 2441-2453.

[11] R. sridevi ―Efficient Method Of Audio Steganography By Modified Lsb Algorithm And Strong Encryption Key With Enhanced Security‖ Journal of Theoretical and Applied Information Technology 2005 – 2009 JATIT.

[12] Nagham Hamid, Abid Yahya, R. Badlishah Ahmad & Osamah M. Al-Qershi , ‘Image Steganography Techniques: An Overview’, International Journal of Computer Science and Security (IJCSS), Volume (6) : Issue (3) : 2012

[13] Anu, rekha, Praveen, ‘Digital Image Steganography ’International Journal of Computer Science & Informatics, Volume-I, Issue-II, 2011

[14] Cox, I., et al., “A Secure, Robust Watermark for Multimedia, ” in information Hiding: First International Workshop, Proceeding ,vol. 1174 of Lecture notes in Computer Science, Springer , 1996,pp.185-206.

[15] Koch, E., and J.Zhao, “Towards Robust and Hiddden Image Copyright Labeling”, in IEEE Workshop on Nonlinear Signal and Image Processing, Jun.1995.

[16] Gurmeet Kaur and Aarti Kochhar, “A Steganography Implementation based on LSB & DCT”, International Journal for Science and Emerging

[17] Technologies with Latest Trends”.

[18] Lisa M. Marvel, Member, IEEE, Charles G. Boncelet, Jr., Member, IEEE, and Charles T. Retter, Member, IEEE, ‘ Spread Spectrum Image Steganography’,IEEE TRANSACTIONS ON IMAGE PROCESSING.

[19] M. Kharazi, H.T. Sencar, and N. Memon. (2004, Apr.). “Image steganography: Concepts and practice.” Aug. 2011

[20] P. Kruus, C. Scace, M. Heyman, and M. Mundy. (2003), "A survey of steganography techniques for image files." Advanced Security Research Journal. Oct., 2011

[21] C.P.Sumathi, T.Santanam and G.Umamaheswari, 'A Study of Various Steganographic Techniques Used for Information Hiding' International Journal of Computer Science & Engineering Survey (IJCSES) Vol.4, No.6, December 2013.

[22] Stefan Katzenbeisser, Fabien A. P. Petitcolas, 'Information Hiding Techniques for Steganography and Digital Watermarking'.

[23] Jessica Fridrich*, Miroslav Goljan "Practical Steganalysis of Digital Images – State of the Art" supported by Air Force Research Laboratory, Air Force Material Command, USAF, under a research grant number F30602-00-1-0521.

[24] J. Fridrich, M. Goljan, P. Lisonek, and D. Soukal, "Writing on wet paper", IEEE Trans.on Signal Processing, Special Issue on Media Security, vol. 53, Oct. 2005, pp. 3923-3935.

[25] A. Joseph Raphael, "Cryptography and Stegano-graphy – A Survey" Int. J. Comp. Tech. Appl., Vol 2 (3), 626-630.

International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 484 - 487, 2014, ISSN: 2319–8656

Efficient Web Data Extraction

Yogita R.Chavan
University of Pune
KKWIEER
Nashik, India

Abstract: Web data extraction is an important problem for information integration as multiple web pages may present the same or similar information using completely different formats or syntaxes that make integration of information a challenging task. Hence the need of a system that automatically extracts the information from web pages is vital. Several efforts have already been carried out and used in the past. Some of the techniques are record level while the others are page level. This paper shows the work aims at extracting useful information from web pages using the concepts of tags and values. To avoid discarding of non-matching first node that represents non auxiliary information in the data region an efficient algorithm is proposed.

Keywords: auxiliary information, data extraction, DOM Tree, record alignment.

1. INTRODUCTION

Web information extraction is one of the very popular research activities aims at extracting useful information from web pages. Such extracted information is then stored into the database that can be used for faster access to the data. Due to the assorted structure of web data, automatic discovery of target information becomes a tedious task.

In order to extract and make use of information from multiple sites to provide value added services, one needs to semantically integrate information from multiple sources. Hence the need of a system that will automatically extract the information from web pages efficiently is vital.

The work aims at studying different web page extraction strategies/techniques and to implement the technique based on tag and value similarity as well as a few enhancements, if possible.

A method of record extraction is referred from CTVS proposed by W. Su et al [6]. This method is further modified using label assignment technique mentioned in DeLa [3] partly to overcome the drawback of not considering an optional attribute found in data region which cause the loss of information. This information is stored in temporary file during data region identification step and regions are then merged using similarity technique [11]. Applying the heuristics, only one data region is selected to extract exact result records. If information stored in temporary file belongs to this selected data region, it is segmented before final record extraction because of which the optional attribute that was not considered during data region identification is considered and information loss is prevented.

