Summer 4-28-2012

Proceeding of International Conference on Computer Science And Engineering ICCSE-2012

Prof. (Dr.) B.V. Dhandra

Editorial

The mushrooming growth of the IT industry in the 21st century determines the pace of research and innovation across the globe. In a similar fashion Computer Science has acquired a path breaking trend by making a swift in a number of cross functional disciplines like Bio-Science, Health Science, Performance Engineering, Applied Behavioral Science, and Intelligence. It seems like the quest of Homo Sapience Community to integrate this world with a vision of Exchange of Knowledge and Culture is coming at the end. Apparently the quotation "Shrunken Earth, Shrinking Humanity" holds true as the connectivity and the flux of information remains on a simple command over an internet protocol address. Still there remains a substantial relativity in both the disciplines which underscores further extension of existing literature to augment the socio-economic relevancy of these two fields of study. The IT tycoon Microsoft addressing at the annual Worldwide Partner Conference in Los Angeles introduced Cloud ERP (Enterprise Resource Planning,) and updated CRM (Customer Relationship Management) software which emphasizes the ongoing research on capacity building of the Internal Business Process. It is worth mentioning here that Hewlett-Packard has been with flying colors with 4G touch pad removing comfort ability barriers with 2G and 3G. If we progress, the discussion will never limit because advancement is seamlessly flowing at the most efficient and state-of-the art universities and research labs like Laboratory for Advanced Systems Research, University of California. Unquestionably apex bodies like UNO, WTO and IBRD include these two disciplines in their millennium development agenda, realizing the aftermath of the various application projects like VSAT, POLNET, EDUSAT and many more. 'IT' has magnified the influence of knowledge management and congruently responding to social and industrial revolution.

The conference is designed to stimulate the young minds including Research Scholars, Academicians, and Practitioners to contribute their ideas, thoughts and nobility in these two integrated disciplines. Even a fraction of active participation deeply influences the magnanimity of this international event. I must acknowledge your response to this conference. I ought to convey that this conference is only a little step towards knowledge, network and relationship.

Editor-in-Chief
Prof. (Dr.) B.V. Dhandra,
Department of Computer Science
Gulbarga University
Gulbarga

A Study of Advanced Comprehensive Ontology Visualization Tools

V. Swaminathan & R.SivaKumar

Department of Computer Science, A.V.V.M. Sri Pushpam College, Bharathidasan University, Trichirappalli, India
E-mail : vswaminathanthanjavur@ yahoo.com, rskumar.avvmspc@ gmail.com

Abstract - The rapid grown of documents, web pages and other types of textual content page a great challenge to modern content management systems. Ontologies offer an efficient way to reduce information overload by encoding the structure of a specific domain thus offering easier access to the information. The continuing need for more effective information retrieval has lead to the creation of the notions of the semantic web and personalized information management, areas of the study that very offer employ ontologies to represent the semantic context of a domain. Consequently, the need for effective ontology visualization for domain management and browse has arisen. Ontology comprehension is a collection of techniques that facilitative the understanding of ontologies. These tools are being applied for further development in various disciplines for better understanding knowledge. The purpose of the work is to present a study on advanced comprehensive ontology visualization tools and categorize their characteristics so that it assists in method selection and promotes further research in the area of ontology visualization. This paper overviews different technology for ontology visualization.

Keywords: *Comprehensive Ontology, Ontology visualization technique, Cognitive Support, Human agents*

I. INTRODUCTION

Visualization is used as a cognitive aid for managing ontology and knowledge representations. The continuing need for more effective information retrieval has lead to the creation of the notions of the semantic web and personalized information management areas of study that very often employ ontology to represent the semantic context of a domain. The Ontology is machine understandable and thus needs some means of graphical visualization for humans to comprehend. Consequently, the need for effective ontology visualization for design, management and browsing has arisen. Visualization tools make it easy to understand large and complex ontology important for easy representation of selected parts. There are several ontology visualizations available through the existing ontology management tools. Ontology is a term initially borrowed from Philosophy, where ontology is a systematic account of existence that is trying to answer the question 'what properties can explain the existence'. In computer science according to Gruber [1], ontology is an explicit specification of a conceptualization. The term "conceptualization" is defined as an abstract, simplified view of the world, which needs to be represented for some purpose. It contains the objects, concepts, and other entities that are presumed to exist in some area of interest, and the relations that hold among them. That is Ontology describes basic concepts in a domain and defines relations among them Many ontology visualization techniques have been already developed such as Protégé class browser, OntoVis, IsaViz, Space Tree, Onto Sphere, Jambalaya, Crop Circles, GopherVR, Protégé TGVizTab and 3D Hyper- bolic Tree. They are reported in literature [3, 4]. However, the research society is in need of knowing the recent developments and advances in ontology tools. Hence this paper focuses on studying the features of recently developed tools such as Altova Semantic Works, FlexViz, Knoodl-OntoVis, Ontopia, Collaborative Protégé and Wandora, which may help the researchers to think in different directions.

The remaining part of the paper is organized as follows: Section 2 presents the overview of ontology comprehensive and standard techniques. Next, Section 3 presents complete survey of different ontology tools. A ccomparative report on advanced features is also presented in Section 4. Finally Section 5 concludes this paper.

II. ONTOLOGY COMPREHENSIONS

The term ontology comprehension is used in different contexts with different meanings by different authors (for example [2, 3, 4]). For this reason it is important to clarify the meaning of ontology comprehension within the context of this work. Ontology comprehension is the interaction between

human agents and the knowledge expressed in ontology. Ontology comprehension is a collection of techniques that facilitate the understanding of ontologies. An ontology understanding technique is an abstract idea that can have many implementations. For example, two computer programs can implement the same technique in different ways [5].Ontology understanding techniques can operate independently, or can interact to enhance understanding. A depiction of ontology comprehension frame work is given in Fig. 1.

2 .1 Standard Techniques

The visualization techniques are presented in tree or graph representation for the presentation of all existing ontology management tools in existing ontology visualizations.

For a method to be eligible for the visualization of an ontology, it has to support the presentation of ontology ingredients i.e. classes (or entity types), relations, instances and properties (or slots).

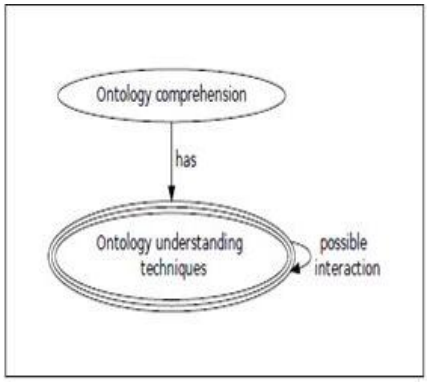


Figure 1: Ontology Comprehension framework.

The methods can be grouped according to different characteristics of the presentation, interaction technique, functionality supported or visualization dimensions. For the needs of this survey the methods were grouped in the following categories, representing their visualization type:

1. Indented List
2. Node–link and Tree
3. Zoomable
4. Space–filling
5. Focus + Context or Distortion
6. 3D Information Landscapes

The methods grouped in these six general categories were further categorized according to the number of space dimensions they employ, i.e. 2D and 3D. The 2D methods use the screen space as a plane and do not use any notion of depth. The 3D methods exploit the third dimension either to create visualizations that are closer to real world metaphors or to improve usage of space and/or usability. More specifically, these methods allow the user to manipulate—rotate and move—3D objects and/or to navigate inside the 3D space.

This second-level grouping was chosen due to the specific needs that characterize the 3D visualizations which are also reflected upon the interaction techniques employed and functionality which can be catered for target user group characteristics and even system requirements. 3D visualization, in general , requires increased system resources in order for navigation and viewing to be smooth and without delays and, as a result, is probably not suitable for web use .

3. SURVEY OF ADVANCED COMPREHENSIVE ONTOLOGY VISUALIZATION TOOLS.

This section surveys the advanced comprehensive ontology visualization tools with a scope of extended features

3.1 Altova Semantic Works

Altova Semantic Works®2012 is the ground-breaking visual RDF and OWL editor for the Semantic Web. Graphically designed RDF instance documents, RDFS vocabularies, and OWL ontologies, then output them in either RDF/XML or N-Triples formats. Semantic Works make the job easy with tabs for instances, properties, classes, etc., context-sensitive entry helpers, automatic format checking, and more. RDFS vocabularies define the allowable properties (predicates) for RDF instances within a particular domain. RDFS also allow defining classes to further classify the relationships between resources. Semantic-Works displays the instances, properties, and classes in an RDFS vocabulary on separate tabs, allowing one to view and edit these different items with ease. The Instance tab lists all resources in the document, and the Properties tab lists all the properties. When a property is selected in the main pane, the domain of that property is displayed in another window. The Class tab lists all the classes available in the vocabulary with a separate window that lists the instances and properties of the selected class as they appear throughout the Redstone can view and edit the details of any item listed by clicking its expand button. Semantic Works display resources graphically according to their associations with other resources. The Semantic Works display is highly configurable. One can adjust the width of the items in the graph, display it with a vertical or horizontal orientation, adjust the distances between parent and child nodes, and even change the font styles and colors

used. To help immediately visualize class relationships, RDFS classes are enclosed in yellow boxes in the graphical display. Holding the mouse over any item or icon display reveals its meaning or corresponding URI. The same entry helpers and context-sensitive choices described in the RDF editing section above are available for RDFS editing, and syntax checking based on the RDFS specification ensure that one's document is valid. To creating a visual RDFS design, Semantic Works R 2012 is auto-generating the corresponding RDF/XML or N-Triples code behind the scenes, and one can view and edit it at any time by clicking the Text tab.

3.2 FlexViz

Graphs are the basic concepts in discrete mathematics and data structure. The applications of graphs are very extensive and vary from common events to complex mathematical or computer science problems. The building blocks of a graph are vertices (nodes) and edges. Ontologies can be represented as a graph by using the basic properties of graphs, such as directed graphs. FlexViz is a graph based visualization tool written is Adobe Flex. It supports single ontology browsing. Nodes of graph are mapped as concepts and relationships (edges) between nodes (e.g. "is a", "depends on") are represented as arcs. It is designed to provide a light-score, interactive, and visually accessible ontologies on Web. [9] Figure 2 shows a demo of FlexViz, web-based visualization primarily render static images or text representations. However, FlexViz greatly enhances user tasks and promotes new technologies by making it very interactive and light-weight.

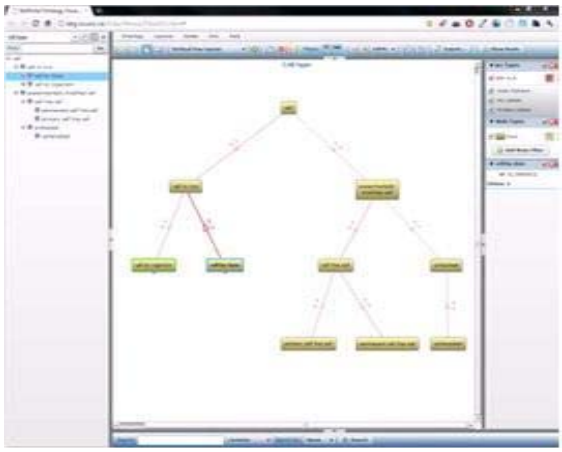


Figure 2: FlexViz demo

[9] FlexViz has many interesting features such as filtering based on nodes (concepts) and edges (relationships), searching, results view in various graphical layout, customization of node and edge (arc) labels , customizing node and arc tool tips, zooming , forcibly stay of node in the screen ,back and forward button to navigate history, nodes expansion to show or hide children, colors of nodes and customization of edges, visualization widget with a fixed ontology and export of graph as a image data or xml file.

3.3 Knoll–Otis

One can start up an Otis page by going to any resource page in Knoll and clicking the "graph" tab. On the main ontology page the "graph tab" takes the user to a blank visualization. From any resource page, the graph will load with that resource in place. The class and property trees are crucial for making diagrams in Otis. If the tree is not visible upon load, the user has to select it from the "View" menu. Once the tree menu is loaded, one can drag classes and properties from it onto the Graph section. The only way to view graphs containing instances is to click the graph tab on the knoll resource page for that instance. The toolbar helps to change the size of the graph, and also lets switch between the pointer and the eraser . To view relations and annotations for resources, the user has to click the three-bar menu symbol that appears when the mouse hovers over the resource. The menu that appears varies depending on whether the resource in question is a class, property, or instance, There are three options in classes first one, Show related classes which places subclasses, disjoint classes, and other related classes on the graph and in some cases, the relations that appear will make use of resources already on the graph and the second one, Show annotations, expands the resource's q2wsdef;klop'container to show its OWL annotations and the final one, Show properties, expands the resource's container to show related properties. The properties appear inside the resource box, may be dragged onto the graph. Note that properties, when dragged onto the graph, will appear as edges rather than as nodes. If one can want to view a property as a resource, one must drag it out from the tree view and also related properties are properties whose domain is declared to be the class in question, as well as properties with property restrictions declared on the class With properties, first one Show related classes , that places sub properties, inverse properties, and other related properties on the graph, in some cases, the relations that appear will make use of resources already on the graph. It also places the classes declared to be the domain and range of the property in question on the graph. With Instances, the first option, Show related instances, places related instances on the graph. The next one, Shows class assertions , expands the resource box to show which classes the instances is a member of and here the Class assertions may be dragged onto the graph, It also shows property assertions and shows

properties asserted on this instance and their values where property assertions may be dragged onto the graph. Once the graph looking the way one wants, the user can choose "Export" under the "File" menu. One can choose PNG or PDF as the format.

3.4 Ontopia

Ontopoly is built as a client/server application. As a client, to use web browser, while the server is a web server bundled with the distribution. The server-side application is built using the Navigator Framework and Web Editor Framework, which are parts of the Ontopia Knowledge Suite. Ontopoly is accessible within Ontopia, an application that provides easy access to Ontopoly, Dominator, and Litigator. Monopoly's primary purpose is to enable the manual creation and maintenance of topic maps that may be based on a variety of ontologies. In order to be able to provide the most intuitive possible user interface for such a generic application, Ontopoly is ontology-driven. What this means is that the forms-based interface for creating and maintaining a topic map is generated automatically from the underlying ontology and the rules that are defined for it. Ontopoly is divided into two main parts: the ontology editor and the instance editor. The application also has an administration interface, i.e., a page for adding metadata to the topic map (description, creator, version, etc.), validating it, etc.; and an interface for exporting to various interchange syntaxes. The Topic Map Index Page allows the user to open an existing topic map, create a new one, or import one from outside Ontopoly. Once selected one of these actions, then it will take to the application pages. The overall architecture and the navigation paths between the various parts of the application are shown in Fig. 3. Monopoly's two primary functions provide a comprehensive Topic Maps editing environment such as configuring a topic map and Populating a topic map. Topic Map Index Page: the user before creates either the ontology or the instance should open a topic map. When one can first access Ontopoly from Ontopia, then take it to the Topic Map Index Page, on the left, is a column for Ontopoly Topic Maps, in the middle is the list of Other Topic Maps, and on the right is the area for creating new topic maps. Another column, Missing Topic Maps, will also appear on this page if an Ontopoly topic map has been deleted from outside syntaxes; it will also import RDF. A topic map created in Ontopoly (an Ontopoly topic map) will differ from one created outside of it (a non-Ontopoly topic map). An Ontopoly topic map carries along with it topics that let it interact with the Ontopoly application and topics that define the schema (the system topics). One can always export a topic map from Ontopoly to remove the system topics, but they are needed for the topic map to be understood within the application Similarly, pre-existing non-Ontopoly topic Maps will need these system topics to be added.

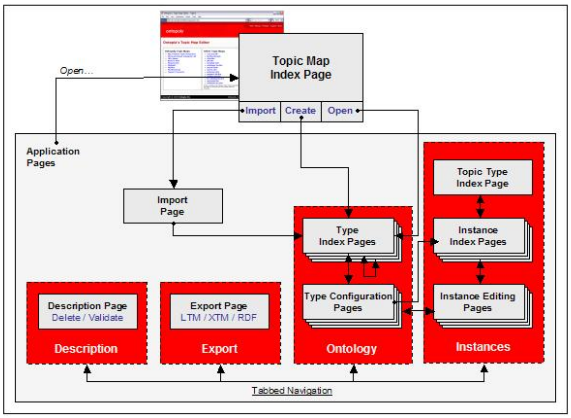


Figure 3: Overview of Monopoly's architecture

There are five kinds of types in Topic Maps: topic types, occurrence types, association types, role types, and name types. Each of these has its own kind of Type Configuration Page, which is accessed via the links at the top on the Type Index Pages. Each Type Configuration Page is unique, but some fields are common to all of them. They are Name, Subject identifier, Read-only and Hidden. The additional properties specific to topic types are Abstract, Super class and Subclass.

3.5 Collaborative Protégé

Collaborative Protégé is an extension of the existing Protégé system that supports collaborative ontology editing. In addition to the common ontology editing operations, it enables annotation of both ontology components and ontology changes. It supports the searching and filtering of user annotations, also known as notes, based on different criteria. One can implement two types of voting mechanisms that can be used for voting of change proposals.

Multiple users may edit the same ontology at the same time. In multi-user mode, all changes made by one user are seen immediately by other users. There are two working modes available for Collaborative Protégé. Both modes such as multi-user mode and standalone mode support multiple users working on ontology. The main feature of Collaborative Protégé is the ability to create notes attached to different things. This is the same idea as if someone would read an article and would add marginal notes on the paper. In the same way, the notes mechanism of Collaborative Protégé allows a user to create his own remarks about a certain part of the

ontology. This feature can also be used to discuss the ontology with other users either in standalone or multi-user mode. The notes are also called annotation. The main functionalities of Collaborative Protégé are annotation of classes and properties and instances with different types of notes (e.g., Comment, Advice, Example, etc.)

The classical Protégé display shows the details of the different ontology components, such as classes, properties (slots) and instances. The collaboration panel adds functionality to support the collaborative development of ontologies. The collaboration panel is made up of two types of tabs such as Entity notes & Changes tabs and Ontology notes, all notes, Search and Chat tabs. In the main Protégé display user will also see the call-out icon if a property (slot) or an instance have annotations attached to them. A user may add notes attached to a specific ontology component in the Entity Notes. By ontology component, to mean classes, properties (or slots) and instances. The image shows an example of note attached to the Domain Concept class. The number of notes attached directly to Domain Concept is shown next to the Notes icon.

3.6<New> Wandora

Wandora is a general purpose information extraction, aggregation, management, and publishing application based on Topic Maps and Java. Wandora has graphical user interface, layered presentation of knowledge, several data storage options, huge collection of data extraction, import and export options, embedded server, and open plug-in architecture. Wandora is a FOSS application with GNU GPL license. Wandora is well suited for constructing information mashups and is capable of extracting and converting a wide range of open data feeds to Topic Map formats. Beyond Topic Maps conversion this feature allows Wandora user to aggregate multidimensional information mashups where information from Flicker interleaves with information from Geo Names and YouTube, for example. Pandora's embedded HTTP server enables easy mashup publication. Wandora can export Topic Maps in Grapheme format. The export starts with File > Export > Export Grapheme. Exported graph can be visualized with hyper graph application for example.

IV. COMPARATIVE REPORT ON ADVANCED FEATURES

Altova supports the following bugs with its 2012 release: Arabic characters not accepted as class names by Semantic Works, classes in taxonomy are declared invalid, enable import of RDF schema in OWL Full, semantic check fails to note incorrect object property , URI ref prefixes toggles on/off spontaneously when editing file in Text View, syntax and semantic check in Semantic Works incorrectly report error on missing type in instance (OWL DL), semantic works crashes when hitting 'del' in the instances detail view while the drop down menu to change a data type is open, export function does not check for validity, failed verification of equivalence class extension, and rdf:XMLLiteral is not offered as data type for a value in RDF/OWL view. On the other hand, Flexi has many interesting features that can be summarized as follows: Filtering based on nodes (concepts) and edges (relationships), searching, results may be viewed in various graphical layout, customization of node and edge (arc) labels is possible, customizing node and arc tool tips, zooming, forcibly staying node in the screen (can cause nodes to overlap), back and forward button to navigate history, nodes can be expanded or collapsed to show or hide children, colors of nodes, edges are customizable, visualization can be displayed as a widget on a web page with a fixed ontology, graph can be exported as a image data or a xml file, source code is released in SourceForge9 which offers customization and extension based on requirement. With Ontvis, one can feel easy-to-use and easy-to-visualize capabilities to show the semantic contents, resource page, graph tab, manipulating the resources in the graph, edge decorations and exporting. As far as Ontopia is concerned it has the following advantages: detailed information of a topic and applicable to any Topic Map and can be generated automatically fast and cheap.

It has also the following disadvantages: for every node a new graph is generated constantly changing visualization, user cannot create a mental model of the information structure, no big picture, usability is poor and only one visualization concept. The Collaborative Protégé has the following advantages: Adding shared notes to ontology, discussion threads, view and discuss changes made to an ontology and live chat during editing. Wandora offers extractors with the following facts: Search engine extractors, HTML structures extractors, simple files extractors, news and syndication extractors, media extractors, micro format extractors, social & bookmark extractors, wake extractors, bibliographical extractors and simple RDF extractors. Similarly it provides various following generators: random graph generator, fully connected graph generator, tree graph generator, linear list graph generator, finite group graph generator, platonic solid graph generator, hypercube graph generator, tiling graph generator (square, triangular, and hexagonal tiling), edge generator and L-system generator.

Table 1: Summary of the ontology visualization characteristics in comprehensive tools.

Tool Name	Month & Year of implementation	Classes and Instances	Taxonomy	Multiple Inheritance	Role relations	Visualization type
Altova Semantic-Works	Oct. 2001	Class tab lists all the classes available in the vocabulary with a separate window that lists the instances	Graph View	Child nodes are placed under both parents	No. Supported through the properties window only.	Intended List
FlexViz	Jan. 2008.	No, Filtering based on nodes and edges	Graph View	No	Represented as arcs	**Zoomable**
Knoodl OntVis	Mar. 2011	Classes are represented as nodes & graph Tab	Tree view	No	No	Zoomable
ontopia	May 2010	Class and instance are created by populating by Topic map	Tree view	Subtypes will be inherited of its fields	Using parent/child combination	Node-link and Tree
Collaborative Protégé	Feb. 2009	Classes and Instances are Presented as Nodes in an indented, expandable and retractable tree.	Tree view	No	No	Node-link and Tree
⟨**New**⟩ Wandora	Sep. 2011	class-instance relations as edges.	Graph View	No	No	Zoomable

V. CONCLUSION

In this paper, comprehensive ontology tools using ontology presentation characteristics were surveyed and analyzed. The visualization of comprehensive ontology tool is a particular sub problem of this area with many implications due to the various features that ontology visualization should present. The current work is an attempt to summarize the research that has been done so far in this area, providing an overview of the comprehensive ontology tools. As seen from the survey and the information provided in the table 1, it can be concluded that there is no one specific method that seems to be the most appropriate for all applications and, consequently, a viable solution is providing the user with several visualizations, so as to be able to choose the one that is the most appropriate for one's current needs.

REFERENCES

[1] Gruber. T.R., (1993), "A Translation approach to portable ontology specifications, knowledge acquisition special issue: current Issues in knowledge modeling", Vol. 5. Issue 2, 199–220.

[2] Chandra. V.K, Stickle. M.E, Thom ere. J.F, Salinger. R.J, (2000) et al. "Using prior knowledge: Problems and solutions". In National conference on artificial intelligence, pp. 436–442.

[3] Gibson. A, Wolstencroft. K and Stevens. R, (2007), "Promotion of ontological comprehension: Exposing terms and metadata with web 2.0", In Workshop on social and collaborative construction of structured knowledge at 16th International World Wide Web Conference, Retrieved on 2010/01/01 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=11.142.8462{\&}rep=rep1{\&}type=pdf.

[4] Keet. C, (2007), "Enhancing comprehension of ontologies and conceptual models through abstractions", Artificial intelligence and human-oriented computing, pp. 813–821.

[5] Johann Rath Bergh, (2010), "Ontology Comprehension", Division of Computer Science Stellenbosch University, Private Bag X1, 7602Matieland, South Africa.

[6] http://www.altova.com/semanticworks.html.

[7] http://www.altova.com/semanticworks/rdf-editor.html.

[8] Vivek Srivastava, (2011), "Methods to visualize ontology", Koblenz. [9] Sean M. Falconer, Chris Callendar, and Margaretanne Storey, (2009)

"FLEXVIZ: Visualizing Biomedical Ontologies on the Web", In International Conference on Biomedical Ontology, Software Demonstration, Buffalo, NY.

[10] http://knoodl.com/ui/groups/knoodl/wiki/Help/entry/diagrams.

[11] Ontopoly: The Topic Map Editor-User's Guide.

[12] A Practical Guide to Building OWL Ontologies Using Prot´eg´e 4 and CO-ODE Tools (2009).

[13] Noy. N. F, Chugh. A, Liu. W, and Musen. M. A, (2006), " A framework for ontology evolution in collaborative environments", In 5th Intl. Semantic Web Conference, ISWC, volume LNCS 4273, Athens, GA, Springer.

[14] Tudorache. T,. Noy. N. F, and Musen. M. A, (2008), "Supporting collaborative ontology development in protégé", In 7th Intl. Semantic Web Conference, ISWC 2008, Karlsruhe, Germany.

[15] Tania Tudorache, Jennifer Vendetti, Natalya F. Noy, (2008), "Web Protégé: A Lightweight OWL Ontology Editor for the Web".

[16] http://protegewiki.stanford.edu/wiki/WebProtege

[17] http://protegewiki.stanford.edu/wiki/Collaborative Protege

[18] http://www.wandora.org/wandora/wiki/index.php?httitle= Image: Exportxmlexample.gif.

❑❑❑

Utilization of Timestamp Records For An Hospital Centralization With Consideration of Emergency Cases

Phani Kishore Rompicharla, Vadde Swathi, Lanke Naga Baby Jyotsna, Kanakollu Naga Sai Deepika

Department of Computer Science and Engineering,
Dhanekula Institute of Engineering & Technology, Ganguru, Vijayawada, India
E-mail : phani.rompicharla@gmail.com, potluri.swathi@gmail.com, naga_jyotsna@yahoo.com, kns.deepika@gmail.com[4]

Abstract - These days, enhancing the quality of healthcare services, such as shortening waiting time and/or providing open-access policy, becomes an important issue even in mid-size hospitals. Even in small to midsize clinics, computerized health care management systems have replaced traditional paper based patient charts and have stored into a database not only patient information but also service-process related information. We introduced some departments like tools, medicine and emergency in order to reduce the patient waiting time deals with the issue of utilizing such per-existing but incomplete daa for simulation study. Such an assumption may be justified in cases were data requirement for simulation is precisely defined and all necessary data have been collected according to such requirement. Aiming at the healthcare management systems that maintain the log of operation activities using timestamps. We considered the reduction of patient waiting time as main problem criteria and worked out for it. Traditional simulation studies of a hospital centralization is very often assume only till the data centralization and data assumption. Aiming at any of the multi-specialty hospitals of INDIA, to reduce the time wasted for the patients we proposed the present system which can suggest te timestamp that the patient should wait until the doctor will be free for that particular patient. We also designed by consideration of emergency case interruption for the doctors.

I. INTRODUCTION

1.1 TERMINOLOGY:

In this section we first want to introduce the different terms that we were going to use in our paper as fallows.

Time Stamp : A time stamp is a sequence of characters, denoting the date and/or time at which a certain event occurred. A timestamp is the time at which an event is recorded by a computer, not the time of the event is recorded by a timestamp (e.g., entered into a log file) should be very very close to the time of the occurrence of event recorded.

Open Access Policy: Open access (OA) refers to unrestricted access via the internet to the required hospital website. It is more advanced and very easy method to admit in a hospital now a days even md-size hospitals are providing these services.

Transition Rate: In this particular research paper the world transition rate means the time taken by the patients to get through from one department to other department of the hospital.

Arena: It is a simulation software that is used for the represents of the any/all simulation processes. Arena is a discrete event simulation software simulation and automation software developed by Systems Modeling and acquired by Rockwell Automation in 2000. It uses the SIMA processor and simulation language. As of 2010, it is in version 13.0. It has been suggested that Arena my join other Rockwell software packages under the Factory Talk brand. In Arena, the user builds an experiment model by placing modules (boxes of different shapes) that represent processes or logic. Connector lies are used to join these modules together and timing, the precise representation of each module and entity relative to real-life objects is subject to the modeler.

Service Time: In this particular needed service i.e., to meet the doctor for consultation in the hospitals.

Arrival Rate: The mean number of new calling units i.e., the new patients arriving at a service facility per unit time.

Sensor Networks: Simulation studies of outpatient clinics often involve significant data collection challenges. We describe an approach for data collection using sensor networks which facilitates the collection of a large volume of very detailed patient flow data through healthcare clinics. Such data requires extensive preprocessing before it is ready for analysis. We present a general data preparation framework for sensor network

generated data with particular emphasis on the creation and analysis of patient path strings.

Magnetic Resonance Imaging: MRI is a fairly new technique that has been used since the begning of the 1980s. The MRI scan used magnetic and radio waves, meaning that there is no exposure to X-rays or any other damaging forms of radiation. An MRI scan is also able to provide clear pictures of parts of the body and spinal cord. Because the MRI scan gives very detailed pictures it is the best technique when it comes to finding tumors (begin or malignant abnormal growths) in the brain. If a tumor is present the scan can also be used to find out if it has spread into nearby brain tissue.

Computed Tomography: A CT scan is a method of taking an image of brain. It is a procedure that produces a clear, two-dimensional image of the brain that shows abnormalities such as brain tumors, blood clots, strokes, or damage due to head injury. A CT scan can help identify the cause of Alzheimer like symptoms either by finding an abnormality or by ruling out certain conditions.

Out Patients: people waiting for consultations or procedures not admitted to hospital are defined as outpatients. Outpatient surgery, also known as ambulatory surgery, same-day surgery or day surgery, is surgery that does not require an overnight hospital stay. The term outpatient arises from the fact that surgery patients may go home and do not need an overnight hospital bed. The purpose of outpatient surgery is to keep hospital costs down, as well as saving the patient time that would otherwise be wasted in the hospital.

Distribution: In mathematicl analysis, distributions (or generalized functions) are objects that generalize functions. Distributions make it possible to differentiate functions whose derivatives do not exist in the classical sense. In particular, any locally integral function has a distributional derivative. Distributions are widely used to formulate generalized solutions of partial differential equations. Where a classical solution may not exist or be very difficult to establish, a distribution solution to a differential equation is often much easier. Distributions are also important in physics and engineering where many problems naturally lead to differential equations whose solutions or intial conditions are distributions.

Appointment Scheduling: Based on the patients incoming and outgoing rates and the time available for the doctors the appointment i.e., time given to meet the doctor is scheduled in a hospital. This is called appointment scheduling.

Simulation: Simulation is the imitation of some real thing available, state of affairs, or process. The act of simulating some thing generally entails representing certain key characteristics or behaviors of a selected physical or abstract system. Simulation can be used to show the eventual real effects of alternative conditions and courses of action. Simulation is also used when the real system cannot engaged, because it may not be accessible, or it may be dangerous or unacceptable to engage, or it is being designed but not yet built, or it may simply not exist.

Incoming Logic: There are three parts that we use in the arena software. The incoming logic is a first one and the incoming logic simulates the incoming time rates of outpatients.

Lobby Logic: There are three parts that we use in the arena software. The lobby logic is the second one and the lobby logic simulates the patient movements by interconnecting the incoming logic and the process logic.

Process Logic: There are three parts that we use the arena software. The process logic is the third one and the process logic simulates hospital services information about patient incoming rates, transition rates, and service time are needed for simulation model.

Capacity Planning: Here capacity means the capacity of time or staff and all other modules that effects the patients waiting time in a hospital. Capacity planning is the task of managing the time or staff in order to reduce the waiting time of patients.

Electronic healthcare management system: EHMS is the representation of data in electronic medical records. An electronic medical record (EMR) is a computerized medical record that is created in an organization that delivers care, such as a hospital or physician's office. Electronic medical records tend to be a part of a local stand-alone health information system that allows storage, retrieval and modification of records.

Financial Payoff: The amount necessary to pay a loan in full, with all accrued interest and fees and the prepayment penalty, if applicable. Payoff figures are usually provided to a closing company as correct on a given day. If closing is delayed, the lender has also provided a per diem charge to increase the payoff for every day of delay.

Ultra Sound Scan: Ultra Sound is cyclic sound pressure with a frequency greater than the upper limit of human hearing. Ultrasound is thus not separated from normal (audible) sound based on differences in physical properties, only the fact that humans cannot hear it. Although this limit varies from person to person, it is approximately 20 Kilohertz (20,000 hertz) in healthy, young adults. The production of ultrasound is used in many different fields, typically to penetrate a medium and measure the reflection signature of the medium, a property also used by animals such as bats for hunting.

The most well-known application of ultrasound is its use in sonography to produce pictures of fetuses in the human womb. There are a vast number of other applications as well.

II. THEORETICAL STUDY:

We can see a large number of patients waiting in queues for a long time in hospitals for treatment every day. Service sector has been developing day by day to keep up with the changing world conditions. This development accompanies with planning and management problems. Methods developed for the services provided in hotels, markets, restaurants, factories and hospitals are the new topics of literature. Among these sector is the most reviewed one. There have been rapid changes in the health sector. Several studies have been carried out about hospitals.

Hospital administration striving to provide the best service to the patients with limited staff and equipment imposes some measures to increase the satisfaction and productivity by optimizing the conditions. As technology and science progress, waiting for something causes loss of time for both individuals and institutions. In health sector, patient waiting due to the density causes cost loss.

Patient waiting may also lead progressing of disease and bring social and economic burdens. To minimize this, various measures such as increasing the system working tie or the number of doctors in the system should be taken. Simulation needs data. Collecting data is the key process of simulation. These data cannot be obtained from health units in hospitals. The data used in simulations is non-collectible but available. There are several factors affecting the waiting time in the department. These are insufficient number of junior doctors and working time or greater number of patient admitted to hospital.

III. METHODOLOGY:

To improve resource utilization and to reduce patient waiting time of general hospitals by modifying appointment system, planning the time schedule, and staff assignment. Reduce patient waiting time via appointment scheduling and by open access policy.

We can reduce the patient's waiting time by knowing the service time of each patient as waiting time for a patient is nothing but service times of previous patients. Service time is calculated by using 2 methods, based on two assumptions.

First the service time does not depend on the time or day or the length of waiting queue. Second, a server immediately serves the next patient when its queue is not empty.

3.1 Busy period method:

It is designed for busy periods. We assume that when any patient is waiting in a queue, the server takes no ideal time and immediately serves the first patient in the queue.

$T^s_{t+1} = T^s_t + S^s_t + I^s_t$ ---------- (1)

$T^s_{t+1} - T^s_t = S^s_t + I^s_t$ ---------- (2)

$T^s_{t+1} - T^s_t = S^s_t$ (when $I^s_t = 0$) --- (3)

T^s_t – In time of a patient "t" at server "S"

S^s_t – Service time for patient "t"

I^s_t – Idle time of the server.

Here the idle time of the server is 0 because in busy period method the server mostly will not be in idle time. When identifying the busy period, we use the patient inter-department time (or patient inter-arrival time).

In a busy period, there always are patients waiting for services such that the patient inter departure time will have very little, if any, idle time in it.

3.2 IDLE PERIOD METHOD:

When server operation policy is complex and/or the patient arrival is sparse, the busy period method cannot be used to compute service time distribution. In such cases idle period method is used

We need to trace each movement and calculate the service time by using the timestamps of the patient generated by different servers. In comparison, The busy period method uses multiple patient's time stamps belonging to the target server. When the number of patients being served is small, the availability of servers will be high and patients can move through the series of services without waiting.

In such an idle period, the waiting time can be ignored.

$T^s_t = D^{s-1}_t + I^s_t$ ---------- (4)

$T^s_t = D^{s-1}_t$ (when $I^s_t = 0$) ----- (5)

$S^s_t = D^s_t - T^s_t$ ---------- (6)

$= T^s_{t+1} - T^s_t$

$= D^s_t - D^{s-1}_t$

T^s_t – In time of a patient "t" at server "S"

S^s_t – Service time for patient "t"

I^s_t – Idle time of the server.

D^s_t – Out time stamp.

IV. LIMITATIONS:

The proposed paper is only considering the mid-size hospitals. All the problems faced by the patients are not solved in this proposed paper. Only some of the problems are solved such as reduction of patients waiting time and increasing the number of departments.

A short coming of the idle period method is that the number of sample size may result in less accurate estimation of the service time distribution. Thus patients waiting time is calculated and in order to reduce it we need give appointments with respect to certain times that can be allocated for certain patients.

.

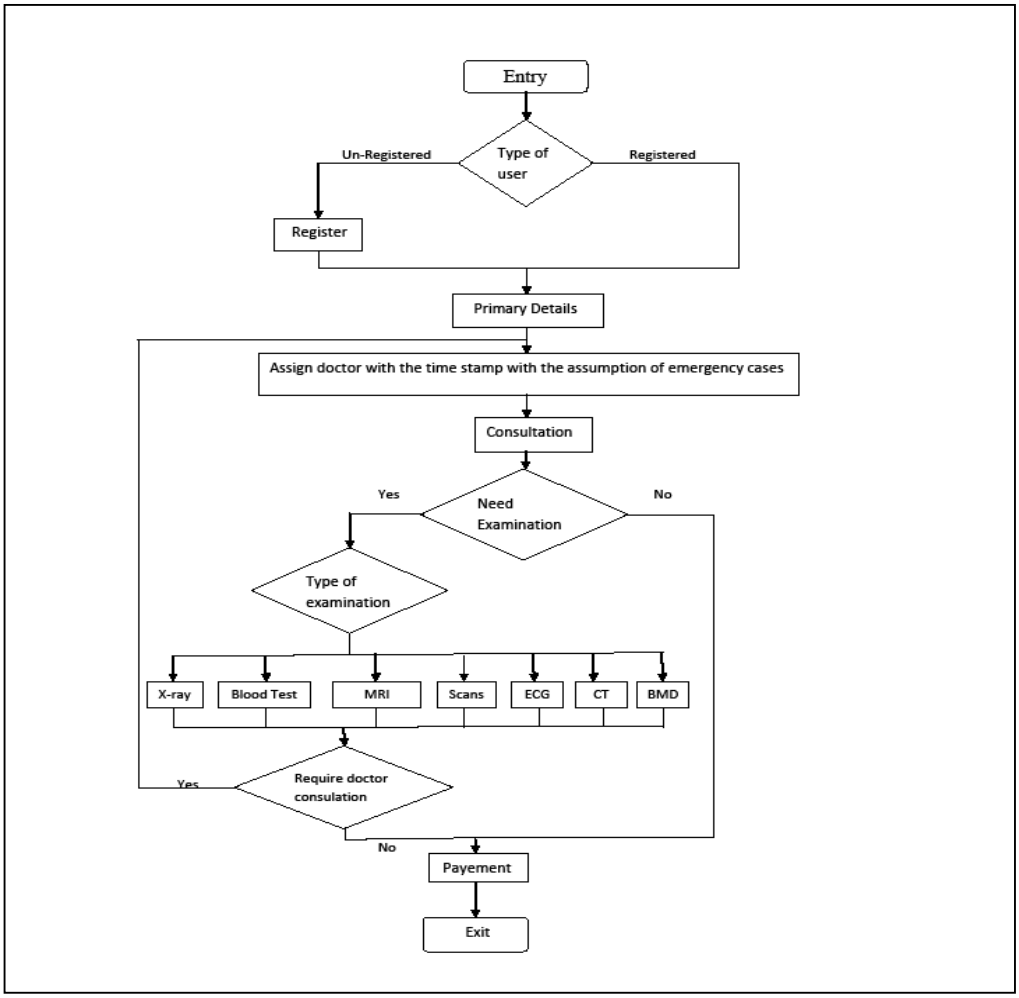


Figure 1: Patient Flow in the Hospital

. A patient will enter into the hospital and get register first if he/she is not yet registered, then they will give some primary details about the disease basing on such details we (the hospital staff) have to assign a time slot for the patient. Allotting the slots is the major thing that we need to consider to reduce the patient's waiting time here we have to assume some emergency cases before allocating the slot. It should be make sure that even for the first patient the time should not be 0 (assuming that there will be emergency cases always) and for second (next) patient to go to the same doctor the emergency case time and first patient's time will be added, after that the patient will be send for the consultation if doctor advices for any examinations like X-Ray, Blood test, ECG, Scanning (MRI, CT etc) then patient has to go for the test and if the results demands to go for doctor's consultation again patient has to consult doctor and take correct prescription then go to payment counter after payment patient will exit.

V. RESULTS

5.1 SERVICE TIME DISTRIBUTION:

We need to compute the service time of each process. The service time distribution is derived from the three methods described as busy period method Busy Period Method, Idle Period Method and Emergency Period Method. The estimated service times were presented in Table.

Process	Mean	Variance	Standard Deviation	Methode Use
Check-in	128	20417	142	Busy
Consultation	324	48254	219	Busy
X-Ray	316	7084	84	Busy
Ultrasound	998	86600	294	Idle
CT	2011	43277	208	Idle
MRI	5012	859757	927	Idle
BMD	1517	157010	396	Idle
Payment	200	3056	55	Busy
Lab Test	1215	367348	607	Idle

Table 1. Service Time Distribution (in sec.)

5.2 TRANSITION RATES:

We also need to figure out how each of our patients moves inside the hospital. Using the timestamp created either at the beginning or at the end of each process, we compute the transition rates between processes

.

From- To|	Check-in	Consultation	X-Ray	Ultrasound	CT	MRI	BMD	Payment	PT	Lab Test	Exit
Check-in	0	80.60	0.15	0.02	0	0.001	0	32.52	0.79	0.5	2.3
Consultation	3.5	0	52	0.77	0.56	0.85	0.99	87	5.76	3.38	1.7
X-Ray	2.3	67	0	0.13	0.28	0.67	0.12	3.56	45	0.45	0.33
Ultrasound	3.01	16.17	12.41	0	0.13	0.92	1.63	16.98	0.54	23.67	27.44
CT	2.34	65.27	39.21	0.45	0	0.11	5.26	9.09	0.98	1.89	9.32
MRI	12.23	75	6.23	0	0.23	0	1.35	6.23	0.47	6.7	9.5
BMD	0.1	17	79	1.20	0.005	9.45	0	13.55	0.23	4.56	9.5
Payment	0.23	0.76	0.85	0.01	0.97	0.28	0.9	0	48.14	5.51	45.51
PT	7.687	8.53	0.05	0.13	0	0.13	0.005	5.07	0	0.83	84.8
Lab Test	5.9	9.45	10.8	3.56	0.57	5.89	8.65	9.47	8.9	0	57.89
Pharmacy	0.12	69	2.09	0.36	0.5	3.46	2.32	8.64	4	8.1	0.56

Table 2. Transition Rates (in %)

.

VI. CONCLUSION

An effective method of computing the service time distribution from incomplete timestamp information. Proposed method for a case study of a mid-size hospital. The proposed model also consists of adding number of departments for reducing the patients waiting time. The departments added here will be considering the emergency cases of those doctors every time and give the timestamps to the patients in an variant of range of time. So that the wastage of the time of the patient was reduced to maximum extent. In the same way the

reduction of some viral and bacterial diseses in the hospital arena was reduced such that the number of patients waiting in the hospital and the time a particular patient staying in the hospital was reduced very much.

REFERENCES

[1] Yanbing Ju, Aihua Wang, Fengchun Zhu, Analysis of One Hospital Using Simulation, IEEE, 2006.

[2] Vladimir Boginski, In K. Mun, Yuzhou Wu, Katherine P. Mason and Chuck Zhang, Simulation and Analysis of Hospital Operations and Resource, Utilization Using RFID Data, IEEE International Conference on RFID, Gaylord Texan Resort, Grapevine, TX, USA, March 26-28, 2007.

[3] Ren Dawei, Liu Zhaoxi, Zhao Shangwu, Process Analysis of Hospital Outpatient Service Based on Arena, IEEE, 2009.

[4] Semin Sim, Sanju Park, Seogmoon Kim, SeungJae Han, Use of Incomplete Timestamp Records for Hospital Simulation Analysis, IEEE, 2009.

❑❑❑

Performance Evaluation of TCP Improvements Over MANETs

D Rajesh Reddy[1], G Swetha[2] & M Aarathi[3]

[1]ADRIN, Dept. of Space, Secunderabad, India
[2]Dept of Information Technology, Institute of Aeronautical Engg, Hyderabad, India
[3]School of Information Technology, JNTU Hyderabad , Hyderabad, India
E-mail : datlarajeshreddy@gmail.com, swetha9g@gmail.com & m_arathi@gmail.com

Abstract - TCP does not work well in MANETs as it assumes that the packet loss is due to network congestion even for link breakages due to node mobility, which inturn degrades the TCP performance. In this paper, we have evaluated performance of the various TCP improvements such as TCP-F, TCP-ELFN, TCP-BuS, Split-TCP and Hop-by-Hop TCP over MANETs based on three performance metrics (Packet Delivery Ratio, Average End-to-End-Delay and Throughput) using GloMoSim Simulator. The simulation results shown that TCP-Bus outperforms over other TCP improvements where as split approaches focused on reliability.

Keywords*-Average End-to-End delay; Mobile Ad-hoc Networks; TCP; Throughput; Packet Delivery Ratio*

I. INTRODUCTION

A Mobile Ad hoc Network (MANET) is a self-configuring network that is formed and deformed on the fly by a collection of mobile nodes without the help of any prior infra-structure or centralized management. These networks are characterized as infrastructure less, mobile, autonomous, multi-hopped, self-administered, and having dynamic topology. Nodes within each other's radio range communicate directly via wireless links, while those that are far apart use other nodes as relays in a multi-hop routing fashion. As mobile ad hoc networks are characterized by a multi-hop network topology that can change frequently due to mobility.

TCP as the standard transmission protocol for most of the networking applications for the reliable packets delivery and support the mechanisms of flow and congestion control. TCP assumes that the packet loss is due to network congestion. It then invokes appropriate congestion control actions including window size reduction. Although this assumption is reasonable for wired networks, but for wireless networks especially MANETs the possible causes of packet losses include wireless link errors, MAC layer losses due to channel contention, and link breakages due to node mobility. All these causes that are not related to congestion unnecessarily trigger the congestion control mechanism, which will degrade the TCP performance [1].

A. TCP improvements over MANETs:

The following are some of the solutions to improve the performance of TCP in MANETs:

TCP-F [7]: TCP Feed-back relies on the network layer at intermediate nodes to detect the route failures due to the mobility. TCP-F puts the TCP sender in one of the two states: active state, snooze state. In the active state, TCP sender follows the standard TCP behavior. As soon as an intermediate node detects a link failure, it explicitly sends a route failure notification (RFN) packet to the sender and records this event. After receiving the RFN, the sender enters the snooze state. In snooze sate the sender stops sending further packets and freezes the values of state variables such as retransmission timer and congestion window size. The sender remains in the snooze state until the intermediate node notifies it of the restoration of the route through a route reestablishment notification (RRN) packet. Then it enters the active state again.

TCP-BuS [2]: TCP-BuS is similar to TCP-F in detection mechanisms. Two control messages (ERDN and ERSN) related to route maintenance are introduced to notify the TCP sender of route failures and route reestablishment. These indicators are used to differentiate between network congestion and route failures as a result of node movement. ERDN (Explicit Route Disconnection Notification) message is generated

at an intermediate node upon detection of a route disconnection, and is propagated toward the sender. After receiving an ERDN message, the sender stops transmission. Similarly, after discovering a new partial path from the failed node to the destination, the failed node returns an ERSN (Explicit Route Successful Notification) message back to the sender. On receiving ERSN message, the sender resumes data transmission. TCP-BuS considers the problem of reliable transmission of control messages. If a node A

reliably sends an ERDN message to its upstream node B, the ERDN message subsequently forwarded by node B can be overheard by A (assuming same transmission ranges of A and B). Thus, if a node has sent an ERDN message but cannot overhear any ERDN message relayed by its upstream node during a certain period, it concludes the ERDN message is lost and retransmits it. The reliable transmission of ERSN is similar.

ELFN [4]: Explicit Link Failure Notification is another technique based on feedback. The objective is to provide the TCP sender with information about link and route failures so that it can avoid taking congestion control actions. ELFN is based upon DSR (Dynamic Source routing) routing protocol. To implement ELFN message, the route error message of DSR was modified to carry a payload similar to the "host unreachable" ICMP message. When a TCP sender receives an ELFN, it disables its retransmission timers and enters a "stand-by" mode, which is similar to the snooze state of TCP-F. Instead of using an explicit notice to signal that a route has been reestablished, a packet is sent periodically to probe the network to see if a route has been established. After finding a new route, the sender leaves the stand-by mode, restores its retransmission timers and continues as normal.

Split-TCP [9]: Proxies split a TCP connection into multiple local segments. They buffer packets and deliver them to the next proxy or to the destination. Each proxy receives packets from either the source or from the previous proxy, sends LACK (Local ACK) s for each packet to the sender (source or proxy) of that packet. These proxies buffer the packet and forward the packet towards the destination, at a rate proportional to the rate of arrival of LACKs. The source keeps transmitting according to the rate of arrival of LACKs from the next proxy, but removes the packet from its buffer only when the end-to-end ACK for that packet received from the destination.

Hop-by-Hop TCP [10]: Hop-by-Hop TCP consists of two parts: an End-to-End TCP working on the source and destination nodes, and a One-Hop TCP working on every node. The sender module of a One-Hop TCP is working at the sender end of a link, and the receiver module is working at the receiver end. Each link needs only one pair of One-Hop TCP for all End-to-End TCP sessions. One-Hop TCP is a light-weight version of TCP running on each node. It forwards the received packets to the next node and sends the Local ACK to the previous node. One-Hop TCP adds the IP address of current node to the packet header such that the receiver knows where to send Local ACK, sets CWND to 1 and removes all CWND adjustment mechanism.

II. PERFORMANCE METRICS

We have considered three performance metrics, the packet delivery fraction, average end-to-end delay and throughput. All metrics are measured quantitatively. Following is description of each metrics.

Packet Delivery Ratio(PDR): It is ratio of successfully delivered data packets to packets generated by CBR sources. Packets delivery ratio describes how successfully protocol delivers packet from source to destination.

PDR = (ΣCBR pkt rec / ΣCBR pkt sent) * 100

Average end-to-end delay: This performance metric defines all possible delays. There are many factors causing delay in network, such as, queuing delay, buffering during routes discovery, latency and retransmission delay. Lower delay means better performance.

$$\text{Average End-to-End delay} = \frac{1}{N}\sum_{n=1}^{N}(Rn - Sn)$$

Sn = Time, when data packet *n* was sent

Rn = Time, when data packet *n* was received

N = Total number of data packets received

Throughput: It is ratio of the total number of bits received to the total time taken.

Throughput= Total number of bits received / Total time taken

Where total time taken = time at which last packet received – time at which first packet sent.

All performance metrics are analyzed under varying network parameters such as pause time, varying speed.

TABLE-1 : SIMULATION PARAMETERS

Parameters	Environment
Area	1000m*1000 m
Simulation time	900 sec
Transmission range of a node	35.293 m
Transmission Power	-10.0 dBm
Traffic Source	CBR

Temperature	290.0 K
Node density	50
Node Placement	Random
Mobility model	Random Waypoint mobility model

III. SIMULATIONS AND RESULTS

A. Simulation Setup

We have implemented five TCP improvements TCP-F, TCP-ELFN, TCP-BuS, Split-TCP and Hop-by-Hop TCP in GloMoSim [5]. All the simulations are done in GLOMOSIM-2.03 and the simulation parameters are described in Table I. DSR is used as a routing protocol. In the simulation environment, we transmitted the constant bit rate (CBR) traffic source with 2048-byte data packets at one second intervals.

B. Results

Pause-time model: This test studied the effects of increasing pause time on the performance of TCP improvements. As pause time increases, mobility (movement) decreases. When a pause time occurs, node stops for a while and selects another direction to travel. If speed is defined as constant then for every occurrence of pause time, speed of node remains constant. In this model pause time changes from 0s to 250s while other parameters (nodes = 50, speed = 25 m/s, data sending rate = 2Mbps and no. CBR flows = 10) are constant.

Figure 1 shows throughputs for all the TCP improvements with respect to pause time. TCP-BuS outperforms over all TCP improvements. Split-TCP and Hop-by-Hop TCP does not forward subsequent packet until LACK is received by the corresponding proxy node and next Hop respectively for the transmitted packet. Cwnd is always set to 1 in Hop-by-Hop TCP. Hence it has relatively low Throughput among all TCP improvements.

Figure 2 shows Avg. End-to-End delay for all the TCP improvements with respect to pause time. TCP-BuS outperforms over other TCP improvements. Since the intermediate node in both Split-TCP and Hop-by-

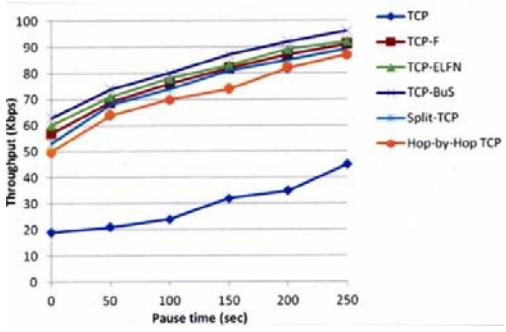


Figure 1.Pause time Vs Throughput

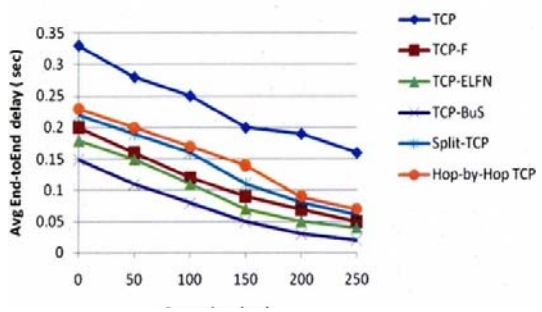


Figure 2.Pause time Vs Average End-to-End delay

Hop TCP also access the TCP header information from the packet apart from Sender and Receiver, the Average End-to-End Delay is more compared to TCP-F, TCP-ELFN and TCP-BuS.

Speed Model: Speed of nodes play an important role in Mobile Ad Hoc Networks. In this model the node's speed changes from 10 m/s to 30 m/s with 0sec pause time. Others parameters like sending rate 2Mbps, no. of nodes 50, and CBR flows 10 are kept constant.

Figure 3 shows difference in Throughput for all the TCP improvements with respect to node speed. As speed increases, number of route failures will be more. TCP-BuS buffers the packet at intermediate node which can be forwarded immediately by the intermediate node itself as soon as route is re-established. Wherein other TCP improvements, the sender has to forward the packet after route re-establishment. TCP-BuS has high throughput compared to other TCP improvements.

Figure 4 shows difference in Average End-to-End delay for all the TCP improvements with respect to

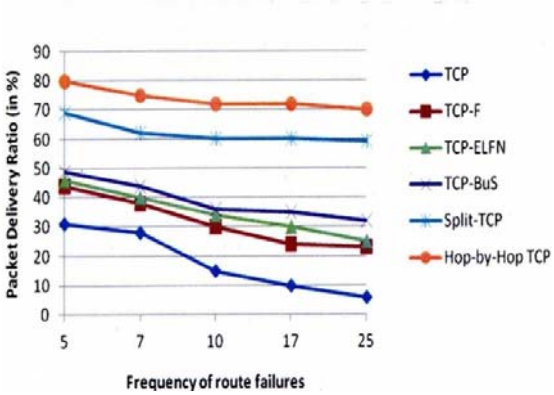


Figure 3.Speed Vs Throughput

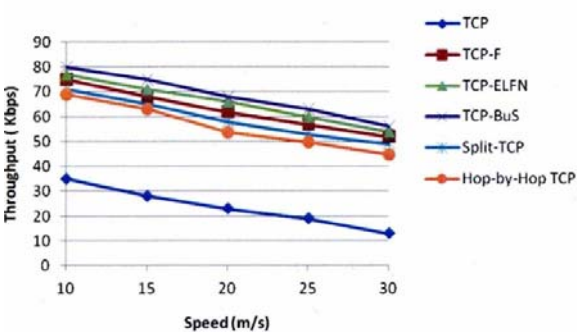


Figure 4. Speed Vs Average end-to-End delay

node speed. It shows that the average end-to-end delay increases as speed increases. Since each Hop in Hop-by-Hop TCP and each proxy node in Split-TCP uses TCP functionality and awaits LACK for subsequent packet transmissions, these improvements take relatively more time for packet transmission.

Frequency of route failure model: Figure 5 emphasizes on the impact of route failures in the performance of TCP improvements over MANETs. As frequency of route failures increases, rapid changes in network topology would be anticipated. Because the subsequent packet transmission relay on previous LACK, Hop-by-Hop TCP is more reliable followed by Split-TCP. So Packet Delivery Ratio for Hop-by-Hop is high.