2. LITERATURE SURVEY

Due to the necessity and quality of deep web data web database extraction has received much attention from the Data mining and Web mining research areas in recent years. Earlier work focused on wrapper induction methods called as non-automatic methods require human assistance to build a wrapper. Wrappers are the hand coded rules i.e. a customized procedure of information extraction. In this method an inductive approach is used where user learns or marks part or all of the items to extract the target item containing set of training pages. A system then learns the wrapper rules and uses them to extract the records from the labeled data.

Advantages

-No extra data extracted

Disadvantages

- Labor intensive and time consuming

- Performs poorly when the format of query result page changes

- Thus, not scalable to a large number of web databases.

Systems WIEN, Stalker, XWRAP and SoftMealy follow wrapper induction technique.

More recently, automatic data extraction systems like RoadRunner, IEPAD, DeLa and PickUp have been proposed. C.H. Chang et al used a method of pattern discovery for information extraction that generates extraction rules which utilize a decoded binary string of the HTML tag sequence and tries to find maximal repeated patterns using a PAT tree which then become generalized using multiple string alignment technique. At the end the user has to choose one of these generalized patterns as an extraction rule.

This method identifies and extracts the data using repeating patterns of closely occurring HTML tags. It is convenient for set of flat tuples from each page and also produces poor results for complex and nested structure data structure[5].

V. Crescenzi, et al proposed a method for automatic data extraction that extracts a template by analyzing two web pages of an equivalent category at a time. In this method one page is used to derive initial template and it then tries to match the second page with the template.

Challenges of this method are deriving the initial template needs to be done manually [10].

Kai Simon et al used visual perceptions for automatic web data extraction that project the contents of the HTML page onto a 2-dimensional X/Y co-ordinate plane due to which it is able to compute two content graph profiles, one for each X and Y planes. These used to detect data regions by locating valleys between the peaks as the separation point between two data.

Drawback of this method lies in the assumption that the data regions are separated by the defined empty space regions. This may not always true [7].

Hongkun Zhao et al proposed a technique for fully automatic wrapper generation for search engines that extracts content line features from the HTML page, where this content line is a type of text which could be visually bounded by a rectangular box.

Several sample pages are used to extract the correct data region from the HTML page using parsing technique. But result records with irregular block structures are excluded in this method and also parsing the sample static and dynamic HTML regions become overhead.

3. IMPLEMENTATION

Algorithm

1. Query Result Page DOM Tree Construction

2. Data Regions Identification

3. Query Result Records Extraction

4. Records Alignment Pair Wise

5. Nested Structure Processing

6. Final Database Table

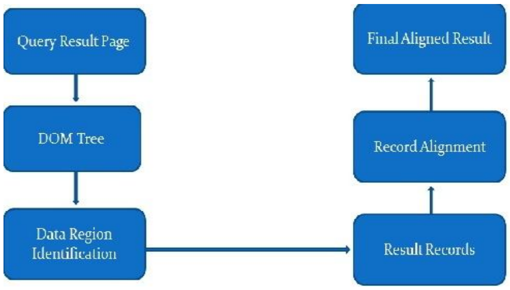


Figure 1: Block Diagram of the System

Implementation further divided into two main parts.

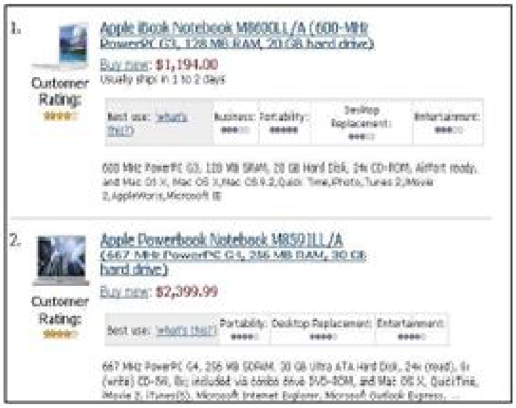
1. Records Extraction

2. Records Alignment

3.1 Record Extraction

3.1.1 Query Result Page DOM Tree Construction

From the source code associated with the page (that is HTML code), a DOM (Document Object Model) tree is to be constructed. Let us start with an example, the query result page for query- Apple Notebook Figure 2 shows a page with two images Apple iBook Notebook and Apple Powerbook Notebook after firing query Apple Notebook along with some non-useful information of links of advertisements. From the HTML source code, DOM tree shown in figure 3 will be generated.