IV. CONCLUSION

TCP improvements are compared in terms of, Throughput and Average end-to-end delay with respect to Pause time and node speed and also compared Packet Delivery Ratio with respect to frequency of route failures.

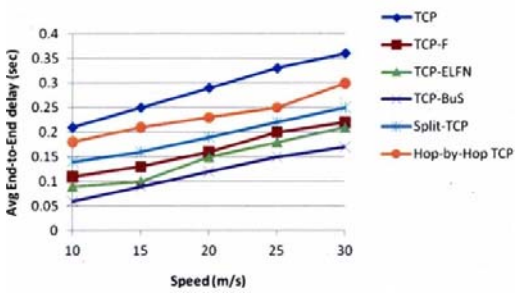


Figure 5. Frequency of route failure Vs Packet Delivery Ratio

It is observed from the simulation that TCP-BuS is giving high throughput and low end-to-end delay compare to other TCP improvements. Split approaches showing high Packet Delivery Ratio compare to other TCP improvements. From this, we are concluding that end-to-end approaches exhibit high performance where as Split approaches are highly reliable.

VI. REFERENCES

[1] C. Siva Ram Murthy, B.S. Manoj "Ad hoc wireless networks-Architectures and Protocols", Pearson Education, 2004.

[2] D. Kim, C.-K. Toh, and Y. Choi,"TCP-BuS: Improving TCP Performance in Wireless Ad Hoc Net-works" Journal of Communications and Networks, Vol. 3, No. 2. Jun. 2001.

[3] Dr. Panos Bakalis, Bello Lawal," Performance Evaluation of CBR and TCP Traffic Models on MANETusing DSR Routing Protocol", International Conference on Communications and Mobile Computing, 2010.

[4] G. Holland and N. Vaidya, "Analysis of TCP performance over mobile ad hoc networks" In Proceedings of ACM MobiCom'99, Seattle, Washington, Aug. 1999.

[5] Glomosim-2.03 software, Available: "http://pcl.cs. ucla.edu/projects/glomosim/academic/download. html".

[6] H. M. El-Sayed, "Performance evaluation of TCP in Mobile ad-hoc networks",The Second International Conference on Innovations in Information Technology,2005.

[7] K. Chandran, S. Raghunathan, S. Venkatesan, and R. Prakash, "A feedback-based scheme for improv-ing TCP performance in ad hoc wireless networks", IEEE Personal Communications Magazine, 8(1):34{39, Feb. 2001}.

[8] K. Natarajan, Dr. G. Mahadevan "A Comparative Analysis and Performance Evaluation of TCP over MANET Routing Protocols", Int Jr of Advanced Computer Engineering and Architecture Vol. 1, No. 1, June 2011

[9] S. Kopparty, S. Krishnamurthy, M. Faloutous, and S.Tripathi, "Split TCP for mobile ad hoc networks", Proc. of IEEE GLOBECOM, Nov. 2002.

[10] Yao-Nan Lien and Yi-Fan Yu, "Hop-by-Hop TCP over MANET", Asia-Pacific Services Comuting Conference, 2008, APSCC'08.IEEE. Dec. 2008.

An Analytical Model on Wireless Sensor Networks

[1]Snehansu Bank, [2]Subrata Saha & [3]Indrajeet Banerjee

[1]PDSIT, Bengal Engineering & Science University, Shibpur, Howrah
[2]Institute of Engineering & management, Salt Lake, Kolkata,
[3]IT Department, Bengal Engineering & Science University, Shibpur, Howrah

Abstract - In this paper, we focus on optimization problem using minimum number of sensor nodes for achieving maximum coverage and connectivity. Efficient mathematical model is proposed to obtain solution of the problem exactly or in an approximate manner. Here we have used weibull distribution in its modified form. We construct the model in probabilistic environment. Model is constructed under true multi objective frame work to (1) reduce number of active nodes and (2) maximize the coverage area. We conduct extensive numerical verification to obtain empirical formulae based on the results.

Key words: *Wireless Sensor Network, Multi objective optimization, Weibull distribution.*

I. INTRODUCTION

Sensor networks have a long history, which can be initially traced back as far as the 1950's. It is recognized that the first obvious sensor network is the Sound Surveillance System (SOSUS). With the emergence of integrated sensors embedded with wireless capability, most of current sensor networks compose a collection of wirelessly-interconnected sensors, of which each is embedded with sensing, computing and communication components. These types of sensor networks are referred to as wireless sensor networks. In design and implementation of WSNs, many system parameters are needed to be optimized. This problem is equivalent to minimize the total number of active nodes, subject to ordinary operations of the system. In this paper, we consider a problem to find such a minimum set of active sensors and their positions, such that proper routing operations can still be facilitated. Potential applications of WSNs include, but are not limited to, health applications [1], agricultural monitoring [2], environment monitoring [3], smart home applications [4]. For a more comprehensive list of WSN applications, see Sohraby [5]. WSNs were studied from the application areas to the system architectures.

A WSN or a wireless ad hoc network is often represented by a graph, in which vertices correspond to the communication nodes in the wireless network, and an undirected edge between a pair of vertices indicates that the corresponding nodes can communicate directly. Edges of the graph are undirected because all the nodes in the WSN are assumed to have homogeneous transmission range. With this graphical representation, a network is called connected if its associated graph is connected. A connected network implies that any pair of nodes can communicate with each other, possibly through multiple hops by using relay nodes. In general, there are two types of approaches to deploy sensors in a WSN, the deterministic deployment, i.e., sensors are placed exactly on pre-engineered positions, and the random deployments, i.e., sensors are deployed on random positions. Although in order to reduce installation costs, there are requirements for applying random deployment to deploy large WSNs which contain very large numbers of nodes. Communication networks are often modeled as random graphs. Mathematically, a random graph is a graph that is generated by some stochastic processes. It was originally derived as a tool to prove in the combinatorial mathematics. However, it is now widely accepted as a modeling method in communications. One type of the most extensively studied random graphs are Erdoős-Re'nyi[6] graphs. An Erdoős-Re'nyi graph G(n, p) is a graph with n vertices, and each of the ${}_nC_2$ possible edges exists independently with a probability p. As Erdoős-Re'nyi graphs assign randomness to edges, i.e., there is a link between each pair of nodes with certain probability, it is not an ideal model to describe real networks with geometric properties, in which distance between nodes is an important parameter to affect the existence of communication links. Moreover, as discussed by Chlamtac and Farago [7], Erdoős-Re'nyi graphs do not consider correlations between different links, mainly due to their incapability of describing the distance-dependent characteristics. A spontaneous candidate for random network modeling is the so called Random Geometric Graphs. With node set V, a geometric graph G = (V, r) is equivalent to a graph G1 = (V, E), in which V is the vertices set, and

$$E = \{(u, v) \mid u, v \in V, \|u - v\| \le r\}$$

To ensure that the WSN can operate properly, sensor placement scheme needs to be carefully determined. There are two types of sensor deployment methodologies, random deployment and deterministic deployment, e.g., sensors are thrown in mass on random positions and placed exactly on carefully engineered positions. For instance, to apply the random placement, we can drop all the sensors from a vehicle in batch, thereby reducing the labour cost of the deployment.

However, this approach brings randomness in the positions of sensors, and the randomness may degrade the network performance. Most of existing literature in the field focuses on determining the "optimal" deterministic placement pattern. The optimality is defined on different context according to the applications and goals of the WSNs. A natural objective for the optimal sensor placement is to minimize the number of required sensors needed, subject to the constraint that the whole sensing field is monitored by the deployed sensors. It is equivalent to finding the minimum number of nodes, such that every position in the sensing field is within the sensing coverage by at least one node. Zhang and Hou[8] proved that arranging sensors at the centers of regular hexagons is optimal for a WSN with a large sensor field, given that all the sensor nodes have identical limited sensing range. To find the optimal node placement pattern subject to the coverage and the connectivity constraints, modeled by Biagioni and Sasaki [9] the problem as an optimization problem that minimizes the number of required sensor nodes, under the constraints that any position of the sensor field is under surveillance of at least one node and all nodes are connected. However, the authors did not solve the problem optimally. They only studied a wide range of regular sensor network deployment patterns. These patterns include circular, and star topologies, as well as grid topologies, such as triangular, square and hexagonal grids. Iyengar et al [10] proved that a strip-based pattern is nearly optimal for large networks in two dimensional space. This paper investigates optimization problems for design and planning of Wireless Sensor Networks (WSNs). The goal of the paper is to find a minimum relay set in a random Wireless Sensor Network, which is equivalent to finding a Minimum Connected Dominating Set (MCDS) corresponding to the network topology. This problem is of interest since MCDS plays an important role in achieving efficient broadcasting, and energy-conserving routing. The second problem focuses on computing the minimum number of sensors to achieve a certain level of connectivity in a random WSN. This problem is called the connectivity problem, which is also important as full connectivity is critical for a WSN to successfully operate. In addition, we evaluate the probability of sensor network connectivity with a given number of sensor nodes, and estimate the tolerable node position perturbation to maintain a required level of connectivity. **Mathematical Model:** Assume the network is an m × n sensor field and k sensors are deployed in the random deployment stage. Each sensor has a detection range r, and sensor Si is deployed at point (X_i, Y_j). For any point at (x,y), we denote the Euclidian distance between Si and P as $d(S_i,P)$

$$d(Si, P) = (x_i - x)^2 + (y_i - y)^2 \quad \text{-----(i)}$$

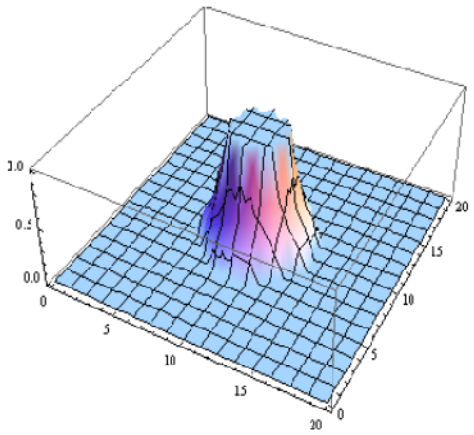
In the binary sensor model coverage function $C_{xy}(S_i)$ of a grid point P by sensor S_i can be expressed as

$$C_{xy} = \begin{cases} 1 & \text{if} \quad d(S_i, P) \le r \\ 0 & \text{otherwise if} \end{cases} \quad ---(ii)$$

The coverage C_{xy} (S_i) needs to be expressed in probabilistic terms; hence a precise detection model is introduced.

$$C_{xy}(s_i) = \begin{cases} 0 & \text{if} \quad r + r_e \le d(S_i, P) \\ Ae^{-\lambda \alpha^{\beta}} & \text{if} \quad r - r_e < d(S_i, P) \le r + r_e \\ 1 & \text{if} \quad r - r_e \ge d(S_i, P) \end{cases} \quad -(iii)$$

Where = d(Si , P) – (r - re). Where re is the measure of uncertainty in sensor detection. , are the parameter of sensor detection probability when a target is at distance greater than re but within maximum from the sensor. Different values of the parameters & yields different translations reflected by different detection probability. The graphical representation of the function is given below:



This can be viewed as the characteristics of various types of physical sensor. If we consider a

grid point with co-ordinate (x,y) lying in the overlap region of sensor S_i & S_j , then the sensors within a cluster operate independently in their sensing activities, if neither S_i nor S_j covers grid point at (x,y) . The probability that the grid point (x,y) will be covered is $C_{xy}(S_i,S_j) = 1- (1- C_{xy}(S_i)) (1- C_{xy}(S_j))$. If C_t be the desired coverage threshold for all grid point then

$$\min_{x,y} \{C_{xy}(S_i,S_j)\} \geq C_t$$

The total coverage will be obtained as,

$$TC=1-\prod_{S_i \in \Lambda}(1-C_{xy}(S_i,S_j))$$

$$\Lambda = \{S_1,S_2,\ldots,S_k\}$$

Now the coverage rate will be

$$TCR = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{C_{xy}}{m \times n}$$

Then our problem as for $\Lambda = \{S_1,S_2,\ldots,S_k\}$

$$Z = \arg\max \{f_1(z), (1-f_2(z))\}$$

$$\text{Where } Z = \{(x_1,y_1),(x_2,y_2),\ldots(x_m,y_n)\} \in Z$$

$$Y = (y_1,y_2\} \in Y$$

Z is called the decision space. Y is called the objective space.

Where $f_1(z)$ is the coverage rate of z and $f_2(z)$ is the sensor used rate.

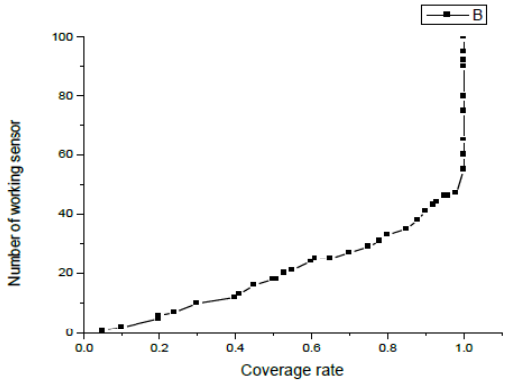
$$f_1(z) = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{C_{xy}(z)}{m \times n} \quad \text{and} \quad f_2(z) = \frac{|Z|}{|K|}$$

Since the problem is converted into a multi objective optimization problem, to find compromise solution, here we use the concept of Nash equilibrium. First we determine optimal solution of each objective then we optimize utility function defined by Minimize

$$U = (f_1^* - f_1)^{\alpha}.(f_2 - f_2^*)^{1-\alpha}$$

Where f1* is the optimal solution of f1 when we do not consider f2. Similarly f2* is the optimal solution when we do not consider f1 **Numerical Results**

In this section, we present some experimental results. Throughout these experiments, we set the communication radius as twice the sensing radius to ensure coverage connectivity. In the first experiment, we place 100 potential sensor nodes in a 100 X 100 square field. The cost associated with each allocation is 1000 units. Among them, 39 sensors are distributed in a hexagonal structure to cover the whole area, which is the optimal sensor distribution. The other 61 sensors are randomly deployed in the area, so these nodes' x and y-coordinates are random. The locations of the sensor nodes are selected as the working set. The recombination rate and mutation rate are set as 0.9 and 0.01 respectively. Back's stated that the optimal value of mutation rate is about the inverse of the chromosome length. Better solutions are obtained in subsequent generations shown in Fig. For instance, the 120th generation improves the coverage rate with a small compromise in the average number of working nodes.



II. CONCLUSION

In this paper, a mathematical model is proposed for wireless sensor networks to optimize coverage rate of the network and the number of active nodes from an existing facility. The model is developed under probabilistic environment by using Weibull distribution detection function. The model is constructed as a Multi-objective optimization problem. We use genetic algorithm to determine the solution. For future extension, the constraint of location may be incorporated to the model.

REFERENCES

[1] G. Amato, S. Chessa, F. Conforti, A. Macerata, and C. Marchesi. Health Care Monitoring of Mobile Patients. ERCIM News, (60):69–70, Jan. 2005.

[2] J. Burrell, T. Brooke, and R. Beckwith. Wineyard Computing: Sensor Networks in Agricultural Production. IEEE Pervasive Computing, 3(1):38–45, 2004.

[3] R. Cardell-Oliver, K. Smettem, M. Kranz, and K. Mayer. A Reactive Soil Moisture Sensor Network: Design and Field Evaluation. International Journal of Distributed Sensor Networks, 1(2):149–162, 2005.

[4] N. Xu. A Survey of Sensor Network Applications. http://courses.cs.tamu.edu/rabi/cpsc689/resources/sensor%20nw-survey.pdf. Accessed in Feb, 2008.

[5] K. Sohraby, D. Minoli, and T. Znati. Wireless Sensor Networks: Technology, Protocols and Applications. John Wiley & Sons, Inc., 2007.

[6] P. Erdoős and A Re'nyi . On the Evolution of Random Graphs. Publ. Math. Inst. Hungarian Acad. Sci., 5:17–61, 1960.

[7] I. Chlamtac and A. Farago'. A New Approach to the Design and Analysis of peer-to-peer Mobile Networks. ACM Wireless Networks, 5(3):149–156, May 1999.

[8] H. Zhang and J. C. Hou. Maintaining Sensing Coverage and Connectivity in Large Sensor Networks. Ad Hoc and Sensor Wireless Networks, 1:89–124, March 2005.

[9] E. S. Biagioni and G. Sasaki. Wireless Sensor Placement for Reliable and Efficient Data Collection. In Proc. 36th Annual Hawaii International Conference on System Sciences, Jan. 2003.

[10] R. Iyengar, K. Kar, and S. Banerjee. Low-coordination Topologies for Redundancy in Sensor Networks. In Proc. 6th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc'05), pages 332–342, USA, May 2005.

❑❑❑

Data Mining for IDS

Seyed Hasan Mortazavi Zarch
E-mail : hassanmortazaviy@yahoo.com

Abstract - There has been much interest in applying data mining to computer network intrusion detection system. This paper provides lessons learned in this task. Based upon our experiences in getting started on this type of project, we suggest data mining techniques to consider and types of expertise and infrastructure needed. This paper has two intended audiences: network security professionals with little background in data mining, and data mining experts with little background in network intrusion detection system.

I. Intrusion Detection System(IDS):

- An intrusion detection system is used to monitor network traffic, check for suspicious activities and notifies the network administrator or the system. In some instances, the IDS might also react to malicious or anomalous traffic and will take action such as barring the user or perhaps the IP address source from accessing the system.
- IDS are available in many different types and will approach the mission of uncovering shady traffic in various ways. You can find host-based (HIDS) and network-based (NIDS) systems. Additionally, there are also IDS which detect movements by searching for particular signatures of well-known threats, just like the way how antivirus software generally detects and safeguards against malware and also IDS which detect by assessing traffic patterns against the baseline and look for anomalies. Besides that, there are that basically observe and alert, plus systems that carry out an action or even actions in reaction to a recognized threat. The following will take a brief look on each intrusion detection system:
- NIDS These are installed at a tactical point or maybe points inside the network in order to monitor all traffic on the network. In reality you would check out all incoming and outgoing traffic, but doing this could produce a bottleneck which would damage the all-round speed of your computer network.
- Signature Based This can be used to monitor the packets on the system and then do a comparison against the database of attributes or signatures from recognized malicious threats. It is similar to how most anti-virus software would detect malware. However, there is a downside with this system because there will be a lag in between when new threats are identified in the wild and also the signature for finding that threat being used on your IDS. In that lag period the IDS will be unable to identify any new threat.
- HIDS These operate on individual devices or hosts on the system. This will monitor all the incoming and outgoing packets on the device only and can notify the administrator or user of any suspicious activity.

II. IDS before Data Mining

When we first began to do intrusion detection on our network, we didn't focus on data mining, but rather on more fundamental issues: How would the sensors perform? How much data would we get? How would we display the data? What kind of data did we want to see, and what queries would be best to highlight that data? Next, as the data came in, sensor tuning, incident investigation, and system performance commanded our attention. The analyst team grew to handle the load, and training and team coordination were the issues of the day. But the level of reconnaissance and attack on the internet was constantly increasing, along with the amount of data we were collecting and putting in front of our analysts. We began to suspect that our system was inadequate for detecting the most dangerous attacks—those performed by adversaries using attacks that are new, stealthy, or both. So we considered data mining with two questions in mind:

Σ Can we develop a way to minimize what the analysts need to look at daily?

Σ Can data mining help us find attacks that the sensors and analysts did not find?

III. Data Mining

It is the crucial step in which clever techniques are applied to extract patterns potentially useful. Here are a few specific things that data mining might contribute to an intrusion detection project:

- Remove normal activity from alarm data to allow analysts to focus on real attacks
- Identify false alarm generators and "bad" sensor signatures
- Find anomalous activity that uncovers a real attack
- Identify long, ongoing patterns (different IP address, same activity)
- To accomplish these tasks, data miners use one or more of the following techniques:
- Data summarization with statistics, including finding outliers
- Visualization: presenting a graphical summary of the data
- Clustering of the data into natural categories [Manganaris et al., 2000]
- Association rule discovery: defining normal activity and enabling the discovery of anomalies [Clifton and Gengo, 2000; Barbara et al., 2001]
- Classification: predicting the category to which a particular record belongs [Lee and Stolfo, 1998]

IV. START BY MAKING YOUR REQUIREMENTS REALISTIC

The seductive vision of automation is that it can and will solve all your problems, making human involvement unnecessary. This is a mirage in intrusion detection. Human analysts will always be needed to monitor that the automated system is performing as desired, to identify new categories of attacks, and to analyze the more sophisticated attacks. In our case, our primary concern was relieving the analyst's day to day burden.

Real-time automated response is very desirable in some intrusion detection contexts. But this puts a large demand on database performance. The database must be fast enough to record alarms and produce query results simultaneously. Real time scoring of anomaly or classification models is possible, but this should not be confused with real-time model building. There is research in this area [Domingos and Hulten, 2000], but data mining is not currently capable of learning from large amounts real-time, dynamically changing data. It is better suited to batch processing of a number of collected records. Therefore, we adopted a daily processing regime, rather than an hourly or minute-by-minute scheme.

V. SELECT A BROAD AND CAPABLE PROJECT STAFF

Our staff will need skills in three areas: network security, data mining, and database application development.

- Of course the security staff need a solid grounding in networking and intrusion detection, but they also need to be able to tackle big, abstract problems.
- The data miners should have a good grounding in statistics and machine learning, but they will also need to learn detailed concepts involved in computer networking.
- Σ The database developers will need good skills in efficient database design, performance tuning, and data warehousing.

VI. INVEST IN ADEQUATE INFRASTRUCTURE

- Significant infrastructure is required to do this sort of work. In addition to the normal processing of the data from the intrusion detection system, you will need:
- **A Database:** Because you will need to store a great deal of data, update this data regularly, and obtain rapid responses to complex queries, we recommend that you select a high-end production-quality database management system.
- **Storage Space:** In addition to the handling of normal IDS data, you will need data and working space associated with data mining. Additional data includes calculating and saving metadata, as well as sometimes copying existing data into more convenient data types. Working space will hold the various sample data sets that will be extracted for

experimentation, as well as working files containing intermediate and final results. Plan for data mining to double your storage requirements.

- **Compute capability:** Data mining tools are very CPU and memory intensive. Naturally, the more memory and CPU power the better. We have found that we needed at least four times the memory and CPU power over what would be needed for an IDS database without the data mining.
- **Software:** In addition to what is required for the basic system (production quality database, Perl, database middleware, database administration and tuning aids), plan for acquisition of specialized tools.

VII. PLAN, COMPUTE, AND STORE APPROPRIATE ATTRIBUTES

- Data records consist of many attributes. When doing data mining for intrusion detection one could use data at the level of TCPDUMP [Lee and Stolfo, 1998] or at the alarm level [Manganaris, et al. 2000]. In both types of data you will find fields for source IP address, destination IP address, source port number, destination port number, date/time, transfer protocol (TCP, UDP, ICMP, etc.), and traffic duration (or equivalently, both start and end times). These 'base' attributes give a good description of the individual connection or alarm, but they often are insufficient to identify anomalous or malicious activity because they do not take into account the larger context. The individual connection records in a denial of service attack are not, by themselves, malicious, but they come in such numbers that they overwhelm your network. A single connection between an outside machine and a single port on a machine inside your network is also not malicious—unless it is part of a series of connections that attempt
- all the active ports on that machine. For this reason you will want to add additional
- fields containing values derived from the base fields.
- Another type of derived data, called an aggregation, is a summary count of traffic matching some particular pattern. For example, we might want to know, for a particular source IP address X, and a particular IP address Y, how many unique destination IP addresses were contacted in a specific time window Z. A high value of this measure could give an indication of IP mapping, which is a pre-attack reconnaissance of the network. Aggregations are generally more expensive to compute than other kinds of derived data that are based upon only a single record.

A third type of derived data is a flag indicating whether a particular alarm satisfies a heuristic rule. Because data mining methods handle many attributes well, and because we don't know for sure which one will be useful, our approach is to compute a large number of attributes (over one hundred) and store them in the database with the base alarm fields.

VIII. INSTALL DATA FILTERS

In our sensor log table, upwards of 95% of the traffic fit the profile of an IP mapping activity. That is, a single source IP was attempting a connection to hundreds or even thousands of destination IPs. Before security specialists can start providing input to the data mining effort, this traffic must be filtered. It is a straightforward task to create a filter that can find these patterns within a data table of traffic.

At MITRE, this preliminary filter is called HOMER (Heuristic for Obvious Mapping Episode Recognition). The heuristic operates on aggregations by source IP, destination port, and protocol and then check to see if a certain threshold of destination IPs were hit within a time window. If the threshold is crossed, an incident is generated and logged to the database. The reduction obtained by HOMER is significant. For example, for the period of Sep. 18 to Sep. 23, 2000, MITRE network sensors generated 4,707,323 alarms (71,094 of priority 1). After HOMER there were 2,824,559 (3,690 of priority 1) - a reduction of 40% (94% of priority 1).

IP mapping activity does not pose much of a security threat in itself, but it can be a prelude to more serious activity. Thus, HOMER provides one other important function. Even though the bulk traffic due to the mapping activity is not shown to the analyst, the source host itself is placed on the radar screen of our system. Please note that some normal activity (e.g., name servers, proxies) within an organization's intranet can match the profile of an IP mapping. HOMER handles this situation by means of an exclusion list of source IPs.

A second heuristic under development, called GHOST (Gathering Heuristic for Obvious Scanning Techniques), plays a slightly different role than HOMER. Port scanning is a more targeted form of information gathering that attempts to profile the services that are run on a potential intrusion target. The GHOST heuristic uses a different set of fields, and has its own configurable time window and port threshold, which if exceeded, triggers a security incident.

IX. limit the Overall Architecture for Intrusion Detection system

Our current architecture for intrusion detection is shown in Figure 1. Network traffic is analyzed by a variety of available sensors. This sensor data is pulled periodically to a central server for conditioning and input to a relational database. HOMER filters events from the sensor data before they are passed on to the classifier and clustering analyses. Data mining tools filter false alarms and identify anomalous behavior in the large amounts of remaining data. A web server is available as a front end to the database if needed, and analysts can launch a number of predefined queries as well as free form SQL queries from this interface. The goal of this operational model is to have all alarms reviewed by human analysts.

.

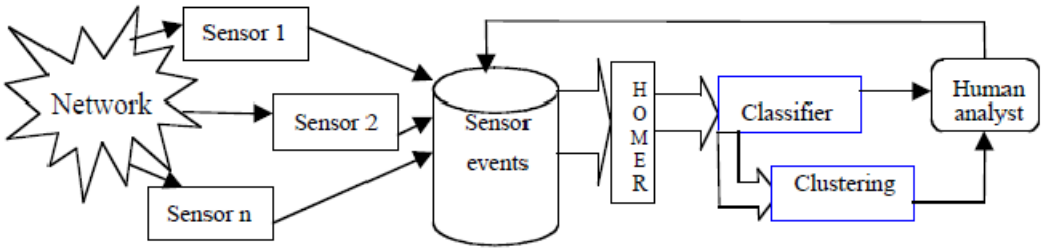


Figure 1. How sensors feed into overall intrusion detection system

.

Without automated support, this task is increasingly difficult due to the volume of alarms. In one recent day at MITRE for example, sensors generated about 3.4 million alarms, of which about 48,000 are labeled priority 1. Attacks and probes can be frequent and noisy, generating thousands of alarms in a day. This can create a burden on the network security analyst, who must perform a triage on the enormous flood of alarms.

X. BUILD CLASSIFICATION RULES

Classification is used to assign examples to pre-defined categories. Machine learning software performs this task by extracting or learning discrimination rules from examples of correctly classified data. Classification models can be built using a wide variety of algorithms. Henery [1994] classifies classification algorithms into three types:

- extensions to linear discrimination (e.g., multi-layer perceptron, logistic discrimination),
- decision tree and rule-based methods (e.g. C4.5, AQ, CART), and
- density estimators (Naïve Bayes, k-nearest neighbor, LVQ).

In this work we have, so far, used only decision tree and rule-based methods because of their familiarity to us and because of their ability to give human understandable results.

Good Examples. The 'quality' of the training data is one of the most important factors in achieving good classifier performance. Training data quality is a function of the number of examples, how representative the examples are, and the attributes used to describe them.

Labeled Data. Supervised classification uses labeled training examples to build a model. The labels usually come from a human expert (or experts) who manually review cases. In our application of classification to intrusion detection we obtained labeled examples by building a web-based interface that required a label to be assigned to a new incident each time it was constructed by an analyst. Using this feedback we were able to collect 12,900 labeled examples of seven different classes of incidents from August 2000 and 16,885 for September 2000.

Classes. Another factor in getting good examples is to have a well-defined set of classes. It is important to maintain consistency in assigned labels over time, both for a single person and across multiple people. Label inconsistency can make classification very difficult especially if identical examples are labeled ambiguously.

XI. PERFORM ANOMALY DETECTION

Both intruder techniques and local network configurations will change. In spite of efforts to update defenses, new attacks may slip through defenses and be

labeled as either normal network traffic, or else filtered as a known but benign probe. Anomaly detection techniques can help humans prioritize potentially anomalous records for review. Catching new attacks can not depend on the current set of classification rules. Since classification assumes that incoming data will match that seen in the past, classification may be an inappropriate approach to finding new attacks. Much of the work in outlier detection has been approached from a statistical point of view and is primarily concerned with one or very few attributes. However, because the network data has many dimensions, we have investigated use of clustering for anomaly detection.

Clustering is an unsupervised machine learning technique for finding patterns in unlabeled data with many dimensions (number of attributes). We use k-means clustering to find natural groupings of similar alarm records. Records that are far from any of these clusters indicate unusual activity that may be part of a new attack.

The network data available for intrusion detection is primarily categorical (i.e., Attributes have a small number of unordered values). Clustering approaches for categorical data, such as in [Guha et al., 1999] are not generally available commercially. Unsupervised approaches for detecting outliers in large data sets for the purposes of fraud or intrusion detection are starting to appear in the literature, but these approaches are primarily based on ordered data. Knorr and Ng [1998] recently developed a distance-based clustering approach for outlier detection in large data sets. Ramaswarny, et al. [2000] define a new outlier criterion based on the distance of a point to its k_{th} nearest neighbor. Breunig et al. [2000] define a new local outlier factor, which is the degree to which a data point is an outlier.

XII.MAKE YOUR SYSTEM EFFICIENT

There are a number of practical considerations in building an effective intrusion detection system. Some of these derive from the use of data mining, but many of them would be present in any intrusion detection system:

- **A central repository must be designed and enabled.** The repository must allow for inputs from a potentially large number of diverse network sensors, preferably within a single data table. Any derived data, such as data mining attributes, should also be stored in this central location. It must also support the creation and tracking of security incidents.
- **Efficient querying is essential to feed the daily operations of security analysts.** A bottleneck in querying the data will affect everything else in the system. Some steps that can be taken to improve query efficiency include the inclusion of a database performance guru on the project team, statistical/ trend analysis of query performance over time, elimination of time-consuming queries, or the retirement of old data from the database.
- **Efficiency can also be improved by selecting appropriate aggregations of attributes and statistics.** A manual analysis of network activity will reveal that a large volume of atomic network activity breaks down into a much smaller set of meaningful aggregates. At MITRE, two of the more useful aggregates were (source IP, destination port), used for catching some IP mapping activity, and (source IP, destination IP), used for catching port scanning activity. But, any combination of fields or attributes could also be used, resulting in a wealth of choices. Regardless of the fields used, aggregates reduce the downstream volume of data.
- **While most attributes and aggregates are used to feed an automated process, don't forget the analysts.** Analysts must have efficient tools to spot check the automatically generated security incidents, and to manually comb through the raw sensor data for new or complex patterns of malicious activity. The MITRE interface is centered on a set of predefined queries of the sensor database, and a browser of the incident database. With this tool, an analyst can create new security incidents or update existing incidents with new status information.
- **Due to the high volume and frequency of data inputs, and the variety of both automated and human data sources, there will invariably be some process failures.** When a failure does occur, the condition must be caught and the security team notified. Scripts that verify the integrity of the data tables, and repair inconsistencies, are useful. If possible, the process should be halted until the error is corrected. But, in some situations, the ability to operate normally regardless of errors, and then rollback and correct statistics and attributes at the team's convenience, may be a more practical recovery strategy.
- **Scheduling is an important aspect of the operational environment.** Each organization must

decide for itself how much of its intrusion detection system truly needs to be "real- time". The calculation of real time statistics must be completed in a matter of seconds, and the amount of data available in this manner will always be limited. But daily batch processing of data may be adequate in many cases.

XIII. SUMMARY

We have described our experiences with integrating data mining into a network intrusion detection capability. We believe that when starting such a project you should:

- Choose your requirements carefully and be realistic.
- Assemble a team with broad, relevant capabilities.
- Invest in adequate infrastructure to support data collection and data mining.
- Design, compute, and store appropriate attributes with your data.
- Reduce data volume with filtering rules.
- Refine the overall architecture for your system, taking into account both automated processing and human analysis.
- Use data mining techniques such as classification, clustering, and anomaly detection, to suggest new filter rules.
- Make sure that automated data processing can be done efficiently.

REFERENCES

[1] Bloedorn, E., L. Talbot, C. Skorupka, A. Christiansen, W. Hill, and J. Tivel [2009]. "Data Mining applied to Intrusion Detection: MITRE Experiences," submitted to the 2001 IEEE International Conference on Data Mining.

[2] Breunig, M. M., H. P. Kriegel, R. T. Ng, and J. Sander [2008]. "LOF: Identifying Density-Based Local Outliers", Proceedings of the ACM Sigmod 2000 Intl. Conference On Management of Data, Dallas, TX.

[3] Clifton, C., and G. Gengo [2009]. "Developing Custom Intrusion Detection Filters Using Data

Mining", 2000 Military Communications International, Los Angeles, California, October 22-25.

[4] Domingos, P., and G. Hulten [2008]. "Mining High Speed Data Streams", in Proceedings of the Sixth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, p. 71-80.

[5] Guha, S., Rastogi, R., and Shim, K. [1998]. "ROCK: A Robust Clustering Algorithm for Categorical Attributes", Proceedings of the 15th Int. Conference On Data Eng., Sydney, Australia.

[6] Henery, R. J. [1994]. "Classification," Machine Learning, Neural and Statistical Classification, Michie, D., Spiegelhalter, D. J., and Taylor, C. C. (Eds.), Ellis Horwood, New York.

[7] Knorr, E. M., and R. T. Ng [1998]. "Algorithms for Mining Distance-Based Outliers in Large Datasets", VLDB'98, Proceedings of the 24th Int. Conference on Very Large Databases, Aug 24-27, 1998, New York City, NY, pp. 392-403.

[8] Lee, W., and S. Stolfo [1998]. "Data Mining Approaches for Intrusion Detection", in Proceedings of the 7th USENIX Security Symposium, San Antonio, TX.

[9] Manganaris, S., M. Christensen, D. Zerkle, and K. Hermiz [2000]. "A data mining analysis of RTID alarms", Computer Networks, 34, p. 571-577.

[10] Ramaswarny, S., R. Rastogi, and K. Shim, [2010]. "Efficient Algorithms for Mining Outliers from Large Data Sets", Proceedings of the ACM Sigmod 2000 Int. Conference on Management of Data, Dallas, TX.

❑❑❑

Decision Tree Induction
- A Heuristic Problem Reduction Approach

D. Raghu, K.Venkata Raju & Ch.Raja Jacob

Department of Computer Science Engineering,Nova college of Engineering&Technology,
Jangareddy gudem, West Godavari District, Andhra Pradesh, India-.534447
E-mail : raghuau@gmail.com, venkatsagar05@gmail.com, rchidipi@gmail.com

Abstract - ID3, a greedy decision tree induction technique constructs decision trees in top-down recursive divide and conquer manner. This supervised learning algorithm is used for classification. This paper introduces a new Heuristic technique of Problem Reduction that provides optimal solution based on evaluation of heuristic function applied to every node in the tree. The attribute selection can be done by considering the best attribute having least Heuristic function value by which we provide the optimal Decision Tree. This provides fast and more efficient method of constructing the Decision Tree taking predominant function of heuristic that best selects an attribute for classification.

Keywords— Decision tree, Problem Reduction, Heuristic function, Attribute Selection Criteria.

I. INTRODUCTION

Data Mining is a powerful technology used in the data warehouses. Data mining is to discover the relationship and rules existing in data, to predict the feature trends based on the existing data, finally to fully explore and use these wealth knowledge hiding in the databases [2],[3].Decision tree induction is most widely used practical method for inductive learning; it plays an important role in the process of data mining and data analysis.

As the Classification by Decision tree induction uses attribute selection measure is taken as information gain heuristic. This paper proposes a new heuristic approach for attribute selection, which is done by applying the Heuristic function to each attribute that best selects an attribute for further split of the tree.

II. ID3 ALGORITHM

ID3 algorithm [4] is a decision tree learning algorithm based on information entropy proposed by Quinlan in 1986. The core of ID3 algorithm is: selecting attributes from all levels of decision tree nodes; using information gain as attribute selection criteria; each selecting an attribute with the largest information gain to make decision tree nodes; establishing branches by the different values of the node; building the tree nodes and branches recursively according to the instances of various branches; until a certain subset of the instances belonging to the same category.

Set S is a collection of data samples. Assume that class label attribute has m different values, define different classes C_i (i=1, 2...m). Let s_i represent the sample number of classes C_i. For a given sample the expected information for the classification is calculated as follows [1]:

$$I(s_1, s_2 \ldots s_m) = -\sum_{i=1}^{m} p_i \log p_i \qquad (1)$$

Where p_i is the probability of any sample belonging to class C_i. Assumed attribute A has v distinct values :{ a_1, a_2 ...a_v} divides s into v sub-sets {s_1, s_2 ...s_v} by attribute A. Making A as test attribute, this subset is the branch getting from the nodes including set S. Assumed s_{ij} is the sample number of class C_i in subset s_j. The Entropy divided by A is given as follows [3]:

$$E(A) = -\sum_{J=1}^{v} \frac{s_{1j}+s_{2j}+\ldots+s_{mj}}{s} I(s_{1j}, s_{2j} \ldots s_{mj}) \qquad (2)$$

Here $\frac{s_{1j}+s_{2j}+\ldots+s_{mj}}{s}$ is the weight of the j subset.

And is equal to the sample number of subset dividing the total number of S. The smaller entropy is the higher purity of the divided subset. For the given subset S_j, its expected information is given as [3]:

$$I(s_{1j}, s_{2j}... s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log p_{ij} \quad (3)$$

Where $p_{ij} = S_{ij}$ is the probability that the sample S_j

Belongs to the class C_i . Get information gain according to the expected information and entropy. Information Gain is calculated by [3]:

$$Gain(A) = I(s_1, s_2... s_m) - E(A) \quad (4)$$

Calculate the information gain off each attribute according to ID3 algorithm and select the attribute with highest information as the test attribute for a given set. Create the node of the selected test attribute ark according to this attribute create a branch for each value of this attribute and accordingly divide the sample.

III. PROBLEM REDUCTION IMPLEMENTATION

In this approach the new algorithm applies heuristic function on each attribute and picks the best attribute with least heuristic function value.f(n) The heuristic function is obtained from Problem Reduction algorithm which is defined by [5]:

f(n) = estimated cost from this node to the goal node the algorithm terminates when f(n)=0 or f(n)>FUTILITY value. In general ID3 the attribute selection criteria is done by the process of information gain heuristic. This Paper proposes a new attribute selection criteria by applying heuristic function f(n) on each attribute.

Attribute Selection Criteria:

Step-1: Calculation of f (n)

f(n)=E(n) with respect to goal attribute Buys_Computer

Step-2: Calculation of FUTILITY value

$$FUTILITY = (\sum_{i=1}^{m} f(i))/m$$

where m is the number of attributes.

Thus step-1 through step-2 are recursively applied for building the Decision Tree. The algorithm terminates when f(n) =0 or f(n) is greater than FUTILITY.

IV. ALGORITHM VALIDATION

Table I: The Data Set used in the Algorithm

RID	AGE	INCOME	Student	Credit	Buys
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31-40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31-40	medium	no	excellent	yes
13	31-40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Table 1 gives the data set [3] affecting Buys_computer. There are four attributes: age, income, student, credit. These four attributes are used to specify Buys_computer and the algorithm is constructed as follows:

I. Initial Attribute:

Table II: The Result Set for the Initial Attribute

f	Attribute
0.6935361388961918	Age
0.9110633930116763	Income
0.7884504573082896	Student
0.8921589282623617	Credit_Rating
Futility Value:	0.8213022293696297
Attribute Selected=age	

Now according to the calculated values the training set is classified depending on attribute “Age” which splits up into 3 values. Since of the two attributes whose f(n) is less than the futility “AGE” is having less values when compared to “STUDENT”.

II. Next Attribute with age<=30

Table III: The Result Set for attribute age<=30

f	Attribute
0.4	Income
0.0	Student
1.0	Credit_Rating
Futility Value:	0.4666666666666666
Attribute Selected=student	

So the next attribute on split age<=30 is "Student". As the f(n) value is 0 this node is labelled as solved.

III. Next Attribute with age=31-40

Table IV: The Result Set for attribute age=31-40

f	Attribute
0.0	Income
0.0	Student
0.0	Credit_rating
Futility value:	0.0
No Further Split	

Since all the f(n)values is zero there is no further splits and it is labelled as solved.

IV. Next Attribute with age>40

Table V: The Result Set for attribute age>40

f	Attribute
0.9509775004326937	Income
0.9509775004326937	Student
0.04675342168646712	Credit_rating
Futility Value:	0.6495694741839515
Attribute Selected:	credit

The attribute that was selected is "CREDIT_RATING " on split age>40 since the f(n) value is less than FUTILITY value.

V. *Next Attribute with age<=30 and student=yes*

Table VI: The Result Set for attribute Student=yes

f	Attribute
0.0	Income
0.0	Credit Rating
Futility Value:	0.0
No Further Split	

Since all the f(n) values are 0 the node is labelled as solved and there is no further splits

VI. Next Attribute with age<=30 and student=no

Table VII: The Result Set for attribute Student=no

f	Attribute
0.0	Income
0.0	Credit_Rating
Futility Value	0.0
No Further Split	

Since all the f(n) values are 0 the node is labelled as solved and there is no further splits

There ends on split age<=30 and skips to age=31-40 and there is no further split on that branch algorithm continues on branch age>40.

VII. Next Attribute with age>40 and credit=fair

Table VIII: The Result Set for attribute Credit=fair

f	Attribute
0.0	Income
Futility Value:	0.0
No Further Split	

Since all the f(n) values are 0 the node is labelled as solved and there is no further splits

VIII. Next Attribute with age>40 and credit=excellent

Table IX: The Result Set for attribute Credit=excellent

f	Attribute
0.0	Income
Futility Value:	0.0
No Further Split	

Since all the f(n) values are 0 the node is labelled as solved and there is no further splits

The algorithm Terminates and the final decision tree is as follows

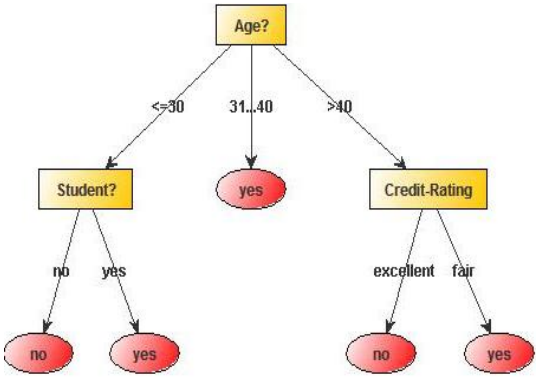


Fig.1 Decision Tree

V. CONCLUSION

The proposed paper created a new innovative approach of attribute selection criteria taking best splitting attribute with least heuristic functional value. It reduces the number of calculations by generating the decision tree with out calculating the gain factor. This paper can be extended to larger databases which can provide best splitting criteria.

REFERENCES

[1] Liu Yuxun, Xie Niuniu *"Improved ID3 Algorithm"* 3RD IEEE Conference (ICCSIT.2010.5564765):465-468.

[2] G.Eason, B.Noble, and I.N.Sneddon, "On *certain integrals of W.J.Frawley, G.Piatetsky Shapiro, C.J.Matheu. Knowledge discovery in databases: An overview[M]"*, USA, AAAI/MIT Press, 1991 1-27

[3] Han Jiawei, Micheline Kamber *Data Mining Concepts and Techniques*

[4] J.R.QUINLAN.*Induction of Decision Trees[J]*.Machine Learning 1968:81-106.

[5] E.Rich, K.Knight, Shiv Shankar B Nair *Artificial Intelligence.*

❑❑❑

Binary Decision Trees for Uncertain Data Using Probability Distribution

M. Rajani Kanth & Katru Rama Rao

Dept of Computer Science and Engineering,
Sri Sunflower college of Engineering and Technology, Lankapalli
Email:rajinikanthmeka7@yahoo.co.in, katruramarao@gmail.com

Abstract - Traditional decision tree classifiers work with data whose values are known and precise. We extend such classifiers to handle data with uncertain information. Value uncertainty arises in many applications during the data collection process. Example sources of uncertainty include measurement/quantization errors, data staleness, and multiple repeated measurements. With uncertainty, the value of a data item is often represented not by one single value, but by multiple values forming a probability distribution. Rather than abstracting uncertain data by statistical derivatives (such as mean and median), we discover that the accuracy of a decision tree classifier can be much improved if the "complete information" of a data item (taking into account the probability density function (pdf)) is utilised. We extend classical decision tree building algorithms to handle data tuples with uncertain values. Extensive experiments have been conducted that show that the resulting classifiers are more accurate than those using value averages. Since processing pdf's is computationally more costly than processing single values (e.g., averages), decision tree construction on uncertain data is more CPU demanding than that for certain data. To tackle this problem, we propose a series of pruning techniques that can greatly improve construction efficiency.

Index Terms—*Uncertain Data, Decision Tree, Classification, Data Mining*

I. INTRODUCTION

Classification is a classical problem in machine learning and data mining. Given a set of training data tuples, each having a class label and being represented by a feature vector, the task is to algorithmically build a model that predicts the class label of an unseen test tuple based on the tuple's feature vector. One of the most popular classification models is the decision tree model. Decision trees are popular because they are practical and easy to understand. Rules can also be extracted from decision trees easily. Many algorithms have been devised for decision tree construction.

These algorithms are widely adopted and used in a wide range of applications such as image recognition, medical diagnosis, credit rating of loan applicants, scientific tests, fraud detection, and target marketing.

In traditional decision-tree classification, a feature (an attribute) of a tuple is either categorical or numerical. For the latter, a precise and definite point value is usually assumed. In many applications, however, data uncertainty is common. The value of a feature/attribute is thus best captured not by a single point value, but by a range of values giving rise to a probability distribution. A simple way to handle data uncertainty is to abstract probability distributions by summary statistics such as means and variances. We call this approach Averaging. Another approach is to consider the complete information carried by the probability distributions to build a decision tree. We call this approach Distribution-based.

II. RELATED WORK

There has been significant research interest in uncertain data management in recent years. Data uncertainty has been broadly classified as existential uncertainty and value uncertainty. Existential uncertainty appears when it is uncertain whether an object or a data tuple exists. For example, a data tuple in a relational database could be associated with a probability that represents the confidence of its presence. "Probabilistic databases" have been applied to semi-structured data and XML. Value uncertainty, on the other hand, appears when a tuple is known to exist, but its values are not known precisely. A data item with value uncertainty usually represented by a pdf over a finite and bounded region of possible values One well-studied topic on value uncertainty is "imprecise queries processing". The answer to such a query is associated with a probabilistic guarantee on its correctness. For

example, indexing solutions for range queries on uncertain data, solutions for aggregate queries such as nearest neighbour queries, and solutions for imprecise locationdependent queries have been proposed. There has been a growing interest in uncertain data mining.well-known k-means clustering algorithm is extended to the UK-means algorithm for clustering uncertain data. As we have explained, data uncertainty is usually captured by pdf's, which are generally represented by sets of sample values. Mining uncertain data is therefore computationally costly due to information explosion (sets of samples vs. single values). To improve the performance of UK-means, pruning techniques have been proposed. Examples include minmaxdist pruning and CK-means. Apart from studies in partition-based uncertain data clustering, other directions in uncertain data mining include density-based clustering (e.g., FDBSCAN), frequent itemset mining and density based classification. Density-based classification requires that the joint probability distribution of the data attributes be known.

Each data point is given an error model. Upon testing, each test tuple is a point-valued data. These are very different from our data model, as we do not require the knowledge of the joint probability distribution of the data attributes. Each attribute is handled independently and may have its own error model. Further, the test tuples, like the training tuples, may contain uncertainty in our model. Decision tree classification on uncertain data has been addressed for decades in the form of missing values.

2.2.1The algorithm in this paper :

A. Averaging

A straight-forward way to deal with the uncertain information is to replace each pdf with its expected value, thus effectively converting the data tuples to point-valued tuples. This reduces the problem back to that for pointvalued data, and hence traditional decision tree algorithms such as ID3 and C4.5[3] can be reused. We call this approach AVG (for Averaging). We use an algorithm based on C4.5. Here is a brief description. AVG is a greedy algorithm that builds a tree top-down. When processing a node, we examine a set of tuples S. The algorithm starts with the root node and with S being the set of all training tuples. At each node n, we first check if all the tuples in S have the same class label c. If so, we make n a leaf node and set Pn(c) = 1, Pn(c0) = 0 8c0 6= c.

Otherwise, we select an attribute Ajn and a split point zn and divide the tuples into two subsets: "left" and "right". All tuples with vi;jn _ zn are put in the "left" subset L; the rest go to the "right" subset R. If either L or R is empty (even after exhausting all possible choices of Ajn and zn), it is impossible to use the available attributes to further discern the tuples in S. In that case, we make n a leaf node. Moreover, the population of the tuples in S for each class label induces the probability distribution Pn. In particular, for each class label c 2 C, we assign to Pn(c) the fraction of tuples in S that are labelled c. If neither L nor R is empty, we make n an internal node and create child nodes for it. We recursively invoke the algorithm on the "left" child and the "right" child, passing to them the sets L and R,respectively.

B. Distribution-based

For uncertain data, we adopt the same decision tree building framework as described above for handling point data. After an attribute Ajn and a split point zn has been chosen for a node n, we have to split the set of tuples S into two subsets L and R. The major difference from the point-data case lies in the way the set S is split. Recall that the pdf of a tuple ti 2 S under attribute Ajn spans the interval [ai;jn; bi;jn]. If bi;jn _ zn, the pdf of ti lies completely on the left of the split point and thus ti is assigned to L.

Similarly, we assign ti to R if zn < ai;jn. If the pdf properly contains the split point, i.e., ai;jn _zn < bi;jn, we split ti into two fractional tuples tL and tR in the same way as described in Section III-B and add them to L and R, respectively. We call this algorithm UDT (for Uncertain Decision Tree).

III. PROPOSEDWORK:

The algorithms described above have been implemented in Java using JDK 1.6 and a series of experiments were performed on a PC with an Intel Core 2 Duo 2.66GHz CPU and 2GB of main memory, running Linux kernel 2.6.22 i686. experiments on the accuracy of our novel distribution-based UDT algorithm has been presented already in Section IV-B. In this section, we focus on the pruning effectiveness of our pruning algorithms and their run-time performance.

3.1 Execution Time

We first examine the execution time of the algorithms, which is charted in Figure 6. In this figure, 6 bars are drawn for each data set. The vertical axis, which is in log scale, represents the execution time in seconds. We have given also the execution time of the AVG algorithm (see Section IV-A). Note that AVG builds different decision trees from those constructed by the UDT-based algorithms, and that AVG generally builds less accurate classifiers. The execution time of AVG shown in the figure is for reference only. From the figure, we observe the following general (ascending) order of efficiency: UDT, UDT-BP, UDT-LP, UDT-GP, UDT-ES. This agrees with the successive enhancements

of these pruning techniques discussed in Section V. Minor fluctuations are expected as the pruning effectiveness depends on the actual distribution of the data. The AVG algorithm, which does not exploit the uncertainty information, takes the least time to finish, but cannot achieve as high an accuracy compared to the distribution-based algorithms

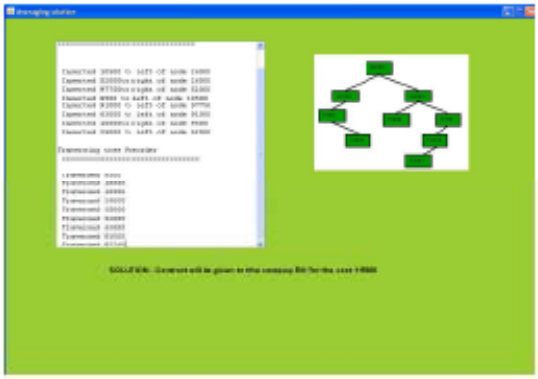
3.2. Pruning Effectiveness

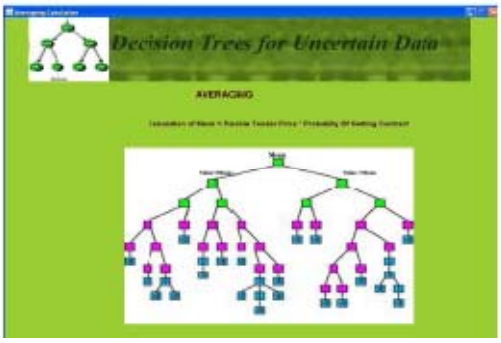
Next, we study the pruning effectiveness of the algorithms. It shows the number of entropy calculations performed by each algorithm. As we have explained, the computation time of the lower bound of an interval is comparable to that of computing an entropy. Therefore, for UDT-LP, UDT-GP,and UDT-ES, the number of entropy calculations include the number of lower bounds computed.

The figure shows that our pruning techniques introduced in this Section are highly effective. Comparing the various bars against that for UDT, it is obvious that a lot ofentropy calculations are avoided by our bounding techniques (see Section V-B). Indeed, UDT-BP only needs to perform 14%–68% of the entropy calculations done by UDT. This corresponds to a pruning of 32%–86% of the calculations.UDT-LP does even fewer calculations: only 5.4%–54% of those of UDT.

By using a global pruning threshold, UDTGP only needs to compute 2.7%–29% of entropy values compared with UDT. By pruning endpoints, UDT-ES further reduces the number of entropy calculations to 0.56%–28%. It thus achieves a pruning effectiveness ranging from 72% up to as much as 99.44%. As entropy calculations dominate the execution time of UDT, such effective pruning techniques significantly reduce the tree-construction time.

IV. EXPERIMENTAL RESULTS





V. CONCLUSION AND FUTURE WORK

We have extended the model of decision-tree classification to accommodate data tuples having numerical attributes with uncertainty described by arbitrary pdf's. We have modified classical decision tree building algorithms (based on the framework of C4.5[3]) to build decision trees for classifying such data.

We have found empirically that when suitable pdf's are used, exploiting data uncertainty leads to decision trees with remarkably higher accuracies.

We therefore advocate that data be collected and stored with the pdf information intact. Performance is an issue, though, because of the increased amount of information to be processed, as well as the more complicated entropy computations involved. Therefore, we have devised a series of pruning techniques to improve tree construction efficiency. Our algorithms have been experimentally verified to be highly effective. Their execution times are of an order of magnitude comparable to classical algorithms. Some of these pruning techniques are generalisations of analogous techniques for handling point-valued data.

Other techniques, namely pruning by bounding and end-point sampling are novel. Although our novel techniques are primarily esigned to handle uncertain data, they are also useful for building decision trees using classical algorithms when there are tremendous amounts of data tuples.

6.REFERENCES

[1] R. Agrawal, T. Imielinski, and A. N. Swami, "Database mining: A performance perspective," IEEE Trans. Knowl. Data Eng., vol. 5, no. 6, pp. 914–925, 1993.

[2] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

[3], C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993, ISBN 1-55860-238-0.

[4] C. L. Tsien, I. S. Kohane, and N. McIntosh, "Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit," Artificial Intelligence in Medicine, vol. 19, no. 3, pp. 189– 202, 2000.

[5] G. L. Freed and J. K. Fraley, "25% "error rate" in ear temperature sensing device," Pediatrics, vol. 87, no. 3, pp. 414–415, Mar. 1991.

[6] O. Wolfson and H. Yin, "Accuracy and resource consumption in trackingand location prediction," in SSTD, ser. Lecture Notes in Computer Science, vol. 2750. Santorini Island, Greece: Springer, 24-27 Jul. 2003, pp. 325–343.

[7] W. Street, W. Wolberg, and O. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in SPIE, vol. 1905, San Jose, CA, U.S.A., 1993, pp. 861–870. [Online]. Available: http: //citeseer.ist.psu.edu/street93nuclear.html

[8] N. N. Dalvi and D. Suciu, "Efficient query evaluation on probabilistic databases," The VLDB Journal, vol. 16, no. 4, pp. 523–544, 2007. [9] E. Hung, L. Getoor, and V. S. Subrahmanian, "Probabilistic interval XML," ACM Transactions on Computational Logic (TOCL), vol. 8, no. 4, 2007.

[10] A. Nierman and H. V. Jagadish, "ProTDB: Probabilistic data in XML," in VLDB. Hong Kong, China: Morgan Kaufmann, 20-23 Aug. 2002, pp. 646–657.