1. Apple iBook Notebook M8600LL/A (600-MHz PowerPC G3, 128 MB RAM, 20 GB hard drive)
Buy new: $1,194.00
Usually ships in 1 to 2 days
Customer Rating:

Best use: what's this?	Business:	Portability:	Desktop Replacement:	Entertainment:

600 MHz PowerPC G3, 128 MB SRAM, 20 GB Hard Disk, 24x CD-ROM, AirPort ready, and Mac OS X, Mac OS X, Mac OS 9.2, Quick Time, iPhoto, iTunes 2, iMovie 2, AppleWorks, Microsoft IE

2. Apple Powerbook Notebook M8591LL/A (667 MHz PowerPC G4, 256 MB RAM, 30 GB hard drive)
Buy new: $2,399.99
Customer Rating:

Best use: what's this?	Portability:	Desktop Replacement:	Entertainment:

667 MHz PowerPC G4, 256 MB SDRAM, 30 GB Ultra ATA Hard Disk, 24x (read), 8x (write) CD-RW, 8x included via combo drive DVD-ROM, and Mac OS X, QuickTime, iMovie 2, iTunes(3), Microsoft Internet Explorer, Microsoft Outlook Express, ...

Figure 2: Query Result page for query - Apple Notebook

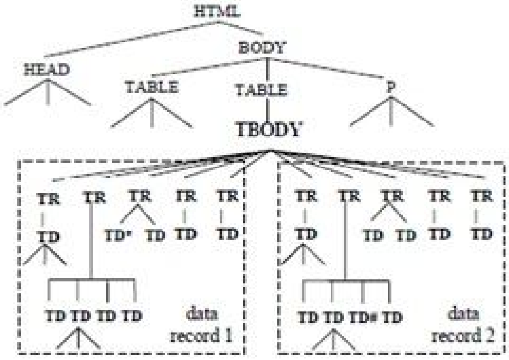


Figure 3: DOM Tree for query page- Apple Notebook

3.1.2. Data Region Identification

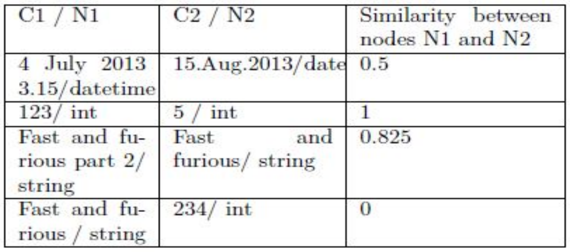
By calculating similarity of nodes, data regions are identified. Using node similarity algorithm where each node represents information and all of them form a graph. Finding their adjacency matrix, incidence matrix similarity between them is calculated. To calculate similarity the edit distance between the nodes is considered and two nodes n1 and n2 are similar if the edit distance between them is greater than or equal to threshold value 0.6 suggested by Simon and Lausen. Similar nodes are then recognized and nodes with the same parent form a data region. Multiple data regions may be formed during this step.

3.1.3. Query Result Records Extraction

Applying heuristics that the search result section usually located at the centre of the query result page and it usually occupies large space in query result page, a data region is identified to extract the information.

3.2 Records Alignment

For record alignment a novel method consisting of three consecutive steps for alignment proposed by W. Su [6] is referred. This method includes two steps



C1 / N1	C2 / N2	Similarity between nodes N1 and N2
4 July 2013 3.15/datetime	15.Aug.2013/date	0.5
123/ int	5 / int	1
Fast and furious part 2/ string	Fast and furious/ string	0.825
Fast and furious / string	234/ int	0

3.2.1. Pairwise Alignment

The similarity between two data values f1 and f2 with data type nodes n1 and n2 is defined as where p(n i) is the parent node of ni in the data type tree, sample of which is shown in figure 4 below.

In the first row of above table, in column C1, information is of type date time where as in column C2 only date is mentioned. Referred formula and data type tree, datetime is the parent node of date node.

Condition N1 = p(N2) and N1 is not string satisfies and similarity 0.5 in considered in column 3. In the second row of the example, both values are of integer type, so similarity one is entered. In the third row, both values are of string type so cosine similarity is taken into consideration. In the last as both values are of different types zero similarity is considered.

3.2.2. Nested Structuring Alignment

After the first step of pairwise alignment all data values of the same attribute are put into the same table column. The logic of finding connected components of an undirected graph is used for this purpose. In nested structuring multiple data values of the same attribute are put in the different row of the table.

In the end, the information will be stored in the form of a table.

4. RESULTS and DISCUSSION

4.1 Data set

1. E- COMM contains 100 deep websites E-commerce in six well-liked domains such as hotel, job, movie, automobile, book and music whereas each domain has 10 to 20 websites.
2. PROFUSION contains 100 websites collected from profusion.com

Above datasets can be used to evaluate the working of the system.