❑❑❑

Routing Protocols in Wireless Networks

D. Kishore Kumar[1], R. Sujitha Rani[2] & YSSR Muthry[3]

[1]Department of Computer Science & Engineering, GITAM Institute of Technology,
GITAM University: Visakhapatnam, Andhra Pradesh
[2&3]Department of Computer Science & Engineering, Shri Vishnu Engineering College For Women,
Vishnupur, Bhimavaram, Andhra Pradesh,
E-mail : kishore_dasari@yahoo.com[1], suji.rimmanapudi@gmail.com[2], yssrmoorthy@gmail.com[3]

Abstract - Multi-hop ad hoc wireless network is one of the greatest breakthroughs in the field of science and technology. Mobile Multi-hop Ad Hoc Networks are collections of mobile nodes connected together over a wireless medium. These nodes can freely and dynamically self-organize into arbitrary and temporary, "ad-hoc" network topologies with the help of "The Dynamic Source Routing Protocol" which is an easy and well-organized routing protocol. There are two mechanisms namely Route Discovery and Route Maintenance by which nodes can discover and maintain source routes to arbitrary destinations in the ad hoc network. The use of source root allows a sender of a packet to partially or completely specify the route the packet takes through the network and it further allows for easier troubleshooting and better route-tracing,. It enables a node to discover all the possible routes to a host for further use as well. This paper is an attempt to evaluate the operation of DSR on a variety of movements and communication patterns in a physical outdoor ad hoc networking test bed. This paper further presents the description of the design of DSR and important properties of protocol.

Keywords - *Multi-hop Networks; mobile nodes; routing protocolse;testbed.*

I. INTRODUCTION

The availability and the usage of the wireless networking is increasing in a wider range and the demand for ad hoc networks is rising high as it is a collection of mobile hosts with wireless network interfaces which form a temporary network without the aid of any established infrastructure or centralized administration.

The *Dynamic Source Routing* protocol is the most competent routing protocol for wireless mesh networks as it uses source routing instead of depending on the routing table at each intermediate device. This protocol enables the network to be self-organized and self-configured. Dynamic source routing protocol (DSR) is an on-demand protocol designed to restrict the band width consumed by control packets in ad hoc wireless networks by eliminating the periodic table-update messages required in the table-driven approach. The major difference between this and the other on-demand routing protocols is that it is beacon-less and hence does not require periodic hello packet (beacon) transmissions, which are used by a node to inform its neighbors of its presence.

DSR Protocol uses a reactive approach which eliminates the need to periodically flood the network with table update messages which are required in a table-driven approach. In a reactive (on-demand) approach such as this, a route is established only when it is required and hence the need to find routes to all other nodes in the network as required by the table-driven approach is eliminated. The intermediate nodes also utilize the route cache information efficiently to reduce the control overhead.

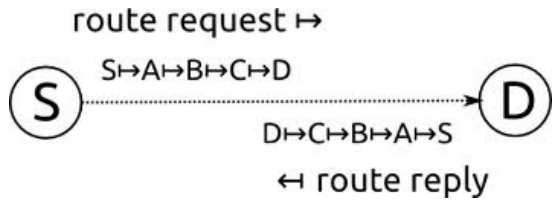
One of the most prominent features of an ad hoc network is that, two hosts can communicate with the help of other networks which are taking part in the same transmission, even though they both are not within the wireless transmission range of each other This paper describes the design and performance of dynamic routing protocol which offers a number of prospective advantages in comparison to conventional protocols such as distance vector in an ad hoc network. In DSR, nodes discover a source route dynamically across multiple network hops to any destination in the ad hoc network. The complete order list of notes can be carried by each data packet.

II. DESIGN AND BASIC OPERATION OF DSR PROTOCOL:

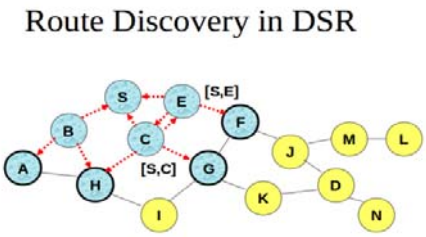
There are two basic mechanisms of DSR, Route Discovery and Route Maintenance.

Route Discovery is the mechanism by which a node can discover a route to any host in the ad hoc network. The primary use of route discovery is to figure out a route by which packets can be sent from one node to another node.

Route Maintenance is for the purpose of identifying whether there are any changes in the network topology such that the nodes can be aware of the available routes for sending the packets.



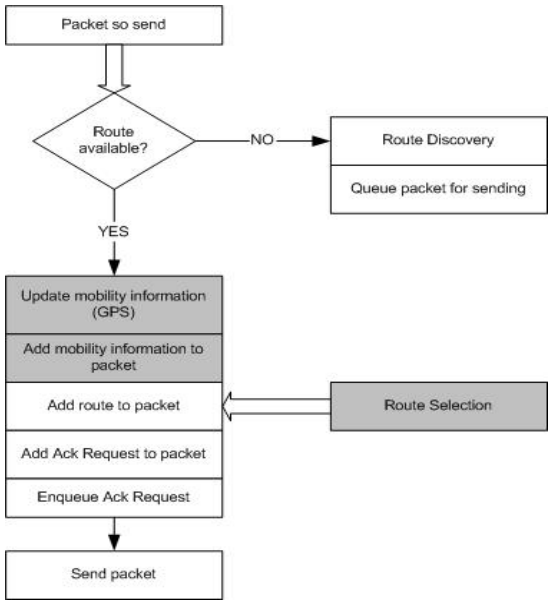
Route discovery allows any host in the ad hoc network to dynamically discover a route to any other host in the ad hoc network, whether directly reachable within wireless transmission range or reachable through one or more intermediate network hops through other hosts. The route request packet identifies the host, referred to as the *target* of the route discovery, for which the route is requested. A route reply packet listing a sequence of network hops received by the initiating host witnesses the successful discovery of the route. Each route request will also contain a route record, which consists of an accumulated record of the sequence of hops taken by the route request packet along with the original initiator and the target of the request. Each route request packet will have a unique request id. A list of the initiator address and request id pairs is maintained to avoid duplicate route requests.



In the process of a route request packet, a host checks the initiator address and request whether it already exists or not. If the request exists, it discards the route request and stops the process. Otherwise, if this host's address is already listed in the route record in the request, then it discards the route request packet and does not process it further. If there is a match between the target and host's address, then the route by which the route record reached this node will be sent back to the source in a route reply packet in a reverse order. If not, Intermediate-Node reply will take place. In the case of these node replies, it checks whether this node already exists or not, if that request is not there before, it will be sent back to the source.

III. ROUTE MAINTAINANCE:

Continuous routing updates are done in conventional routing protocols as it puts together route discovery with route maintenance. When there is any change in the status of a link, the changes will be reflected by the periodic updates to all other routers which may result in the calculation of new routes. Conventional routing protocols put together route discovery with route maintenance by continuously sending periodic routing updates. If the status of a link or router changes, the periodic updates will eventually reflect the changes to all other routers, presumably resulting in the computation of new routes.



Uni-directional links and asymmetric routes are allowed and they can be easily supported in the operation of Route discovery and Route Maintenance. There is a possibility that the performance of link between two nodes may not work equally in wireless networks. For this, DSR provides a solution and increases the overall performance. Inter-networking between different types of wireless networks is best done with DSR. For instance, there are some nodes consisting short-range radios and there may be some

other nodes with short-range and long-range radios and all these can be measured as a single ad hoc network by DSR.

Route maintenance can be easily provided in many wireless networks, which use a hop-by-hop acknowledgement at the data link so that they can detect and retransmit the lost or corrupted packets, as the performance of that route can be traced out by the host transmitting the packet and a route error packet is sent to the original sender of the packet if any transmission problem occurs. End-to-end acknowledgements could also be used rather than the hop-by-hop acknowledgements in Route Maintenance in situations like where the working of wireless transmissions between two hosts doesn't work well in both directions. The root maintenance is possible when there are two hosts which can communicate. In these situations, the status of one host's route is indicated to the other host with the help of existing transport or application replies or requested network acknowledgements. The difference between hop-by-hop and end-to-end acknowledgements is, in hop-by-hop, the route error packet indicates the error where as in end-to-end acknowledgement it is the sender who will know the error.

IV. OPTIMIZATIONS:

The routing information can be ascertained from a node that is forwarding any packet to the route cache. Routing information from the source route or the returned in Route Reply may also be cached by any node. It is possible to cache the routing information whether the node is sent to a broadcast or multicast MAC address or the packet is addressed to this node.

If a node receives a non target request, then it searches and sends a Route Reply to the initiator itself rather than forwarding the Route Request and the sequence of hops which is used as a transport is listed in the route record.

DSR plays a major role in a heterogeneous network which isa network connecting computers and other devices with different operating systems. It shows it's undeniable impact in Mobile IP which is a standard that allows users with mobile devices whose IP address are associated with one network to stay connected when moving to a network with a different IP address.

DSR would allow the packets to be routed transparently from the ad hoc network to nodes in the Internet and vice-versa so as to facilitate seamless interpretation between these networks. Internet can participate in an ad hoc network through DSR when the nodes in the ad hoc network are connected to the Internet to enable interoperation and this is called a 'gateway' between these two networks.

V. EVALUATIONS:

We have taken into consideration a few evaluations on the performance of DSR done by *David B. Johnson David A. Maltz Josh Broch* and have made an analysis through detailed studies using discrete event simulation and implemented an actual operation in an ad hoc networking test bed environment. An environment which consists of a set of wireless and mobile networking extensions was selected.

1. A number of different simulation studies have been done with this environment, analyzing the behavior and performance of DSR and comparing it to other proposed routing protocols for ad hoc networks.
2. Only some of the basic results have been summarized and all simulations were run in ad hoc networks of 50 mobile nodes moving according to the random waypoint mobility model within a flat rectangular area.
3. The simulation time was 15 minutes and the scenario file was used to capture all movement and application-layer communication generated in advance. The Lucent Wave LAN direct sequence spread spectrum radio has been approximated by the characteristics of the physical radio such as the antenna gain, transmit power, and receiver sensitivity.
4. During the simulation, the beginning of each node happens at a random location and has an independent movement in the random waypoint mobility model.
5. The state of (inactive notion) is called as the pause time and each node remains stationary during that time and then motion happens through a straight line to a new randomly chosen location.
6. An evaluation of two different node speeds has confirmed that DSR delivers almost all data packets, regardless of pause time, with packet delivery ratio rising to equal 100% at pause time 900 (a stationary network) and the routing overhead is low and going up slowly as pause time decreases.
7. DSR has shown a marked ability for delivering more than 99.5% of all packets at the lower movement speed of 1 meter/second. In most cases the delivery ability was greater than 99.8% of all packets. A random generation of the scenarios used in the simulations caused a slight decrease of 30%. Even at pause time 0 and a high speed of 20 meters/second, DSR showed an efficiency rate of 98%.

VI. SUMMERY OF TESTBED AND DSR IMPLEMENTATION:

The behavior of DSR was tested on a real network using the FreeBSD version of UNIX. The entire code that was used to implement DSR exists in the kernel in a module that straddles the IP layer. The dsr0 interface allows packets from the normal IP and utilizes its own mechanisms for their delivery via the actual physical network interfaces. A packet model for DSR modeled after the extension header and option format used by IPv6 has been used for a combination of multiple types of DSR information in a single packet and for piggybacking on existing packets.

Total 5 mobile nodes implemented as cars have been taken for the testbed analysis and all these cars were continuously driven in a loop and the route between the two stationary nodes was constantly changing as the cars moved. A laptop computer implementation of the DSR routing protocol has been used in each car which served as an endpoint in various higher layer protocol connections and applications. The operation of the testbed was of bulk file transfer, telnet, constant bit rate UDP streams loading the network similar to voice or video.

VII.CONCLUSION

This paper gives a brief analysis of Dynamic Resource Protocol's routing performance. It is clear that Ad hoc and wireless networking has been gaining critical attention from researchers as the available wireless networking and mobile computing hardware bases are now capable of supporting the promise of this technology. DSR is a broadly used routing protocol for mobile ad hoc networks and is able to correctly deliver almost all originated data packets, even with continuous, rapid motion of all nodes in the network. Dynamic source routing is a protocol which adapts quickly to routing changes when host movement is frequent, yet requires little or no overhead during periods in which hosts move less frequently. A key reason for this good performance is the fact that DSR operates *entirely* on demand [Johnson 1994], with *no* periodic activity of *any kind* required at *any level* within the network. For example, DSR does not use any periodic routing advertisement, link status sensing, or neighbor detection packets, and does not rely on these functions from any underlying protocols in the network. The operation of DSR relies entirely on demand and this is the major reason for the good performance of DSR which doesn't require periodic activity at any level.

REFERENCES

[1] F. Xue and P. R. Kumar, "The number of neighbors needed for connectivity of wireless networks," Wireless Networks.

[2] J. D´ıaz, V. Sanwalani, M. Serna, and P. Spirakis, "The chromatic and clique numbers of random scaled sector graphs," Tech. Rep., Research report,.

[3] David A. Maltz, Josh Broch, and David B. Johnson. Experiences Designing and Building a Multi-Hop Wireless Ad Hoc Network Testbed. Technical Report .

[4] CMU-CS-99-116, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania

[5] [Monarch] Carnegie Mellon University Monarch Project. CMU Monarch Project Home Page. Available at http://www.monarch.cs.cmu.edu/.

[6] David A. Maltz, Josh Broch, Jorjeta Jetcheva, and David B. Johnson. The Effects of On-Demand Behavior in Routing Protocols for Multi-HopWireless Ad Hoc Networks.

[7] IEEE Journal on Selected Areas of Communications, IEEE Computer Society LAN MAN Standards Committee. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.

[8] James Geier, Martin DeSimio, and Byron Welsh. Network Routing Techniques and their Relevance to Packet Radio Networks.

❑❑❑

Detection of Wormhole Attack in Wireless Sensor Networks

Kuldeep kaur, Vinod Kumar & Upinderpal Singh

Department of Computer Science & Engineering, Lovely Professional University Punjab, India
E-mail : deepu_judge14@yahoo.co.in, vinod.15779@lpu.co.in & upinder.singh418 @gmail.com

Abstract - The wireless sensor network is the collection of sensor nodes which collect information from the environment, the environment may be the building, industrial, battle field or elsewhere. Due to the wireless nature of the sensor nodes they are prone to various attacks like wormhole attack, grayhole, packet flooding, sinkhole attack, blackhole attack, sync attack, Sybil attack. In this paper I proposed the solution to detect the wormhole attack. In this solution I use the concept of digital signature in the packet header information .Using this solution the sensor nodes can be authenticate and can avoid the wormhole attack as possible in wireless sensor networks.

Keywords - *Hello, Sensor.*

I. INTRODUCTION

A Wireless Sensor Network [1] is a self-configuring network of small sensor nodes communicating among themselves using radio signals, and deployed in quantity to sense, monitor and understand the physical world. The wireless sensor nodes are called motes. A huge number of these devices configure the network and these motes have following capabilities: -

1) Computational capabilities.
2) Sensing capabilities.
3) Communication capabilities.

As we know that wireless sensor network technology is a technology in which sensor works under the rigorous conditions where human cannot survive for long. The major challenge in the field of wireless sensor technology is the energy consumption along with good bandwidth. This issue requires innovative design techniques to use the available bandwidth and energy efficiency.

II. ARCHITECTURE FOR NODES IN WIRELESS SENSOR NETWORKS

The nodes have to meet the requirement of a specific application. They should be small cheap, portable and energy efficient. The basic components of a node are:

1) **Sensor and actuator** - An interface to the physical world designed to sense the environmental parameters like pressure and temperature.
2) **Controller –**It is used to control different modes of operation for processing of data.
3) **Memory** - Storage for programming data.
4) **Communication** - A device like antenna for sending and receiving data over a wireless channel.
5) **Power Supply**- Supply of energy for smooth operation of a node like battery[2].

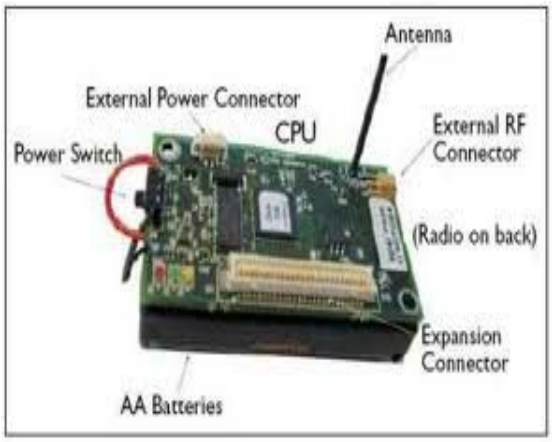


Fig. 1: Architecture of sensor node

III. ATTACKS IN WIRELESS SENSOR NETWORKS

The open nature of the wireless communication channels, the lack of infrastructure, the fast deployment practices, and the hostile environments where they may

be deployed, make them vulnerable to a wide range of security attacks. The attacks such as[1]

1) Spoofed, altered, or replayed routing information
2) Selective forwarding
3) Sinkhole attacks
4) Sybil attacks
5) Wormholes
6) HELLO flood attacks

Spoofed, altered, or replayed routing information

The most direct attack against a routing protocol is to target the routing information exchanged between nodes. By spoofing, altering, or replaying routing information, adversaries may be able to create routing loops, attract or repel network traffic, extend or shorten source routes, generate false error messages, partition the network, increase end-to-end latency, etc.[1]

Selective forwarding

Multi-hop networks are often based on the assumption that participating nodes will faithfully forward received messages. In a selective forwarding attack, malicious nodes may refuse to forward certain messages and simply drop them, ensuring that they are not propagated any further. A simple form of this attack is when a malicious node behaves like a black hole and refuses to forward every packet .[3]

Sinkhole attacks

In a sinkhole attack, the adversary's goal is to lure nearly all the traffic from a particular area through a compromised node, creating a metaphorical sinkhole with the adversary at the centre. Because nodes on, or near, the path that packets follow have many opportunities to tamper with application data, sinkhole attacks can enable many other attacks . Sinkhole attacks typically work by making a compromised node look especially attractive to surrounding nodes with respect to the routing algorithm. [2][3].

Sybil attack

In a Sybil attack a single node presents multiple identities to other nodes in the network. The Sybil attack can significantly reduce the effectiveness of fault-tolerant schemes such as distributed storage and multipath. Replicas, storage partitions, or routes believed to be using disjoint nodes could in actuality be using a single adversary presenting multiple identities.[2]

HELLO flood attack

Many protocols require nodes to broadcast HELLO packets to announce themselves to their neighbors, and a node receiving such a packet may assume that it is within radio range of the sender. This assumption may be false: a laptop-class attacker broadcasting routing or other information with large enough transmission power could convince every node in the network that the adversary is its neighbor.[3]

Wormhole Attack

In the wormhole attack, an attacker tunnels messages received in one part of the network over a low latency link and replays them in a different part. The simplest instance of this attack is a single node situated between two other nodes forwarding messages between the two of them. However, wormhole attacks more commonly involve two distant malicious nodes colluding to understate their distance from each other by relaying packets along an out-of-bound channel available only to the attacker. An attacker situated close to a base station may be able to completely disrupt routing by creating a well-placed wormhole. An attacker could convince nodes who would normally be multiple hops from a base station that they are only one or two hops away via the wormhole. This can create a sinkhole: since the attacker on the other side of the wormhole can artificially provide a high-quality route to the base station, potentially all traffic in the surrounding area will be drawn through if alternate routes are significantly less attractive. This will most likely always be the case when the endpoint of the wormhole is relatively far from a base station[4].

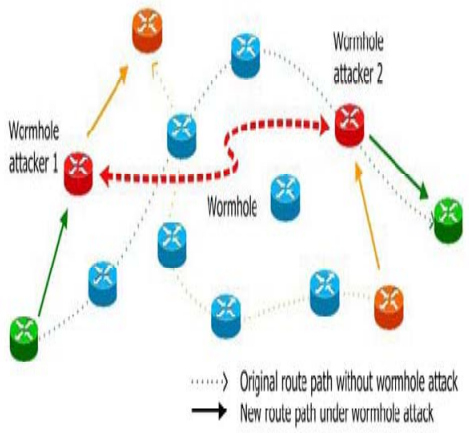


Fig. 2 : Illustration of wormhole attack in wireless network.

IV. PROPOSED SOLUTION

In this section I will explain in detail the proposed solution node authentication using the digital signature. In this algorithm the authentication is provided at each sensor node in the packet header which is forward from source to destination. Only the authenticate nodes can communicate in wireless sensor network. Using this authentication procedure we can detect the malicious nodes which causes the wormhole attack.

ALGORITHM

Source Node

If (Any Packet sent P)

{

Alter Header add columns =Route‘ and =Signatures‘

Insert id into Route Column

Insert Digital Signature into Signature Column

Forward Packet P

}

If (received A Packet)

{

If (Received Packet==Data_Ack)

{

Note the =Signature‘ in the header

Note Route Noted In header

Verify the Digital Signature

If(Verification Successful)

{

Discard the route noted

Else

{

Drop the packet

}

Repeat the procedure for next packet

}

}

Intermediate Node

If (Received a packet P)

{

Insert id into Route Column

Insert Digital Signature into Signature Column

Forward Packet P

}

Destination Node

If (Received a packet P)

{

Note the =Signature‘ in the header

Note Route Noted In header

Verify the Digital Signature

If(Verification Successful)

Note the =Signature‘ in the header

Note Route Noted In header

Verify the Digital Signature

If(Verification Successful)

{

Noted route=Null;

}

Else

{

Noted Route unchanged

}

Create Data_Ack Packet

Insert columns =Route‘ and =Signatures in Data_Ack

}

Insert id into Route Column

Insert Digital Signature into Signature Column

}

V. CONCLUSION

Security related issues in wireless sensor networks have become an important part of research in present scenario. To detecting malicious functions of node and offering efficient counter measure is the difficult task. In the proposed method there is no need for specific hardware and neither is the need for clock synchronization due to use of cryptographic concept digital signature. In that each node authenticate using digital signature. The received node at the destination node is verified and if the digital signature is false the information about that is sent to the sender node using DATA_ACK.

REFERENCES

[1] Guiyi Wei Xueli Wang "Detecting Wormhole Attacks Using Probabilistic Routing and Redundancy Transmission". WASE International Conference on Information Engineering 2010.PP-251-254.

[2] Junfeng Wu, HonglongChe, "Label-Based DV-Hop Localization AgainstWormhole Attacksin Wireless Sensor Networks".Fifth IEEE International Conference on Networking, Architecture, and Storage 2010. Pp-79-88.

[3] Zhibin Zhao, Bo Wei, Xiaomei Dong, Lan Yao, FuxiangGao"Detecting Wormhole Attacks in Wireless Sensor Networks with Statistical Analysis"First International Conference on Integrated Intelligent Computing.pp-283-289.

[4] Prasannajit B1, Venkatesh, Anupama S "An Approach towards Detection of Wormhole Attack in Sensor Networks" 2010 First International Conference on Integrated Intelligent Computing.

[5] Thorne, Kip S. (1994). Black Holes and Time Warps. W. W. Norton. p. 504.ISBN 03-23763.

[6] DeBenedictis, Andrew and Das, A. (2001). "On a General Class of Wormhole Geometries". Classical and Quantum Gravity 18 (7): 1187–1204.

[7] Forman G., Zahorjan J,"The challenges of mobile computing",*IEEE Computer* ; 27(4):38-47.

[8] Bayrem Triki, Slim Rekhis, Noureddine Boudriga ," Digital Investigation of Wormhole Attacks in Wireless Sensor Networks", Network Computing and Applications, IEEE International Symposium, July 2009, pp. 179-186 .

[9] Dezun Dong, Mo Li, Yunhao Liu, Xiangke Liao ," Connectivity-Based Wormhole Detection in Wireless Ad Hoc and Sensor Networks", Parallel and Distributed Systems, International Conference on , December 2009, pp. 72-79 .

[10] B. Prasannajit, Anupama S. Venkatesh, K. Vindhykumari, S.R. Subhashini, G. Vinitha ," An Approach Towards Detection of Wormhole Attack in Sensor Networks", Integrated Intelligent Computing , August 2010, pp. 283-289.

❑❑❑

Deployment Analysis in Underwater Acoustic Wireless Sensor Networks

[1]T.S.Yenganti & [2]Adane

[1]Dept. Of M.Tech(CSE), R.T.M.Nagpur University, Tulsiramaji Gaiwkad-Patil College Of Engineering and Tech, Nagpur

[2]Department of IT, R.T.M. Nagpur University, Ramdeobaba College of Engineering , Nagpur

E-mail : yengantiwar@rediffmail.com,

Abstract - This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an IEEE conference. The au In this paper, different deployment strategies for two-dimensional and three-dimensional communication architectures for Under Water Acoustic Sensor Networks (UW-ASNs) are proposed, and statistical deployment analysis for both Architectures is provided. The objectives of this paper are to determine the minimum number of sensors needed to be deployed to achieve the optimal sensing and communication coverage, which are dictated by the application, provide guidelines on how to choose the optimal deployment surface area, given a target region. study the robustness of the sensor network to node failures, and provide an estimate of the number of redundant sensors to be deployed to compensate for possible failures

Keywords— *Design, Performance, Reliability. Underwater Acoustic Sensor Networks, Deployment.*

I. INTRODUCTION

Underwater sensor networks are envisioned to enable applications for oceanographic data collection, ocean sampling, environmental and pollution monitoring, offshore exploration, disaster prevention, tsunami and seaquake warning, assisted navigation, distributed tactical surveillance, and mine reconnaissance. There is, in fact, significant interest in monitoring aquatic environments for scientific, environmental, commercial, safety, and military reasons. While there is a need for highly precise, real-time, fine grained spatio-temporal sampling of the ocean environment, current methods such as remote telemetry and sequential local sensing cannot satisfy many application needs, which call for wireless underwater acoustic networking. Under Water Acoustic Sensor Networks (UW-ASN) consist of sensors that are deployed to perform collaborative monitoring tasks over a given region. UW-ASN communication links are based on *acoustic wireless technology*, which poses unique challenges due to the harsh underwater environment, such as limited bandwidth capacity, high and variable propagation delays , high bit error rates, and temporary losses of connectivity caused by multipath and fading phenomena . We consider two communication architectures for UW-ASNs, i.e., the *two-dimensional architecture*, where sensors are anchored to the bottom of the ocean, and the *three-dimensional architecture*, where sensors float at different ocean depths covering the entire monitored volume region. While the former is designed for networks whose objective is to monitor the ocean bottom, the latter is more suitable to detect and observe phenomena that cannot be adequately observed by means of ocean bottom sensor nodes. We propose different deployment strategies, and provide a mathematical analysis to study deployment issues concerning both architectures, with the objectives below:

i) Determine the minimum number of sensors needed to be deployed to achieve the target sensing and communication coverage, which are dictated by the application;

ii) Provide guidelines on how to choose the optimal deployment surface area, given a target region;

iii) Study the robustness of the sensor network to node failures, and provide an estimate of the number of redundant sensors to be deployed to compensate for possible failures.

II. RELATED WORK

The problem of sensing and communication coverage for terrestrial sensor networks has been addresses in several papers. However, to the best of the authors' knowledge, this work is the first to study

deployment issues for underwater sensor networks. Many previous deployment solutions and theoretical bounds assuming spatio-temporal correlation, mobile sensors, redeployment of nodes, and particular deployment grid structures may not be feasible for the underwater environment.

In particular, methods for determining net-work connectivity and coverage given a node reliability model are discussed, and an estimate of the minimum required node-reliability for meeting a system-reliability objective is provided. An interesting result is that connectivity does not necessarily imply coverage. As the node reliability decreases', in fact, the sufficient condition for connectivity becomes weaker than the necessary condition for coverage.

Although provides useful theoretical bounds and insight into the deployment of wireless terrestrial sensor networks, the analysis is limited to grid structures. In two coordination sleep algorithms are compared, a random and a coordinated sleep scheme. It is shown that when the density of the network increases, the duty cycle of the network can be decreased for a fixed coverage. In sensor coverage is achieved by moving sensor nodes after an initial random deployment. However requires either mobile sensor nodes or redeployment of nodes, which may not be feasible for UW-ASNs. In sensing and communication coverage in a three-dimensional environment are rigorously investigated. The diameter, minimum and maximum degree of the reach ability graph that describes the network are derived as a function of the communication range, while different degrees of coverage (1- coverage and, more in general, k-coverage) for the 3D environment are characterized as a function of the sensing range. Interestingly, it is shown that the sensing range r required for 1-coverage is greater than the transmission range t that guarantees network connectivity. Since in typical applications $t \geq r$, the network is guaranteed to be connected when 1-coverage is achieved.

.

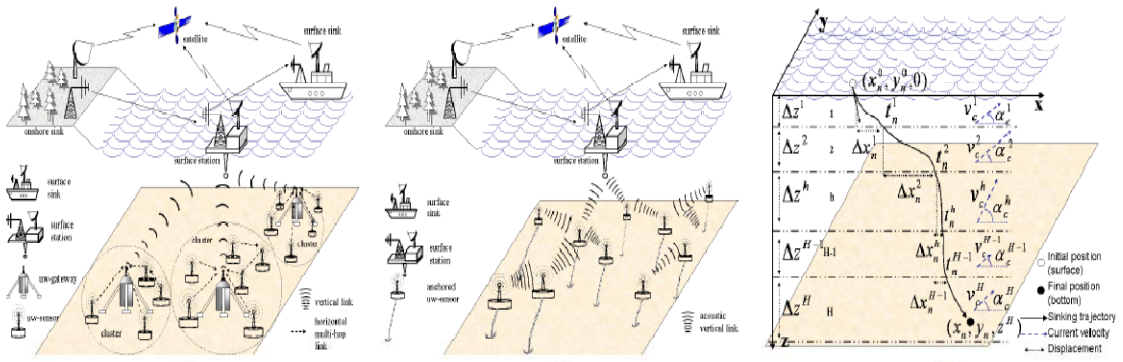


Figure 1: Architectures for two-dimensional (a) and three-dimensional (b) UW-ASNs; (c): Trajectory of a sinking object

.

Although these results were derived for terrestrial networks, they can also be applied in the underwater environment. Thus, in this paper, we will focus on the sensing coverage when discussing deployment issues in 3D UW-ASNS, as in three-dimensional networks it implicitly implies the communication coverage .

III. COMMUNICATION ARCHITECTURE

We consider two communication architectures for underwater sensor networks, i.e., a *two-dimensional* and a *three-dimensional architecture* , and identify the relevant deployment challenges. As in terrestrial sensor networks, in UW-ASNs it is necessary to provide *communication coverage*, i.e., all sensors should be able to establish multi-hop paths to the sink, and *sensing coverage*, i.e., the monitored area should be covered by the sensors. More formally, the *sensing range r* of a sensor is the radius of the sphere that models the region monitored by the sensor (sensing sphere). A portion $A\eta$ of the monitored region A is said to be *k-covered* if every point in $A\eta$ falls within the sensing sphere of at least k sensors. The *k-coverage ratio* ηk of a monitored region A is the fraction of the volume/ area that is k-covered by a 3D/2D UW-ASN, respectively. In the following, we will consider the case of $k = 1$ both for 2D and 3D networks to obtain simple *1-cover age* $\eta 1$ of the

region, since underwater sensors may be expensive devices and spatio-temporal correlation may not be assumed.

1) Two-dimensional UW-ASNs

A reference architecture for two-dimensional underwater sensor networks is shown in Fig. 1(a), where deployed sensor nodes are anchored to the bottom of the ocean. Underwater sensors may be organized in a cluster-based architecture, and be interconnected to one or more *underwater gateways* (uw-gateways) by means of wireless acoustic links. Uw-gateways are network devices in charge of relaying data from the ocean bottom network to a surface station. They are equipped with a long-range *vertical* transceiver, which is used to relay data to a *surface station*, and with a *horizontal* transceiver, which is used to communicate with the sensor nodes to send commands and configuration data, and to collect monitored data. The surface station is equipped with an acoustic transceiver, which may be able to handle multiple parallel communications with the uw-gateways, and with a long-range radio transmitter and/or satellite transmitter, which is needed to communicate with an *onshore sink* and/or to a *surface sink*.

2) Three-dimensional UW-ASNs

Three-dimensional underwater networks are used to detect and observe phenomena that cannot be adequately observed by means of ocean bottom uw-sensor nodes, i.e., to perform cooperative sampling of the 3D ocean environment. In this architecture, sensors float at different depths to observe a given phenomenon. One possible solution would be to attach each sensor node to a surface buoy, by means of wires whose length can be regulated to adjust the depth of each sensor node. However, although this solution enables easy and quick deployment of the sensor network, multiple floating buoys may obstruct ships navigating on the surface, or they can be easily detected and deactivated by enemies in military settings. Furthermore, floating buoys are vulnerable to weather and tampering or pilfering. A different approach is to anchor winch based sensor devices to the bottom of the ocean, as depicted in Fig.1(b). Each sensor is anchored to the ocean bottom and is equipped with a floating buoy that can be inflated by a pump. The buoy pulls the sensor towards the ocean surface. The depth of the sensor can then be regulated by adjusting the length of the wire that connects the sensor to the anchor, by means of an electronically controlled engine that resides on the sensor .

IV. DEPLOYMENT IN A 2D ENVIRONMENT

In this section, we provide a mathematical analysis of the graph properties of sensor devices that are deployed on the surface of the ocean, sink, and reach the ocean bottom. To achieve this, we study the trajectory of sinking devices (sensors and uw-gateways) when they are deployed on the ocean surface with known initial conditions (position and velocity). This allows us to capture both the case when sensor nodes are *randomly deployed* on the ocean surface, e.g., scattered from an airplane, or the case when sensors are *accurately positioned*, e.g., released from a vessel.

To address the deployment challenges presented in the previous section, in Section 4.1 we propose the *triangular-grid* deployment, and derive useful geometric properties. In Section 4.2, we study the dynamics of a sinking object and evaluate its trajectory under the presence of ocean currents. In Section 4.3, we characterize the different sinking behaviour of sensors and uw-gateways, with the objective of describing their average horizontal displacement and study the main communication properties of sensor clusters.

1) Triangular-grid Coverage Properties

In this section, we propose the *triangular-grid* deployment, and derive useful geometric properties. Let us consider the common case of sensors with same sensing range r. The optimal deployment strategy to cover a two-dimensional rectangular area using the minimum number of sensors is to centre each sensor at the vertex of a grid of equilateral triangles, as shown in Fig. 2(a). With this configuration, by adjusting the distance d among sensors, i.e., the side of the equilateral triangles, it is possible to achieve *full coverage*, i.e., $\eta = 1$. In addition, this enables to optimally control the coverage ratio η, defined as the ratio between the covered area and the target area. In particular, as it will be mathematically proven in the following, when $d = \sqrt{3}r$ the coverage ratio η is equal to1, i.e., the uncovered area ABC depicted in Figs. 2(a-b) is zero, and the overlapping areas are minimized. This allows to achieve the full coverage of a target area, but requires the highest number of sensors. Conversely, as the distance among sensors increases, i.e., the number of deployed sensors decreases, the coverage ratio decreases. Therefore, there is a trade-off between the number of deployed sensors and the achievable sensing coverage. We are interested in finding the minimum number of sensors that need to be deployed in order to guarantee a target sensing coverage η^*, which is dictated by the application requirements. To this end, we present the following theorem.

THEOREM 1. *In an equilateral grid the sensing coverage $\eta(d, r)$,i.e., the ratio of the covered area and the target area, is*

$$\eta(d,r) = \eta\left(\frac{d}{r}\right) = \begin{cases} \frac{\mathcal{A}_{DEF} - \mathcal{A}_{ABC}}{\mathcal{A}_{DEF}} = 1 - \frac{\mathcal{A}_{ABC}}{\frac{\sqrt{3}}{4}d^2} & \frac{d}{r} \in [0,2] \\ \frac{3 \cdot \frac{\pi r^2}{6}}{\frac{\sqrt{3}}{4}d^2} = \frac{2\pi}{\sqrt{3}} \cdot \left(\frac{d}{r}\right)^{-2} & \frac{d}{r} \in (2,\infty), \end{cases} \quad (1)$$

where:

$$\mathcal{A}_{ABC} = \frac{\sqrt{3}}{4}\left(\frac{d}{2} - \sqrt{3r^2 - \frac{3}{4}d^2}\right)^2 - 3r^2 \arcsin\frac{\overline{BC}}{2r} + \frac{3}{4}\overline{BC}\sqrt{4r^2 - \overline{BC}^2}, \quad \overline{BC} = \frac{d}{2} - \sqrt{3r^2 - \frac{3}{4}d^2}. \quad (2)$$

PROOF. With reference to Fig. 2(b), which represents a zoomed portion of Fig. 2(a), $AE = r$ and $EH = d/2$, where r is the sensing range and d is the distance between sensors. Since the triangle DEF is equilateral by construction, $HO = (\sqrt{3}/6)d$. Consequently, since $AH = \sqrt{r^2 - d^2/4}$, it holds $AO = HO - AH = (\sqrt{3}/6)d - \sqrt{r^2 - d2/4}$. As triangle DEF is equilateral, triangle ABC is equilateral too. Since $AO = (\sqrt{3}/3)BC$, then $BC = d/2 - \sqrt{3r^2 - (3/4)d^2}$. Therefore, the area of triangle ABC is *A (delta) of ABC* $= (\sqrt{3}/4)BC^2$ In order to express the sensing coverage $\eta(d, r)$ as a function of d and r, we need to compute the area *delta ABC* of the *uncovered region ABC* among the circles with centres in D, E, and F, and radius r. This can be computed as *A ABC = A(delta)ABC − 3.A BTCK*, where *A BTCK* coincides with the difference of the areas of the circular sector *BTCF i.e A BTCK = A BTCF- A(delta) BCF.*

$r^2 \arcsin(\overline{BC}/2r) - (\overline{BC}/4)\sqrt{4r^2 - \overline{BC}^2}$.

Consequently, $\mathcal{A}_{ABC} = (\sqrt{3}/4)\left(d/2 - \sqrt{3r^2 - (3/4)d^2}\right)^2 - 3r^2 \arcsin(\overline{BC}/2r) + (3/4)\overline{BC}\sqrt{4r^2 - \overline{BC}^2}$, where $\overline{BC} = d/2 - \sqrt{3r^2 - (3/4)d^2}$, which gives (1) in the non-trivial case $d/r \in [0,2]$. As far as the case $d/r \subset (2,\infty)$ is concerned, no overlapping areas are formed, and the coverage η can be computed

2) Trajectory of a Sinking Object

In this section, we study the dynamics of a sinking object and evaluate its trajectory under the presence of ocean currents. In particular, we first consider the ideal case in which the velocity of the ocean current does not change with depth; then, we extend the model to capture the more realistic case in which the velocity of thecurrent depends on depth. According to Newton's first law of motion, the acceleration **a** describing the sinking in the water of an object with a density ρ and volume V is determined by the following vectorial motion law,

$$\vec{\mathbf{F}}_{\mathbf{W}} + \vec{\mathbf{F}}_{\mathbf{B}} + \vec{\mathbf{F}}_{\mathbf{R}} + \vec{\mathbf{F}}_{\mathbf{C}} = \rho V \cdot \vec{\mathbf{a}}, \quad (4)$$

where:

- $\vec{\mathbf{F}}_{\mathbf{W}} = \rho V \cdot \vec{\mathbf{g}}$ is the *weight force*, which depends on the density ρ [Kg/m^3] and volume V [m^3] of the sinking object, and on the terrestrial gravitational acceleration $g = 9.81\,\mathrm{m/s^2}$;

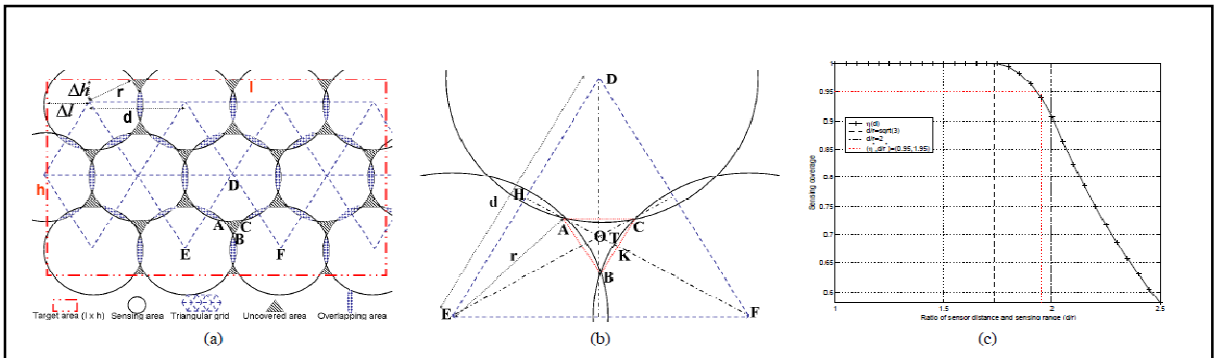


Figure 2: Triangular-grid deployment. (a): Grid structure and side margins; (b): Uncovered area; (c): Sensing coverage

We project (4) onto the x-, y-, and z- axes, which are directed as shown in Fig. 1(c), and we denote the dynamic position of the sinking object as $\mathbf{P} = (x, y, z)$, its velocity as $\dot{\mathbf{v}} = (\dot{x}, \dot{y}, \dot{z})$, and its acceleration as $\mathbf{a} = (\ddot{x}, \ddot{y}, \ddot{z})$. We then consider the velocity of the current $\mathbf{Vc} = (V^x c, V^y c, V^z c)$, which, for the sake of clarity, is first assumed to be independent on the ocean depth (we will then relax this assumption). Under the assumption that no significant vertical movement of ocean water is observed, i.e., the considered area is neither an *upwelling* nor a *downwelling area*, the current along the z-axes can be neglected ($V^z c \approx 0$), and (4) leads to three scalar laws,

$$x: F_C^x = \rho V \ddot{x}; \quad y: F_C^y = \rho V \ddot{y}; \quad z: F_W^z + F_B^z + F_R^z = \rho V \ddot{z}. \tag{5}$$

Specifically, we obtain the following dynamic system equations,

$$\begin{cases} \ddot{x} + \frac{C\sigma A^{xy}}{\rho V}\dot{x} = \frac{C\sigma A^{xy}}{\rho V} v_c^x \\ \ddot{y} + \frac{C\sigma A^{xy}}{\rho V}\dot{y} = \frac{C\sigma A^{xy}}{\rho V} v_c^y \\ \ddot{z} + \frac{K\mu\rho_w A^z}{\rho V}\dot{z} = g\frac{\rho-\rho_w}{\rho}, \end{cases} \tag{6}$$

where *Axy* and *Az* represent the horizontal and vertical cross-sections, respectively. By solving this dynamic system, with the initial conditions of the object on the surface at time t^0 i.e its Position $\mathbf{P}(t0) = (x(t0), y(t0), 0)$ and velocity $\mathbf{v}(t0) = (x(t0), \dot{y}(t0), \dot{z}(t0))$, we obtain the solution.

$$\begin{cases} x(t) = x(t^0) + v_c^x \cdot (t - t^0) + \frac{\dot{x}(t^0) - v_c^x}{C\sigma A^{xy}/\rho V} \cdot [1 - e^{-\frac{C\sigma A^{xy}}{\rho V} \cdot (t - t^0)}] \\ y(t) = y(t^0) + v_c^y \cdot (t - t^0) + \frac{\dot{y}(t^0) - v_c^y}{C\sigma A^{xy}/\rho V} \cdot [1 - e^{-\frac{C\sigma A^{xy}}{\rho V} \cdot (t - t^0)}] \\ z(t) = v_\infty^z \cdot (t - t^0) + [\dot{z}(t^0) - v_\infty^z] \cdot [1 - e^{-\frac{K\rho_w \mu A^z}{\rho V} \cdot (t - t^0)}], \end{cases} \tag{7}$$

where we denoted as $v_\infty^z = \frac{gV(\rho-\rho_w)}{K\rho_w \mu A^z}$ [m/s] the *terminal velocity* along z, which is computed by imposing in (5) the following force equilibrium, $F_W^z + F_B^z + F_R^z = 0$, i.e., $\ddot{z} = 0$ in (6).

V. DEPLOYMENT IN A 3D ENVIRONMENT

In this section, we propose three deployment strategies for three dimensional UW-ASNs to obtain a target *1-coverage* $\eta^*1 = \eta^*$ of the 3D region, i.e., the *3D-random*, the *bottom-random*, and the *bottom-grid* strategies. the sensing range *r* required for 1-coverage is greater than the transmission range *t* that guarantees network connectivity. Since in typical applications $t \geq r$, the network is guaranteed to be connected when 1-coverage is guaranteed. Thus, in the following we focus on the sensing coverage. In all these deployment strategies, winch-based sensor devices are anchored to the bottom of the ocean in such a way that they cannot drift with currents. Sensor devices are equipped with a floating buoy that can be inflated by a pump by means of an electronically controlled engine that resides on the sensor. This way, they can adjust their depth and float at different depths in order to observe a given phenomenon, as described in Section 3.2. In all the proposed deployment strategies, described hereafter, sensors are assumed to know their final positions by exploiting localization techniques.

3D-random. This is the simplest deployment strategy, and does not require any form of coordination from the surface station. Sensors are randomly deployed on the bottom of the 3D volume, where they are anchored. Then, each sensor randomly chooses its depth, and, by adjusting the length of the wire that connects it to the anchor, it floats to the selected depth. Finally, each sensor informs the surface station about its final position.

Bottom-random. As in the previous strategy, sensors are randomly deployed on the bottom, where they are anchored. Differently from the 3D-random scheme, the surface station is informed about their position on the bottom. Then, the surface station calculates the depth for each sensor in order to achieve the target 1- coverage ratio η^*. Finally, each sensor is assigned its target depth and floats to the desired position.

Bottom-grid. This deployment strategy needs to be assisted by one or multiple AUVs, which deploy the underwater sensors to predefined target locations to obtain a grid deployment on the bottom of the ocean. Each sensor is also assigned a desired depth by the AUV and accordingly floats to achieve the target coverage ratio η^*. As shown in Figs. 6(a-c), given a fixed number of sensors we achieve a better coverage ratio with increasing complexity of the deployment strategy.

.

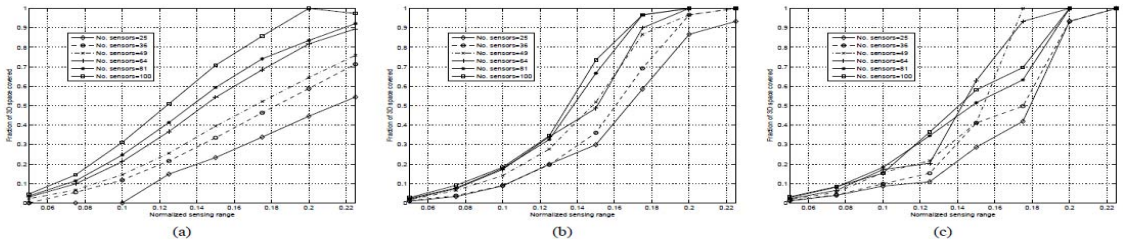


Figure 6: Three-dimensional scenario. (a): 3D coverage with a 3D random deployment; (b): Optimized 3D coverage with a 2D bottom-random deployment; (c): Optimized 3D coverage with a 2D bottom-grid deployment

.

In fact, the coverage ratio obtained with the bottom-grid strategy is greater than the coverage ratio obtained with the bottom-random strategy, which is in turn greater than the coverage ratio of the 3D-random strategy. Moreover, given a target coverage ratio, the minimum number of sensors needed to achieve the desired coverage ratio decreases with the complexity of the deployment strategy.

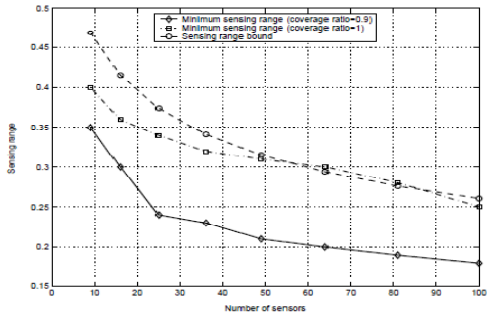


Figure 7: Theoretical and experimental sensing range

Figure 7 shows a comparison between the minimum normalized sensing range that guarantees coverage ratios of 1 and 0.9 with the bottom-random strategy and the theoretical bound on the minimum normalized sensing range derived in [4], where the authors investigate sensing and communication coverage in a 3D environment. According to Theorem 4 , the 3D volume is guaranteed to be *asymptotically almost surely* 1-covered iff $4\ 3\pi\ n\ V\ r3 = \ln n+\ln\ln n+\omega(n)$, with $1 << \omega(n) << \ln \ln n$,where V is the volume of the region to be covered, n the number of deployed sensors, and r their sensing range. Hence, to draw Fig.7 we set $\omega(n) = 1+\ln \ln n\ 2$. This shows that the bottom-random deployment strategy very closely approximates the theoretically predicted bound, i.e., the minimum sensing range that guarantees 1-coverage with probability 1 is almost the same as that predicted by the model.

VI. CONCLUSIONS

In this paper, deployment strategies for two-dimensional and three dimensional architectures for underwater sensor networks were proposed, and deployment analysis was provided. The objectives were to determine the minimum number of sensors to be deployed to achieve the application-dependent target sensing and communication coverage; provide guidelines on how to choose the deployment surface area, given a target region; study the robustness of the sensor network to node failures, and provide an estimate of the number of required redundant sensors.

VII. REFERENCES

[1] I. F. Akyildiz, D. Pompili, and T. Melodia. Underwater Acoustic Sensor Networks: Research Challenges. Ad Hoc Networks (Elsevier), 3(3):257–279, May 2006.

[2] C. Hsin and M. Liu. Network Coverage using Low Duty Cycled Sensors: Random and Coordinated Sleep Algorithms. In Proc. Of IEEE/ACM IPSN, pages 433–442, Berkeley, California, USA, Apr.2004.

[3] J. Proakis, E. Sozer, J. Rice, and M. Stojanovic. Shallow Water Acoustic Networks. IEEE Communications Magazine, pages 114–119, Nov. 2001.

[4] V. Ravelomanana. Extremal Properties of Three-dimensional Sensor Networks with Applications. IEEE Transactions on Mobile Computing, 3(3):246–257, July/Sept. 2004.

[5] R. A. Serway and J. W. Jewett. Physics for Scientists and Engineers. Brooks/Cole, 2004.

[6] S. Shakkottai, R. Srikant, and N. Shroff. Unreliable Sensor Grids:Coverage, Connectivity, and Diameter. In Proc. of IEEE INFOCOM, volume 2, pages 1073–1083, San Francisco, CA, USA, Apr. 2003.

[7] E. Sozer, M. Stojanovic, and J. Proakis. Underwater Acoustic Networks. IEEE Journal of Oceanic Engineering, 25(1):72–83, Jan.

❑❑❑

A Spatial Multiplex based Auto-stereoscopic Display Approach for 3D Visualization

Dilip Kumar Dalei, B. V. Hari Krishna Nanda & N. Venkataramanan

ANURAG, Defence R & D Organization (DRDO), Kanchanbagh, Hyderabad, 500058, India.
E-mail: dilipkumar@anurag.drdo.in, bvhk_nanda@anurag.drdo.in, n_venkataramanan@anurag.drdo.in

Abstract **-** Scientific Data Visualization is a fast growing research field which is used in a vast range of applications like CFD Analysis, Simulation, Medicine, Education and Engineering. The visual output of visualization process can be made more realistic and appealing by using advanced display technologies. Auto-stereoscopy is a kind of 3D display technology which provides a seamless three dimensional experience without any viewing aid like head gear or shutter glasses. In this paper we have presented the analysis of 3D visualization system. We have also explained the implementation details of an experimental 3D visualization system using Auto-stereo approach. The implementation is done in C++ using OpenGL graphics library and CUDA Library.

Keywords: *Auto-stereoscopic display, 3D Display, Spatial Multiplexing, 3D Visualization, CUDA, OpenGL*

I. INTRODUCTION

Scientific Data Visualization can be thought of as a computational process that transforms the data to visual objects enabling researchers to comprehend their simulation, computations and measurements. The field of visualization mainly focuses on creating visual images that convey salient information about data, relationships and underlying processes. These visuals are finally rendered on conventional 2D display for visual analysis. The use of 3D display technology further adds realism to visuals of a visualization system.

Auto-stereoscopic display is a next generation 3D technology that introduces the ability to watch 3D effects without any viewing aid. Moreover, current auto-stereoscopic methods can display much more than two images to provide an adequate rendering in several directions and to be adapted for multi-user purposes [1]. This paper provides an overview of 3D Visualization by exploring state-of-the art visualization method namely iso-surface generation for volume data. Then it focuses on the implementation of an experimental 3D visualization system using Auto-stereoscopic approach.

The remainder of the paper is organized as follows. Section II explains the basic foundation of Auto-stereoscopic display systems and their guiding principles. This is followed by explanation of 3D Visualization system and its implementation in Section III. The results are analyzed and discussed in Section IV. Finally, the paper is concluded in section V.

II. AUTO-STEREOSCOPY

A. Methodology

A user's eye sees an infinite number of different images in a scene. The whole viewing space of user can be theoretically divided into finite no of viewing zone. Each zone displays one image or view of the scene. But the user still sees different images for each eye and the image change upon movement of head. This preserves both stereo and horizontal movement parallax cues necessary for 3D perception. In auto-stereoscopy, the scene is rendered from multiple view points and the results are collected in multiple intermediate images. These images are fused into a single 3D spatial image conforming to the auto-stereoscopic display. To perceive 3D, the final image is accordingly projected into different viewing zones using optical filters (Parallax Barrier/Lenticular sheet).

B. Display systems

The current available technologies for auto-stereoscopic displays can be broadly classified into three categories. These are Spatial Multiplex, Multi-projector and Time Multiplexing. The paper focuses the study on spatial multiplex based auto-stereoscopic display designs [2].

Spatial-multiplexing is an auto-stereo approach which involves image interleaving. In Image interleaving, the source image is split into strips and these strips are merged into a single image.

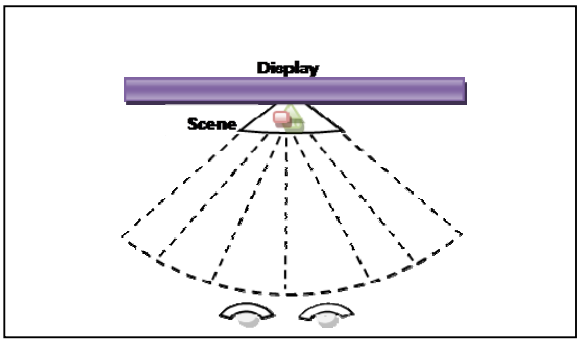


Fig 1: Multi-view Auto-stereoscopic

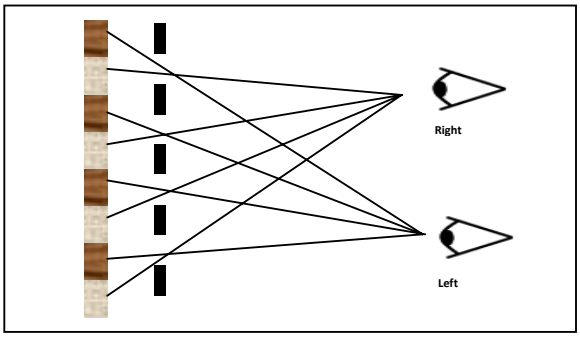


Fig 2: Parallax Barrier Display

A significant issue in spatial multiplex is the loss of resolution of source image. The spatially-multiplexed category of displays is further classified by the mechanism employed to re-direct interleaved images toward the eyes [5]. These are *Parallax Barrier* and *Lenticular* display systems.

- *Parallax Barrier System*

This is probably the oldest auto-stereoscopic technique that preserves horizontal parallax. In this approach, a sheet of alternate opaque/transparent columns is mounted on top of the display. The vertical columns reveal different parts of the underlying image depending upon the viewing direction.

The left and right eye would see different sets of columns as shown in figure 2. Thus, by encoding the left eye view on one set of columns and the right eye view on the other set, the user feels the depth perception in the auto-stereoscopic image created a certain distance from the display. One major disadvantage of this system is low image brightness because some portion of light is blocked by opaque columns. The examples of Parallax Barrier displays are 4D-Vision [11], the Varrier [12], SuperD HDB 24 etc.

- *Lenticular Lenslets System*

The other popular method for auto-stereoscopic displays uses a lenticular array in place of the parallax barrier. It consists a sheet of long cylindrical lenses (lenticulars) placed over a flat display in such a way that the display's image plane coincides with the focal plane of the lenses as shown in the figure 3. This helps the lenses to focus different portion of each image toward the user's eye.

Unlike barrier technology, this system has no loss of brightness due to the ability of lenses to collect all the light from pixels. But it still suffers from the same resolution and viewing zone problem as parallax barrier systems. A relatively low number of views have the advantage of a low resolution loss, but the disadvantage of limited look-around ability. Systems like SynthaGram [8], Phillips 3D-LCD [13], SuperD HDL 24 are examples of lenticular displays. We use both Barrier and Lenticular 3D monitors from SuperD in our work.