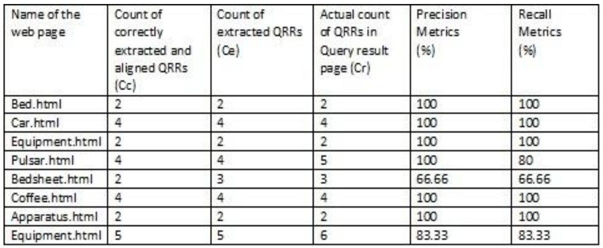
4.2 Result Set

The implementation is done in java using Netbeans where user is allowed to enter a query result page as an input. Above mentioned datasets can be used for the same purpose. And results are then compared with existing systems results. Proposed methodology is used to provide the better results preventing loss of information. When applied on first 10 web pages, the table of results is, where precision metrics= Cc/Ce and recall metrics= Cc/Cr

5. CONCLUSION AND FUTURE WORK

An efficient method for web data extraction is proposed that includes finding data regions and also considering optional attribute (non-auxiliary information) node value and further add it in the final database table. This overcomes the drawback of loss of information in a data region. This increases the performance of the system by extracting the information more effectively.

In this research, it is aimed to obtain extraction of web page's information accuracy by using the efficient algorithm. For this purpose, several fully automatic web extraction approaches are investigated. Extensive studies are done on these approaches to explain why they do not achieve satisfactory data extraction outcome. After performing several experiments as described in the result table, it is observed that efficiency of the system has increased. Second, any optional attribute that appears as the start node in a data region will not be treated as auxiliary information.



Name of the web page	Count of correctly extracted and aligned QRRs (Cc)	Count of extracted QRRs (Ce)	Actual count of QRRs in Query result page (Cr)	Precision Metrics (%)	Recall Metrics (%)
Bed.html	2	2	2	100	100
Car.html	4	4	4	100	100
Equipment.html	2	2	2	100	100
Pulsar.html	4	4	5	100	80
Bedsheet.html	2	3	3	66.66	66.66
Coffee.html	4	4	4	100	100
Apparatus.html	2	2	2	100	100
Equipment.html	5	5	6	83.33	83.33

This research has found that the system outperformed the existing web data extraction systems.

Along with the advantages, this method has shortcomings like it requires at least two query result records in the result page as for forming a template at least two result records expected and the other is while selecting a single data region depending on heuristics discussed other data regions are discarded which may contain useful information needs to be stored in database table. These drawbacks would be tried to be removed in the future.

6. REFERENCES

[1] A. Arasu and H. Garcia-Molina, *Extracting Structured Data from Web Pages*, Proc. ACM SIGMOD International Conference Management of Data, Pp. 337-348, 2003

[2] R. Baeza-Yates *Algorithms For String Matching: A Survey*, ACM SIGIR Forum, Vol. 23, Nos. 3/4, 34-58, 1989

[3] J. Wang And F.H. Lochovsky *Data Extraction and Label Assignment For Web Databases*, Proc. 12th World Wide Web Conference Pp. 187-196,2003.

[4] Y. Zhai And B. Liu *Structured Data Extraction From The Web Based On Partial Tree Alignment*, IEEE Trans. Knowledge and Data Eng., Vol. 18, No. 12 Pp.1614-1628, Dec. 2006

[5] C.H. Chang and S.C. Lui *IEPAD: Information Extraction Based On Pattern Discovery*, Proc. 10th World Wide Web Conference Pp. 681-688, 2001.

[6] Weifeng Su, Jiying Wang *Combining Tag and Value Similarity For Data Extraction and Alignment*, IEEE Transaction On Knowledge And Data Engineering, Vol.24, No.7, July 2012

[7]] K. Simon And G. Lausen *VIPER: Augmenting Automatic Information Extraction With Visual Perceptions*, Proc. 14th ACM International Conference Information and Knowledge Management Pp. 381-388,2005.

[8] Y. Zhai And B. Liu *Structured Data Extraction From The Web Based On Partial Tree Alignment*, IEEE Trans. Knowledge and Data Eng., Vol. 18, No. 12 Pp.1614-1628, Dec. 2006.

[9] D. Buttler, L. Liu and C. Pu *A Fully Automated Object Extraction System for the World Wide Web,* Proc. 21st International Conference Distributed Computing Systems Pp. 361-370, 2001

[10] V. Crescenzi, G. Mecca and P. Merialdo *Roadrunner: Towards Automatic Data Extraction from Large Web Sites,*

Proc. 27th International Conference Very Large Data Bases Pp. 109-118, 2001

[11] Miklos Erdelyi, Janos Abonyi *Node Similarity-Based Graph Clustering and Visualization*, 7th International Symposium of Hungarian Researchers on Computational Intelligence