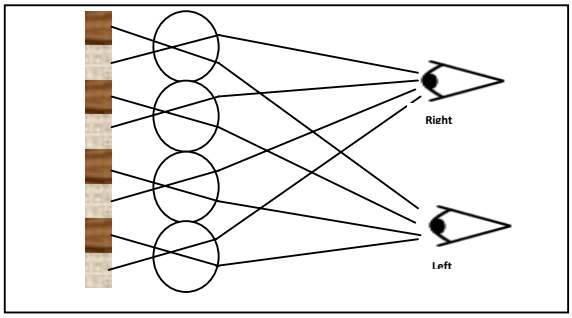


Fig 3: Lenticular Lenlets Display

III. 3D VISUALIZATION

A. *Overview*

3D visualization has received a lot of attention in past few years due to rapid development in technologies like 3D display, Interaction devices, virtual reality etc. The Visualization process basically converts numerical data into convenient and efficient images. The visuals are then rendered on the conventional 2D display systems for human visual analysis. But these 2D displays lack the ability of immersive and realistic 3D experience which is quite natural to human beings.

The presence of 3D environment helps in achieving effective visualization. It provides a kind of virtual reality environment for visualizing phenomena which involves delicate and detail structures of process. A 3D visualization system employs a variety of technologies to create real-life like 3D visual experiences. One such technology that is widely adopted for creating 3D environment is based upon stereoscopic approach. It takes the help of advanced equipments like active/passive stereoscopic display systems, specialized eye wares, head mounted displays. But the stereoscopic systems requires the user to wear external devices like

eye glasses, so it lacks the appearance and feeling the natural visual experience as we move and experience in real surroundings. To overcome this limitation, another display technology is evolved based on auto-stereoscopic approach. The auto-stereoscopic system presents a real 3D space like environment without any need to wear any devices.

B. *Iso-Surface Extraction*

A typical visualization system provides wide range of visualization services like scalar visualization (contour, iso-surface generation, height map, color map), Vector visualization (streamline, stream surface, and 3D glyph), volume rendering etc. In our work we have considered only one visualization technique i.e. iso-surface extraction for volume data.

Iso-surface Extraction is one of the indirect volume visualization technique based on polygonal representation of surface extracted from 3D data. It can be defined as generation of an approximate, piecewise iso-surface (usually composed of a collection of triangles) on a sampled scalar field [15]. For our work on 3D visualization we have considered Marching Cube (MC) based Iso-surface Generation. Marching Cube is a standard sequential-traversal algorithm for extracting surfaces from 3D datasets [15].

The traditional MC method divides the volume grid into small units called cubes and traverses all the cubes sequentially in search of surface polygons. So, this is quite time consuming in case of large size data. The MC is a completely data-parallel algorithm which makes it ideal for execution on Graphics Processing Units (GPU). A lot of GPU-based implementations have been proposed for accelerating the MC. We have used one such method known as Stream Compaction and Expansion. This method is used in the iso-surface implementation provided in NVIDIA's CUDA SDK. This iso-surface module forms the base code for our system implementation. We have re-factored the module for auto-stereoscopic rendering on 3D display systems. The whole implementation is done in C++ using CUDA and OpenGL Library.

C. *Auto-Stereo Implementation*

Many algorithms are proposed to create auto-stereoscopic images on a parallax display. We mainly adopt the Multi-pass/Deferred approach to manipulate the image pixels to create 3D perception. The parallax type of displays generally requires a single multiplexed 3D image constructed from multiple images taken from different viewpoints of the scene.

The whole process of 3D visualization of iso-surface can be divided into two phases, iso-surface computation and rendering. In the computation phase the polygons representing the surface are searched and generated. This involves the process of voxel classification and generation of triangles per voxel. This phase is designed using stream compaction and expansion method suitable for parallel execution on GPU. The implementation is done using NVIDIA's CUDA library. CUDA (Compute Unified Device Architecture) is NVIDIA's innovative solutions for GPU based parallel computing[19].

In the Iso-surface rendering phase we adopt the auto-stereo approach to generate the 3D image. In this approach the scene containing the surface polygons are rendered N times from N different viewpoints and the corresponding N images are collected and processed to form a composited image. As per our 3D monitor's specification, nine views (N = 9) of the scene are captured in our experiment. The images are arranged in a special nine-tiled matrix format as shown in figure 4.

Fig 4: Nine-tile Format: Nine views of surface model arranged in a 3 X 3 Matrix

This image matrix is finally passed through a GPU based composition code which fuses the nine images into a single image suitable for auto-stereoscopic rendering. The composition code is implemented in GLSL [21] which is a OpenGL compliant shading language. The whole rendering part is implemented using OpenGL API[20]. OpenGL is widely used as the industry standard 3D graphics library. It provides a pipeline approach to render a 3D scene.

IV. RESULTS AND DISCUSSIONS

The hardware specification of our Testbed system is Intel Core2 duo with two Nvidia Quadro Fx 5800 graphics card under Linux System. Two HDL/B 24'' 3D monitors from SuperD Corporation are used in the experiment.

We have taken volume datasets available in public domains as our test input [16][17]18]. These are MRI-Head data, Fuel injection data and Bucky Ball data. The details of the dataset are given in the table 1.

Table 1: Volume Dataset

Dataset	Grid Resolution	Data Information
Baby Head	256 x 256 x 98	CT Scan.
Fuel Injection	64 x 64 x 64	Simulation of fuel injection into a combustion chamber.
Bucky Ball	32 x 32 x 32	Simulation of the electron density of a Buckminster-Fullerene.

An iso-surface is extracted from datasets and rendered on 3D display monitors. The final multiplexed auto-stereo image of different surfaces is given in Fig 5 -7. The visuals give an illusion of being in space in front 3D monitors. The usual interaction with the visual object creates the feeling of 3D immersion.

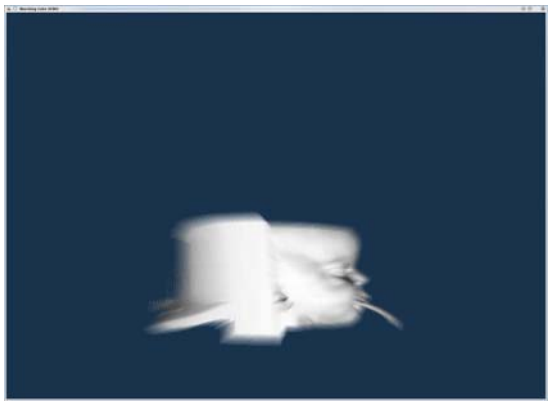


Fig 5: 3D image of Iso-surface of MRI Head

Fig 6: 3D Image of Iso-surface of Fuel Injection Data

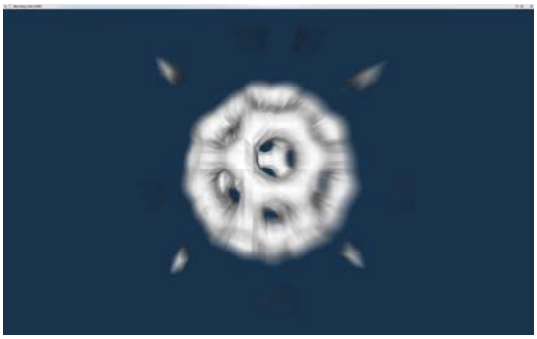


Fig 7: 3D Image of Iso-surface of Bucky Data

V. CONCLUSION

In our experiment we have considered only one visualization technique i.e. iso-surface generation for our experimental 3D visualization system. We observed that incorporating auto-stereo approach in visualization increases the appealing of visual analysis. The feeling of immersion and interaction seems very promising and natural.

In future, we plan to extend the present system to a full-fledged visualization system having all other visualization features like volume rendering, flow visualization etc. Further study can be done on integration of auto-stereoscopic technology in a virtual reality environment.

ACKNOWLEDGEMENT

We thank our team members for constant support and encouragement. We would like to extend our courtesy to all sources [16][17][18] for distributing volume datasets for research purpose.

REFERENCES

[1] D. K. Dalei, Kuldeep Goyal, N. Venkataramanan, " Implementation Issues in Multi-View Rendering on Spatial Multiplex based 3D Display System", ICACCT 2011, Panipat.

[2] N. A. Dodgson, "Auto-stereoscopic 3d displays", Computer, vol. 38, no. 8, pp. 31-36, 2005.

[3] John R. Moore, Neil A. Dodgson, Adrian R. L. Travis, Stewart R. Lang, "Time-multiplexed color auto-stereoscopic display", SPIE Symposium on Stereoscopic Displays and Applications VII, Jan 28–Feb 2, 1996

[4] F. de Sorbier, V. Nozick and V. Biri, "GPU rendering for auto-stereoscopic displays", 4th International Symposium on 3D Data Processing,

Visualization and Transmission(3DPVT'08), June 2008.

[5] Robert L. Kooima, Tom Peterka, Javier I. Girado, Jinghua Ge, Daniel J. Sandin, Thomas A. DeFanti, "A GPU Sub-pixel Algorithm for Auto-stereoscopic Virtual Reality".

[6] Nick Holliman, "3D Display Systems", February 2, 2005

[7] E. Lynn Usery, "Auto-stereoscopy – Three-Dimensional Visualization Solution or Myth?"[

[8] L. Lipton, M. Feldman. A New Auto-stereoscopic Display Technology: The SynthaGram. In Proceedings of SPIE Photonics West 2002: Electronic Imaging, San Jose, California, 2002.

[9] Ian Sexton, "PARALLAX BARRIER DISPLAY SYSTEMS".

[10] Michael Halle, "Auto-stereoscopic displays and computer graphics"

[11] A. Schmidt, A. Grasnick. Multi-viewpoint Auto-stereoscopic Displays from 4D-Vision. In Proceedings of SPIE Vol. 4660, 2002.

[12] D. Sandin, T. Margolis, J. Ge, J. Girado, T. Peterka, T. Defanti. The Varrier™ Autostereoscopic Virtual Reality Display. In ACM Transactions on Graphics, Proceedings of ACM SIGGRAPH, 24, no.3, 2005, pp. 894-903

[13] C. van Berkel, "Image Preparation for the 3D-LCD". In Proceedings of SPIE, Stereoscopic Displays and Virtual Reality Systems. 1999.

[14] Lorensen W E, Cline H E. Marching cubes: A high resolution 3D surface construction algorithm. ACM SIGGRAPH Computer Graphics. 1987;21(4)

[15] Timothy S. Newman, Hong Yi. A survey of the marching cubes algorithm, pp 854-879, computers & Graphics.

[16] Baby-Head Data: Jason Bryan, VolSuite Package, http://www.osc.edu/archve/VolSuite

[17] Fuel Injection Data : SFB 382 of the German Research Council (DFG)

[18] Bucky Ball data: AVS, USA, http://www.avs.com.

[19] NVIDIA Corporation, http://www.nvidia.com.

[20] OpenGL 3D Graphic Library, www.opengl.org.

[21] OpenGL shading Language(GLSL), www.opengl.org.

❑❑❑

An Approach To Hiding Image Into Video Using Magic Square

Bijoly Saha[1] & Sudipta Bhattacharya[2]

[1]Department of Information Technology, Techno India, EM-4/1, Sector-V, Salt Lake,Kolkata-91, W.B, India
[2]Department of CSE/IT, Bengal Institute Of Technology, Tech Town,Kolkata-150, W.B. India
E-mail:-saha.b8@gmail.com , bhattacharya.sudipta@gmail.com

Abstract **-** It is simple to inspect suspicious objects and extract hidden messages by comparing them to the original versions, the restricted portability and accessibility of original cover-signals generally make blind steganalysis more attractive and feasible in many practical applications. The new techniques which are based on magic square are presented in this paper. The proposed method hides a digital data using magic square into a frame of a video in a block-by-block fashion by using the least significant bit replacement. It is one of the proposed techniques extends the conventional spatial domain lsb steganography.It is a very simple technique for embedding and decoding the hidden data. The Experimental results show that the video produced by the proposed method has a good visual quality.

Keywords:*Steganography;LSBReplacement;Magic-square;Encryption*

I. INTRODUCTION

In general, information hiding (also called data hiding or data embedding) include steganography. The art and science of hiding information with a secret meaning inside other seemingly innocuous media is known as steganography. The primary goal of steganography is to set up a covert communication channel in a completely undetectable manner. This implies that the warder should be capable of discriminating suspicious objects from a large number of innocuous ones (i.e., the so-called passive steganalysis). [7].Attempting to detect and de cipher these messages is called steganalysis. Steganography also can be characterized as a specific form of covert channel. Steganalysis involves detecting whether steganography is used and being able to extract the hidden message. [5]

The hiding process has to be such that the modifications of the media are imperceptible. For images, this means that the modifications of the pixel values have to be invisible. For video also, this means that the modifications of the pixel values have to be invisible. Stereography is the act of adding a hidden message to an image or other media file. It is similar to encrypting a document, but instead of running it through a cipher, the document is broken up and stored in unused, or unnoticeable, bits within a frame of selected video [3].

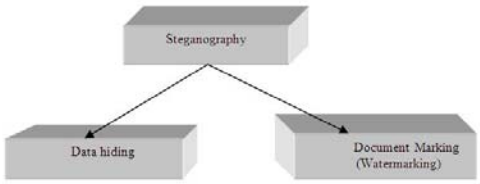


Figure 1 Represents Types of steganography

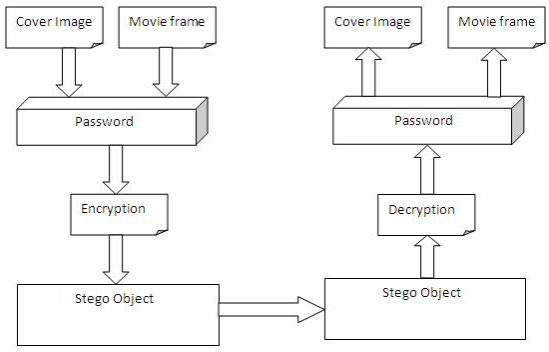


Figure2: Represents Steganography

In cryptography, encryption is the process of transforming information using an algorithm to make it unreadable to anyone except those possessing special knowledge, usually referred to as a key. Cryptography is

used in applications present in technologically advanced society; examples include the security of ATM cards, computer passwords, electronic commerce which all depend on cryptography.[2]

While cryptography is about protecting the content of messages, steganography is about concealing their very existence. In recreational mathematics, a magic square of order n is an arrangement of n^2 numbers, usually distinct integers, in a square, such that the n numbers in all rows, all columns and both diagonals sum to same constant. A normal magic square contains the integers from 1 to n^2 .The term "magic square" is also sometimes used to refer any of various types of word square. Normal magic squares exist all orders $n \geq 1$ except n=2, although the case n=1 is trivial – it consists of a single cell containing the number 1. The smallest non-trivial case shown below is order of 3.[4]

Maintaining the Integrity of the Specifications

16	3	2	13
6	10	11	8
9	6	7	12
4	15	14	1

Figure3:- Represents Magic Square The constant sum in every row , column and diagonal is called the magic constant or magic sum, M .The magic constant of a normal magic square depends only on n and has the value $M(n) = (n^3 + n) / 2$.[1]

This paper proposes a hiding scheme based on a magic square for the purpose of video authentication. This algorithm depending on the spatial domain. In spatial domain technique, the stego data is embedded in the source video by selecting pixel positions and replacing bit. The spatial domain techniques are easy to perceive and it has low time complexity that does not exists in other domains. Here, we are taking some fixed no. of frames from host video and embed an image into selected frame by using magic square and by replacing bit. In this method confidentiality of message means security, integrity of messages, sender authentication, non-repudiation of messages, and many other related issues are maintained.[6]

II.PROPOSED METHOD

In this paper, we presented an algorithm for embedding a image into a video frame sequences. The basic principle is based on spatial domain. In the ***spatial domain*** is the normal image space,

A. Movie Encoding Algorithm:

Here we will discuss how to encode an image in a video frame as cover image using a key i.e. password. This algorithm consist of four main steps.

STEP1:-This algorithm mainly uses single video frame which is selected randomly and that frame format will be .jpg image file formats as a cover image. The minimum size of the .jpg file must be at least 4x4 After getting the input of the selected .jpg cover image file. our next approach will be Generate a matrix containing the pixel values of the selected frame act as a cover image ,I. Now , we will break up the 2-D array i.e. the matrix I into a cell array of adjacent of sub matrices of I ,and store it as Magic Block, such that minimum number of sub matrices have the size of 4X4 . After getting the image which will be hidden and selected frame act as a Cover image 1st we have to check whether the selected stego image can be encoded in the frame(cover image) or not. That means we have to check the image size first, then convert pixel values into a message length.Calculate maximum message length, 3*(row_max* col_max -1)/8 .here row_max and col_max means the row and column size of the image. We are going to store the message length into the 1st pixel of the image. hence we are not considering it for message stream bits encoding, So we will have (row_max*col_max-1)pixels and as we are considering 24 bitmap image hence it will have 3*(row_max*col_max-1) number of maximum message stream bits to be stored. As each byte of message. i.e, character consists of 8 bits. We can store maximum 3*(row_max*col_max-1)/8 no. of characters in the cover image. Next, we will get the message length of the input message stream n and check n ≤ maximum message length. Next part is normalizing the message length and dividing it into 3 parts so that it can be stored in R,G and B components of the 1st pixel. the following process consists of, Divide the message length , n by 10 and store the remainder r = n%10 ;Get the dividend after dividing the n by 10. d= (n-r)/10.Now, square root the d and store the round result r_1 = round ($\sqrt{d}$).Now, as we have rounded the square root and it may not give the accurate d when calculate the square of r1 .To calculate the error, 1st get $d_1 = r_1^2$.Then find the difference , $s = d_1 - d$;If, (s<0) we will increase the value of r_1 to 1.$r_1 = r_1 + 1$;Now ,repeat the process "4" to get new d1 and process "5" to calculate the new $s = d_1 - d$; Here our main focus will be encrypting the message stream and encode it into the image in the following process, It will start from the second pixel of the first pixel row as 1st pixel is used to store the message length First of all we are using here the created Magic_ Block in the STEP- 2 .i.e. the sub matrices of the 2D array of 1st pixel values I, The R component of 1st sub matrices will be firstly used. As the 1st pixel is used to store the message length which is 1st pixel of the 1st sub matrices. Hence, we will start

from the second pixel of the 1st sub matrices. And the R components of the 1 st sub matrices will be 1st used after that the G component and after that the B component. Now we are checking the size of the of 1st sub matrices, whether it is 4X4 or not. If it is 4X4 then it can be matched with the magic square. now, Take a byte from the message stream and a byte from the password stream.XOR the message bit with the password stream stego key i.e. LSB of the byte of the password stream. Find out the position of current pixel. Now this step consist of If the current sub matrices has the same size as magic square of 4 we will obtain the indices of the value of temporary flag which will iterate form 1 to 16 that is the values contains in the magic square. When the flag has the value of 1 that means, it is at the position of (4,4).And now we will match the same position with the current sub matrices and obtain the pixel value. The flag will iterate until it reaches the end of magic square i.e. that is the value of 16 which will be at the position (1, 1).After it the flag will be set to 1 again, And it will iterate in the same process .If current sub matrices do not match the same size of magic square of 4 it will fetch the pixel position simply without matching it to any magic square. Next step is to XOR the encrypted message bit with the LSB of the current pixel (obtained earlier step).and set the new value in the corresponding pixel. Repeat the process in a loop .

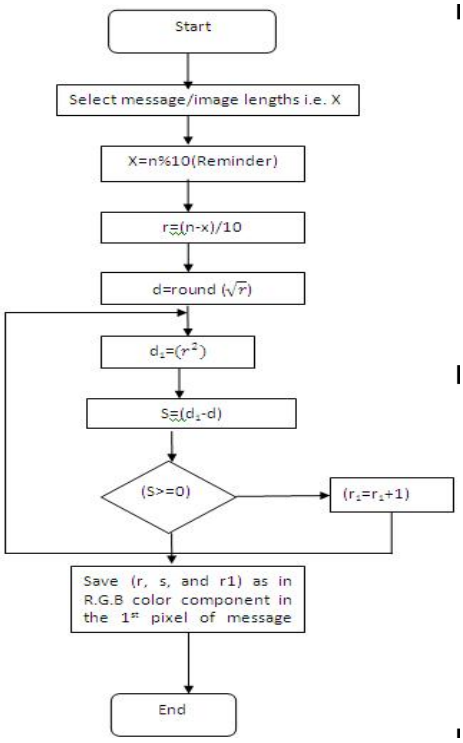


Figure 4:-Length of image which is stored in the 1st pixel after Encoding

If the end of stream reached for the password stream then restart it from the beginning of the stream. If the R component of the current sub matrices is completed it will go the G component of the sub matrices and after that to the G component respectively .If the current sub matrices is completed then go the corresponding next sub matrices .If the end of stream is reached for the watermark image save the file and exit.[1],[2].

B Movie Decoding Algorithm:

Here we will discuss how to decode the hidden image from a video frame used as cover image using a key i.e. password. Input i.e. selecting the video frame no and password, Creating blocks using the image pixel matrix. Message length decoding from the cover image. Decoding and decrypting the hidden image from the frame of that video. This project mainly uses .jpg image file formats as cover image. The minimum size of the .jpg file must be at least 4x4

a. Get the Input which is a video file.

b. Get the input of the password stream, which will be used as a key

STEP 2:-After getting the input of the selected .jpg cover image file our next approach will be as same as in the case of movie encoding algorithm.

a. Generate a matrix containing the pixel values of the selected frame as cover image, I.

b. Now, we will breaks up the 2-D array i.e. the matrix I into a cell array of adjacent of sub matrices of I , and store it as Magic_ Block, such that minimum number of sub matrices have the size of 4X4 i.e. 4 square. where ,we are taking a matrix of 10X10 and creating magic_ block contains four 4X4,two 2X4 , two 4X2 and one 2X2 sub matrices

STEP 3:-This step is for the retrieving the message length from the 1st pixel of the image.

a. Get the value of R , G and B component of the 1st pixel as r , g and b.

b. In this step we will square the value of r and subtracting it from the value of g and then result will be multiplied by 10 and added with the value of b to get the message length. Message_Length = (((r^2) - g).*10) + b

STEP 4:-After getting the message stream length from the first pixel we will go through the following process to decode the message length starting from the second pixel of the current sub matrices ,as (1st pixel is used to decode the message length)we are using here the created Magic_ Block in the STEP- 2 . i.e. the sub matrices of the 2-D array of pixel values, I. The R component 1st

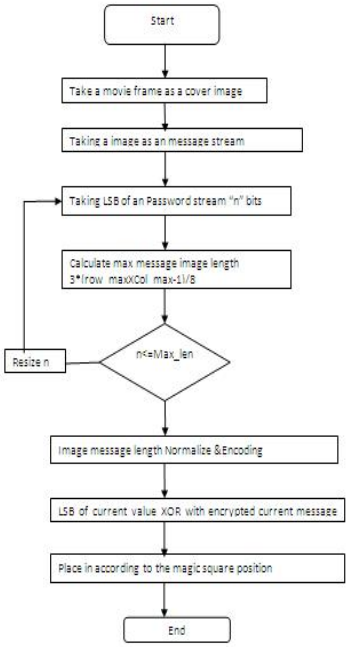
sub matrices will be firstly used. As the 1st pixel is used to store the message length which is 1st pixel of the 1st of the sub matrices. Hence, we will start from the second pixel of the 1st sub matrices. And the R components of the 1st sub matrices' will be 1st used after that the G component and after that the B component. Now we are checking the size of the of 1st sub matrices, whether it is 4X4 or not .If it is 4X4 then it can be matched with the magic square. now, Take the LSB value of the current pixel and the LSB value of the password stream.XOR both the value and we will get a value which will be the message bit .Store it in a stream. Repeat the process for the

a. Until it reaches the offset value to 8.Then the message bit stream will be converted to get the 1st message byte and store it into the message byte stream. After reaching it to the 8 set the offset to the value 1 again and repeat the process

b. If the current sub matrices R component is completed then, go to the G component and after that to the B component. If the current sub matrices are completed then go to the next sub matrices of the magic block.

c. If the password stream reaches to the end-of – stream set it to the beginning of stream.

d. Repeat the process until the it reaches the message length. After it reaches the message length, show the hidden image.



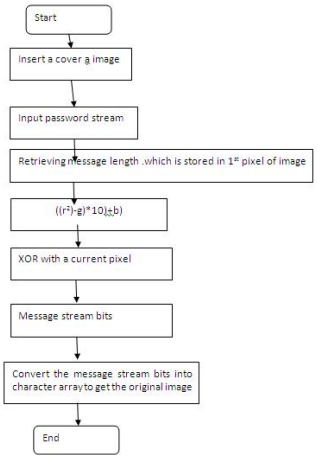


Figure5:-Represents Movie Encoding and Decoding

III. EXPERIMENTAL RESULTS

It is a Performance evaluation. As we have studied the proposed algorithm in the previous section it is clearly seen that it has some clear advantages .It combines both steganography and cryptography, it is more secure. It uses magic square of four as a transformation tool which is used in ancient history, hence providing more security. As we are using 24-bitmap image as .bmp file it certainly does give extra capacity to contain more data within the cover image. As we have calculated, the maximum msg length would be= 3*(row_max*col_max-1)/8.As we are creating sub matrices using the matrix containing pixel values. It is actually time consuming realive to other techniques. This method also normalized and encodes the message length in the first pixel that it provides more security to hidden data.

A movie .avi extension file is used as a cover movie and a .jpeg as a message image and a text password combined a stego object

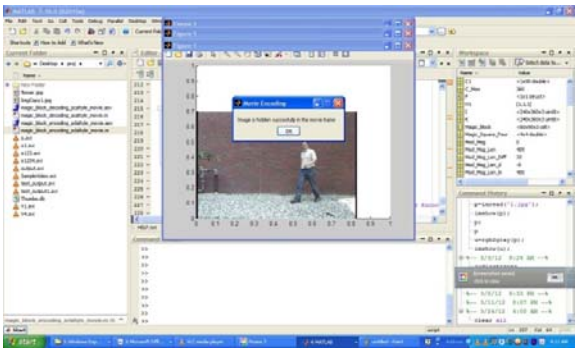


Figure 6:-Represent an image is successfully hidden into a movie frame

Figure 7:-Several test image & movie frame

Peak Signal to Noise Ratio,[8] often abbreviated PSNR, the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation.

$PSNR=20.\log_{10}(MAX/\sqrt{MSE})$.

Table-I

The original Image	PSNR of the stego. image
White Flower	37.5db

IV CONCLUSION:

In this research, we proposed a framework for a robust and secure algorithm for JPEG. Here we are present an approach for hiding a .jpeg image using stereography and cryptography in spatial domain using magic square. This algorithm consists of image embedding and decoding .Here, an image is embedded in a video using a key successfully and we can extract that image from that video. The PSNR of each tested image > 37 db.

REFERENCES

[1] C.C.Chang T.D.Cue, "An Image Authentication Scheme Using Magic Square",2nd International Conference on Computer Science and Information Technology, China IEEE ©2009.

[2] Ching-Yung Lin,Shih-Fu Chang , "A Robust Image Authentication Method Distinguishing JPEG Compression from Malicious Manipulation", IEEE,pp-153-168 © 2001.

[3]. Jiri Fridrich, "Robust Bit Extraction from Images Multimedia Computing and Systems", Center for Intelligent Syst, State Univ. of New York, Binghamton, NY, IEEE pp-536-540, ©1999.

[4] Christian Rey and Jean-Lue, "Blind Detection of Malicious Alterations on Images using Robust Watermarks",pp-67-71, ©2001.

[5] R. Popa, "An Analysis of Steganographic Techniques", Department of Computer Science and Software Engineering, http://ad.informatik.unifreiburg.de/mitarbeiter/will/dlib_bookmarks/digital-watermarking/popa/popa.pdf, ©1998.

[6] C. Kurak and J. McHugh, "A Cautionary Note On Image Downgrading," Proc. IEEE Eighth Ann. Computer Security Applications Conf., IEEE Press, Piscataway, N.J, pp. 153-159, ©1992.

[7] W. Bender et al., "Techniques for Data Hiding," IBMSystems J., Vol. 35, Nos. 3 and 4, pp. 313-336, October, ©1996.

[8] T. Liu and Z.D. Qiu, "The Survey of Digital Watermarking Based ImageAuthentication Techniques", 6th International Conference, pp. 1566-1559, © 2002.

[9] Stephen Mahoney and Neil F. Johnson. "Audio Steganography and Steganalysis", Workshop on Statistical and Machine Learning Techniques in Computer Intrusion Detection, Johns Hopkins University, June 11-13, © 2002.

[10] Neil F. Johnson. "Steganography: Art & Science of Hidden Communication", Office of Naval Research (ONR) Naval-Industry Partnership Conference, Washington, DC, USA, 13-14 August ©2002.

[11] Neil F. Johnson, Edmond G. Kong. "Investigating Hidden Information: Steganography and Computer Forensics", American Academy of Forensic Sciences (AAFS) 54th Annual Meeting, Atlanta, GA, February 11-16, © 2002.

[12] Neil F. Johnson and Sushil Jajodia. "Steganalysis: The Investigation of Hidden Information", IEEE Information Technology Conference, USA, pp 113-116., September, ©1998.

[13] MATLAB® copyright 1984-2009 The Math Work Inc, USA.

❑❑❑

A Heuristic AO Star Approach for Decision Tree Induction

Phani Kishore Rompicharla & D. V. Manjula

Department of Computer Science & Engineering, Kakinada Institute of Engineering and Technology, Korangi, Kakinada,East Godavari District, Andhra Pradesh, India

E-mail : phani.rompicharla@gmail.com, dv.manjula@gmail.com

Abstract **-** A Decision tree is a flow chart like tree structure where each internal node denotes a test on attribute each branch represents an outcome of the test and leaf nodes represents classes or class distribution. The basic algorithm is a Greedy algorithm that constructs decision trees in top-down recursive divide and conquer manner. This algorithm is used for classification and it is a Supervised Learning because class label of each training sample is provided. ID3Optimization algorithm is proposed based on interestingness gain of attribute. We are proposing a new Heuristic technique called AO* that provides optimal solution based on evaluation of heuristic function applied to every node in the tree. The attribute selection can be done by considering the best attribute having least Heuristic function value by which we provide the optimal Decision Tree. This provides fast and more efficient method of constructing the Decision Tree taking predominant function of heuristic that best selects an attribute for classification.

Keywords*— AO*, Decision tree, Heuristic function, Attribute Selection Criteria, Futility Value.*

I. INTRODUCTION

Data Mining is a powerful technology used in the data warehouses. Data mining is to discover the relationship and rules existing in data, to predict the feature trends based on the existing data, finally to fully explore and use these wealth knowledge hiding in the databases [11],[12]. Decision tree induction is most widely used practical method for inductive learning; it plays an important role in the process of data mining and data analysis.

As the Classification by Decision tree induction uses attribute selection measure is taken as information gain heuristic. This paper proposes a new heuristic approach for attribute selection, which is done by applying the Heuristic function to each attribute that best selects an attribute for further split of the tree.

II. ID3 ALGORITHM

ID3 algorithm [13] is a decision tree learning algorithm based on information entropy proposed by Quinlan in 1986. The core of ID3 algorithm is: selecting attributes from all levels of decision tree nodes; using information gain as attribute selection criteria; each selecting an attribute with the largest information gain to make decision tree nodes; establishing branches by the different values of the node; building the tree nodes and branches recursively according to the instances of various branches; until a certain subset of the instances belonging to the same category.

Set S is a collection of data samples. Assume that class label attribute has m different values, define different classes C_i (i=1, 2...m). Let s_i represent the sample number of classes C_i. For a given sample the expected information for the classification is calculated as follows [10]:

$$I(s_1, s_2 \ldots s_m) = -\sum_{i=1}^{m} p_i \log p_i \qquad (1)$$

Where p_i is the probability of any sample belonging to class C_i. Assumed attribute A has v distinct values :{ a_1, a_2 ...a_v} divides s into v sub-sets {s_1, s_2 ...s_v} by attribute A. Making A as test attribute, this subset is the branch getting from the nodes including set S. Assumed s_{ij} is the sample number of class C_i in subset s_j. The Entropy divided by A is given as follows [12]:

$$E(A) = -\sum_{J=1}^{v} \frac{s_{1j}+s_{2j}+\ldots+s_{mj}}{s} I(s_{1j}, s_{2j} \ldots s_{mj}) \qquad (2)$$

Here $\frac{s_{1j}+s_{2j}+\ldots+s_{mj}}{s}$ is the weight of the j subset

And is equal to the sample number of subset dividing the total number of S. The smaller entropy is the higher purity of the divided subset. For the given subset S_j, its expected information is given as [12]:

$$I(s_{1j}, s_{2j} \ldots s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log p_{ij} \qquad (3)$$

Where $p_{ij} = S_{ij}$ is the probability that the sample S_j $\lceil |S_{jj}$

Belongs to the class C_i . Get information gain according to the expected information and entropy. Information Gain is calculated by [11]:

$$\text{Gain}(A) = I(s_1, s_2 \ldots s_m) - E(A) \qquad (4)$$

Calculate the information gain off each attribute according to ID3 algorithm and select the attribute with highest information as the test attribute for a given set. Create the node of the selected test attribute ark according to this attribute create a branch for each value of this attribute and accordingly divide the sample.

III. AO* IMPLEMENTATION

AO* is the simplification of Problem Reduction Algorithm and A* Alogorithm. AO* Algorithm uses a single structure GRAPH rather than the two lists OPEN and CLOSED which were used in A* Algorithm.

In AO* algorithm each node in the GRAPH will point down to its immediate SUCCESSOR. Each node in the GRAPH will also have associated with its a h' value, an estimate of cost of a path from itself to a set of solutions (which were NODE's). So h' value of the given data will be serving as the estimate of goodness of a node. H'(n) The heuristic function is obtained from Problem Reduction algorithm which is defined by [14]:

h'(n) = estimated cost from this node to the goal node the algorithm terminates when h'(n)=0 or h'(n)>FUTILITY value. In general ID3 the attribute selection criteria is done by the process of information gain heuristic. This Paper proposes a new attribute selection criteria by applying heuristic function h'(n) on each attribute.

Attribute Selection Criteria:

Step-1: Calculation of h'(n)

h'(n)=E(n) with respect to goal attribute Buys_Computer
[E(n) is caliculated as given in [14]:]

Step-2: Calculation of FUTILITY value

FUTILITY = The h' value of the starting attribute selection which is not getting selected as an goal node for us in the due course or before getting the all the rules

Thus step-1 through step-2 are recursively applied for building the Decision Tree. The algorithm terminates when h'(n) is equal to 0 or h'(n) is greater than FUTILITY.

IV. IMPROVEMENT OVER BASE ALGORITHMS'

In ID3 Algorithm the number of calculations were very high as they calculate both f(n) which is similar to h'(n) and also we need to calculate GAIN value of each attribute in each iteration which is increasing the number of calculations.

In A* Algorithm also similar to ID3 algorithm and calculate the f(n) and also GAIN Values.

In problem Reduction also the f(n) values were common and only the gain values were replaced with the FUTILITY values. But the FUTILITY value is to be calculated for each iteration.

In the present Algorithm called AO* algorithm we need to calculate only h'(n) value and also the FUTILITY value but only once as said above.

V. VALIDATION OF ALGORITHM

Fig. 1: The Data Set used in the Algorithm

RID	AGE	INCOME	STUDENT	CREDIT_RATING	BUYS_COMPUTER
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Figure 1 gives the data set [3] affecting Buys_computer as YES or NO. There are four attributes: age, income, student, credit. These four attributes are used to verify that the particular attributes were leading to Buys_computer YES or NO and the algorithm is constructed as follows:

1. *Initial Attribute:*

Fig. 2: The Result for the Initial Attribute Selection

To Select the First Split Attribute		
h' value for age	:	0.6935361388961918
h' value for income	:	0.9110633930116763
h' value for student	:	0.7884504573082896
h' value for credit rating	:	0.8921589282623617
The Futility value is	:	0.9110633930116763
The next attribute selected was	:	**age**

Now according to the values calculated above the training set is classified depending on attribute "Age" which splits up into 3 values (age='youth', age='middle_aged' and age='senoir').

Since of the two attributes were having the h'(n) is greater the FUTILITY value (income and credit_rating). Those two attributes were restricted to participate in the next attribute selection. Two attributes whose h'(n) is less than the FUTILITY value (age and student) So in those two "age" is having less value when compared to "student".

2. Next Attribute with age=youth

Fig. 3: The Result for the attribute age=youth

To Select the second Split Attribute age='youth'		
h' value for income	:	0.4
h' value for student	:	0.0
h' value for credit rating	:	0.9509775004326937
The Futility value is	:	0.9110633930116763
The next attribute selected was	:	**student**

As the value of credit_rating is having the value greater than the FUTILITY value the it is restricted and among the other two values (income and student) student is having the best value. So the next attribute selected for this category is "student".

3. Next Attribute with age=youth and student=yes

Fig. 4: The Result for the attribute age=youth and student=yes

To Select the third Split Attribute age='youth' and student='yes'		
h' value for income	:	0.0
h' value for credit rating	:	0.0
The Futility value is	:	0.9110633930116763
The next attribute selected was	:	Their is no further split.... so we need to form rules & move to another branch.

Since all the h'(n) values were zero there is no further splits and it is labelled as SOLVED. We need to traverse back to the parent node and try to solve the next branch.

4. Next Attribute with age=youth and student=no

Fig. 5: Result for the attribute age=youth and student=no

To Select the fourth Split Attribute age='youth' and student='no'		
h' value for income	:	0.0
h' value for credit rating	:	0.0
The Futility value is	:	0.9110633930116763
The next attribute selected was	:	Their is no further split.... so we need to form rules & move to another branch.

Since all the h'(n) values were zero there is no further splits and it is labelled as SOLVED. We need to traverse back to the parent node and try to solve the next branch.

5. *Next Attribute with age=middle_aged*

Fig. 6: The Result for the attribute age=middle_aged

To Select the fifth Split Attribute age='middle_aged'		
h' value for income	:	0.0
h' value for student	:	0.0
h' value for credit rating	:	0.0
The Futility value is	:	0.9110633930116763
The next attribute selected was	:	Their is no further split.... so we need to form rules & move to another branch.

Since all the h'(n) values were zero there is no further splits and it is labelled as SOLVED. We need to traverse back to the parent node and try to solve the next branch.

6. Next Attribute with age=senior

Fig. 7: The Result for the attribute age=senior

To Select the sixth Split Attribute age='senior'		
h' value for income	:	0.9509775004326937
h' value for student	:	0.9509775004326937
h' value for credit rating	:	0.0
The Futility value is	:	0.9110633930116763
The next attribute selected was	:	credit rating

Since in Fig. 7 the values of attributes income and student were having the values greater than FUTILITY value. So they both were restricted in the next attribute selection criteria. Now the remaining credit_rating attribute will be the next attribute.

7. Next Attribute with age=senior and credit_rating=fair

Fig. 8: The Result for the attribute age=senior and credit_rating=fair

To Select the seventh Split Attribute age='senior' and credit_rating='fair'		
h' value for income	:	0.0
The Futility value is	:	0.9110633930116763
The next attribute selected was	:	Their is no further split.... so we need to form rules & move to another branch.

Since all the h'(n) values were zero there is no further splits and it is labelled as SOLVED. We need to traverse back to the parent node and try to solve the next branch. Here the speciality is the we have not considered the attribute student as it has got the solution in the earlier traversals itself so we have not selected considered the student attribute.

8. Next Attribute with age=senior and credit_rating=excellent

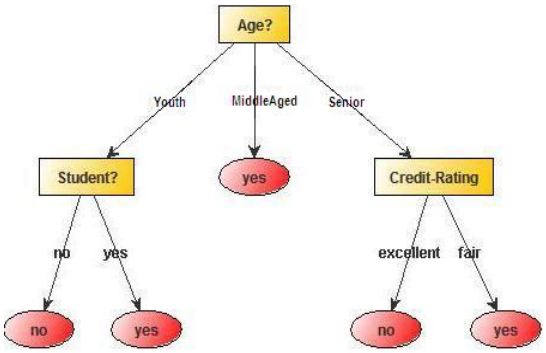
Fig. 9: The Result for the attribute age=senior and credit_rating=excellent

To Select the eigth Split Attribute age='senior' and credit_rating='excellent'		
h' value for income	:	0.0
The Futility value is	:	0.9110633930116763
The next attribute selected was	:	Their is no further split.... so we need to form rules & move to another branch.

Since all the h'(n) values were zero there is no further splits and it is labelled as SOLVED. We need to traverse back to the parent node and try to solve the next branch. Here the speciality is the we have not considered the attribute student as it has got the solution in the earlier traversals itself so we have not selected considered the student attribute.

The algorithm Terminates and the final decision tree is as follows:

Fig. 10 Decision Tree



VI. CONCLUSION

The proposed paper created a new innovative and fastest approach of attribute selection criteria taking best splitting attribute with least heuristic functional value. This will also calculate the FUTILITY value only once in the total decision tree. It reduces the number of calculations by generating the decision tree with out calculating the gain factor. This paper can be extended to larger databases which can provide best splitting criteria and also for the other databases in future were we have the same type of attribute selections.

REFERENCES

[1] "ID3 Optimization Algorithm Based on Interestingness Gain " by Liu zhongtao Wang Hong.

[2] "Decision Tree Induction: An Approach for Data Classification Using AVL-Tree" by Devi Prasad Bhukya and S. Ramachandram.

[3] "A NEW HEURISTIC OF THE DECISION TREE INDUCTION" by Ning Li, Li Zhao, Ai-xia Chen, Qing-wu Meng, Guo-fang Zhang.

[4] "A High Speed Decision Tree Classifier Algorithm for Huge Dataset" by Thangaparvathi, Anandhavalli , Mercy Shalinie

[5] "AN EFFICIENT ALGORITHM FOR INDUCTION WITH RANDOM SAMPLNG" by ALI MIRZA MAHMOOD1,2* MRITHYUMJAYA RAO KUPPA3 V SAI PHANI CHANDU

[6] "Decision Tree Induction using Adaptive FSA" Hemerson Pistori and Joao Jose Neto.

[7] "Hyper-heuristic Decision Tree Induction" by Alan Vella, David Corne and Chris Murphy.

[8] "Optimization of Decision Tree Based on Variable Precision Rough Set" by Weiguo Yi, Jing Duan and Mingyu Lu.

[9] "Improved ID3 Algorithm" by Liu Yuxun and Xie Niuniu.

[10] Liu Yuxun, Xie Niuniu *"Improved ID3 Algorithm"* 3RD IEEE Conference (ICCSIT.2010.5564765):465-468.

[11] G.Eason, B.Noble, and I.N.Sneddon, "On *certain integrals of W.J.Frawley, G.Piatetsky Shapiro, C.J.Matheu. Knowledge discovery in databases: An overview[M]"*, USA, AAAI/MIT Press, 1991 1-27

[12] Han Jiawei, Micheline Kamber *Data Mining Concepts and Techniques*

[13] J.R.QUINLAN.*Induction of Decision Trees[J]*.Machine Learning 1968:81-106.

[14] E.Rich, K.Knight, Shiv Shankar B Nair *Artificial Intelligence.*

❑❑❑

Automated Vehicle Using Speech Recognition

Heena Solanki, PrabhaSharma, Shweta Patil, Pooja Kasture & Vaishali Deshmukh

Depatment Of computer Engineering, Smt. Kashibai Navale College of Eng
E-mail : heenavsolanki@gmail.com, prabhasharma.13@gmail.com, shwetarpatil@yahoo.co.in, poojakastureis@gmail.com, mailtovsd@gmail.com

Abstract **-** Of late road traffic becomes congested and unmanageable. Notwithstanding the fact that government takes all steps to reduce congestion, in view of increasing two wheelers, four wheelers not to speak of autos, all important roads are bound with vehicles. Further many person do not drive with diligent neither they think of themselves nor of others. In the circumstance it is of dire need that with the available infrastructure we should do something for the benefit of the society as a whole .in this context we have developed a mechanism of speech recognition and GPS GSM which reduces the accidents and also helps for physically/visually challenged persons. The proposed system will reduce the accident by neighboring vehicle detection, obstacle detection, controlling the vehicle speed. All these above mentioned facilities are automated.

Keywords*-components: Automated, Speech recognition, Global System for Mobile communication (GSM), Global Positioning System (GPS).*

I. INTRODUCTION

Automation**,** the application of machines to tasks once performed by human beings or, increasingly, to tasks that would otherwise be impossible. Although the term mechanization is often used to refer to the simple replacement of human labour by machines, automation generally implies the integration of machines into a self-governing system. Automation has revolutionized those areas in which it has been introduced, and there is scarcely an aspect of modern life that has been unaffected by it.

The term automation was coined in the automobile industry about 1946 to describe the increased use of automatic devices and controls in mechanized production lines. The origin of the word is attributed to D.S. Harder, an engineering manager at the Ford Motor Company at the time.[1]

II. SURVEY OF EXISTING SYSTEM

Research on autonomous vehicles started as far back as the 1980's but only in the past 10 years or so have any big leaps been made in the industry. Companies like GM are putting more and more money into driver assisted vehicles. Driver assisted vehicles could eventually turn into fully automated vehicles. With the research being done by big companies and schools for robotics, autonomous vehicles are quickly becoming a more realistic possibility as time progresses. The US Government set a goal to have a third of its vehicles in the Armed Forces be unmanned by the year 2015 (Anhalt et al., 2008). However, they did not feel they could meet their deadline without help, so they set up a contest. In 2004, the Defense Advanced Research Projects Agency (DARPA) unleashed what they called the Grand Challenge. The challenge was to build an autonomous vehicle that could maneuver on off-road terrain. The goal was to get a vehicle to a certain destination in a set amount of time. No vehicle was able to make it to the destination that first year, but this opened the doors of innovation. After the first Grand Challenge, researches started putting more time and effort into the topic of fully autonomous vehicles. The second Grand Challenge in 2005 proved much more successful. Many vehicles were able to navigate safely through the desert.

This lead to the most recent challenge in 2007, called the Urban Challenge. The Urban Challenge is the most significant of the three challenges mentioned in this paper. The Government wanted a vehicle that was able to navigate safely through the streets of an urban environment. The vehicle needed to follow the rules of the road, navigate itself successfully to its destination, and be able to compensate for unforeseen situations. Many vehicles were successful which indicates that autonomous vehicles are not an unrealistic or far in the future idea. The vehicles had the ability to do the following: recognize the road and stay on it, detect objects around the vehicle and decipher what they were, handle different terrain, and finally, set goals, tasks, and handle situations that occur on roadway systems. And

that is how this paper will be categorized. The remainder of this paper will focus on different techniques and methods used in certain areas of vehicle autonomy. [6]

III. PROPOSED SYSTEM

In our proposed system we have made the vehicle to move automatically by detecting the obstacles thus avoiding the need of a driver. We have used Microcontroller Atmega 16(8 bit) for automatic motion and the colloidance with the neighboring vehicles are being made by IR sensors. The vehicle will move according to the commands given to the Speech Recognition module. And GPS GSM are being used in case of accident.

IV. METHODOLOGY

The input given to the Atmega 16 microcontroller is IR sensor, Speech recognition, and regulated power supply. The microcontroller will accept he command from the speech recognition module and will move vehicle according to that command. The IR sensors are used to detect the obstacle or the neighboring vehicle. When the IR sensors detect any obstacle it will tell the microcontroller and the microcontroller will control the vehicle speed or stop the vehicle. The microcontroller will control the motors. Two DC motors are used to drive the vehicle and these two motors will receive command from microcontroller and will move according to that command i.e left, right, forward, backward or stop. The position of the vehicle is given using the GPS system. The position of the vehicle will be send to the specified mobile number using the GSM system.

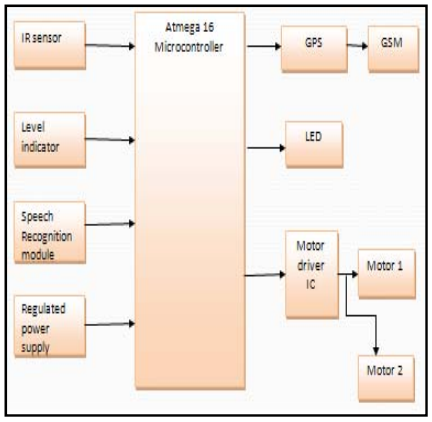
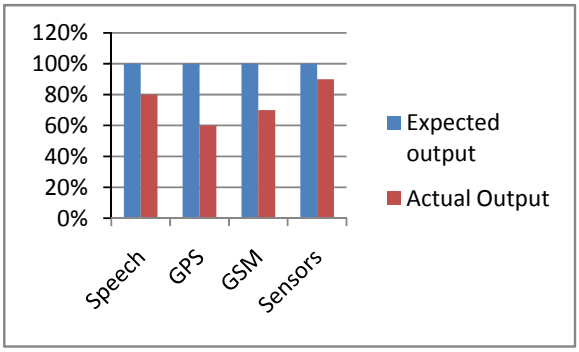


Figure 1: System Architecture

V. EXPERIMENTAL RESULTS

Experiments are performed to verify the effectiveness of the proposed system. The expected output and actual output of each module in the system is compared by giving various inputs to the system. The graph represents the expected output and the actual output. The inputs are given using the boundary value analysis method.



6 MATHEMATICAL MODEL

1) **Speech Recognition:** The speech recognition system is a completely assembled and easy to use programmable speech recognition circuit. Programmable, in the sense that you train the words (or vocal utterances) you want the circuit to recognize.

Assumptions:

1) A= sequence of symbols drawn from some alphabet B.
   $A=a_1,a_2,\ldots,a_m \;\; a_i \;\; B$
2) W= a string of n words each belonging to a fixed and known vocabulary V.
   $W=\arg_w \max P(W|A)$
   Where, P(W|A)= conditional probability of W, given A.

Using Baye's Formula,

$$P(W|A)=\frac{P(W)\,P(A|W)}{P(A)}$$

Where, P(W)= prior probability that the word string W will be uttered.
P(A|W)= probability that when W is uttered the acoustic evidence A will be observed.
P(A)= prior probability of observing A.

P(A)=constant.

So, we have to find out:

$$W=\arg_w \max P(W)\,P(A|W)$$

GPS: GPS is global positioning system which tracks the position of the system. It gives the exact

location in terms of longitude, latitude, altitude and time.

$$p_1 = \sqrt{(X - x_1)^2 + (Y - y_1)^2 + (Z - z_1)^2} - c\,dT$$

$$p_2 = \sqrt{(X - x_2)^2 + (Y - y_2)^2 + (Z - z_2)^2} - c\,dT$$

$$p_3 = \sqrt{(X - x_3)^2 + (Y - y_3)^2 + (Z - z_3)^2} - c\,dT$$

$$p_4 = \sqrt{(X - x_4)^2 + (Y - y_4)^2 + (Z - z_4)^2} - c\,dT$$

The left side represent the pseudo range measurement by receiver. Expression under the square root sign is the true range to the satellite.

Co-ordinates X,Y,Z represents position of the receiver. The term C dT is the contribution to the pseudo range from the receiver clock offset dT.

The set of four equations must be solved simultaneously to obtain the values for X,Y,Z together with the clock offset dT.

Although the equations are written in terms of geocentric Cartesian coordinates the resulting X,Y,Z values can be easily be converted to latitude, longitude and height.

VII. ADVANTAGES AND LIMITATIONS

A. Advantages:

1) System reduces the number of accidents to a great extent.
2) More efficient than the manually operated vehicle.
3) User safety is an important issue for automates vehicle so automated vehicle provide improve safety.

B. Limitations:

1) Higher level of maintenance needed than a manually operated vehicle.
2) Lower degree of flexibility.

VIII. CONCLUSION

This paper presents aim towards the welfare of physically impaired people. Automated vehicles would be a great asset. Great foundation for future autonomous development. These technologies are being improved & enhanced as well as being made more affordable.

IX. FUTURE SCOPE

In the future scope of this system we are willing to do the traffic light detection, automated parking, sign board detection.

X. REFERENCES

[1] "Techniques used in Autonomous vehicle system: A survey" [J. C. Last, Robbie Main University of northern Iowa]

[2] "Safety Automation of cars using Embedded microcontrollers" [Abhinay Ray, Pratosh Deepak Rajkumar]

[3] http://www.sunrom.com/speech-recognition/speech-recognition-system

[4] http://www.rhydolabz.com/index.php?main_page=products_new

[5] http://www.aero.iitb.ac.in/web/uploadfiles/2007-09-28--23-04-e-drive%20(ieee).pdf

[6] http://www.wikipedia.com/automated-vehicle/DARPA

❑❑❑

Antenna Array Synthesis by Implementing Non – Uniform Spacing Using Tsukamoto Fuzzy Logic Controller

Sanmoy Bandyopadhyay
Department of Electronics and Communication Engineering National Institute of Technology, Durgapur
Durgapur, India
e-mail:sanmoy1985@rediffmail.com

Ashok Babu.Chatla
Department of Electronics and Communication Engineering
National Institute of Technology, Durgapur
Durgapur, India
e-mail: ashokbabu23@gmail.com

B.Maji
Department of Electronics and Communication Engineering
National Institute of Technology, Durgapur
Durgapur, India
e-mail: bmajiecenit@yahoo.com

Abstract—**The paper is based on the work of the Antenna Array Synthesis by Space Perturbation Using Tsukamoto Fuzzy Logic Controller. Using Tsukamoto fuzzy logic controller it has been tried to get the new antenna array radiation pattern. Here in Tsukamoto fuzzy controller two inputs namely amplitude and phase shift are given and corresponding defuzzified value of change in spacing between each of the array elements is obtained as the output from the fuzzy controller. Putting new spacing value formed by the combination of the output change in spacing value and old spacing value, in the array factor formula we get the new array radiation pattern. This work can be implemented in future to reduce the SLL of the antenna array where we are not sure of the exact value of amplitude or phase shift or other parameter of array.**

Keywords— **Fuzzy logic, Antenna Array, Fuzzification, Defuzzification, Side Lobe Level, Array Factor, Membership function, Fuzzy Logic Controller, Array Pattern.**

I. Introduction

Antenna array can be defined as the arrangement of several antenna elements in space. Antenna array is necessary to generate the radiation pattern in desired direction. It is used to get the high directivity. Most of time it is necessary to generate the desired radiation characteristic. The task, in general, is to find not only the antenna configuration but also its geometrical dimensions and excitation distribution [1]. The designed system should yield, either exactly or approximately, an acceptable radiation pattern, and it should satisfy other system constraints. This method of design is usually referred to as synthesis. The goal in antenna array geometry synthesis is to determine the physical layout of the array that produces a radiation pattern that is closest to the desired pattern [2]. There are various methods that can be utilized for the antenna array synthesis namely GAs [3 - 5], Simulated Annealing (SA) [6], Tabu Search [7], Taguchi's method [8], Memetic Algorithms (MAs) [9], PSO [10, 11], Fuzzy Genetic Algorithm [12], Fitness-Adaptive Differential Algorithm [2], Fuzzy PSO Algorithm [13], Bacteria Foraging Optimization Technique [14] etc. In our work we have mainly concentrate in doing the antenna array synthesis by space perturbation applying the fuzzy logic [15]. In our work we have applied the Tsukamoto fuzzy logic controller. Using Tsukamoto fuzzy logic controller it has been tried to get the new antenna array radiation pattern [16]. In this we have applied the input in the Tsukamoto fuzzy logic controller and taken the multiple number of corresponding output from the defuzzifier of Tsukamoto fuzzy logic controller, in this work we have taken 19 output from the fuzzy controller and summing it with the old spacing value we have implemented it in the formula of the corresponding array factor and as a result we have got the new antenna array radiation pattern. We have mainly implemented our work using the MATLAB software.

II. Fuzzy Logic

Fuzzy logic can be said as extended case of boolean logic or boolean logic can be said as special case of fuzzy logic [17]. Actually it introduced a third variable in between zero and one. There are many things which can not be explained by boolean logic, such thing can be explained by fuzzy logic. For example suppose we describe the man who is having salary rupees twenty thousand as rich one and who is having salary rupees two hundred as poor one and such thing can be easily represented by boolean function, but the situation that the man is having salary rupees nineteen thousand nine hundred ninety nine, can not be explained by the boolean logic, for this we need the help of fuzzy logic.

Just as human brain doesn't need very accurate information about the system parameters and still it can take effective decision even without knowing the exact mathematical model of the system on the basis of his experience, fuzzy logic based controller gives us that ability to solve many complex, nonlinear hard to mathematically modelled problems with comparatively simpler systems. It is very useful where there is a lack of exact data.

II. Tsukamoto Fuzzy Logic Controller

The Tsukamoto fuzzy model was proposed by Y. Tsukamoto in year 1979. In Tsukamoto fuzzy models [18], the consequent of each fuzzy if-then rule is represented by a fuzzy set with a monotonical Membership Function. As a result, the inferred output of each rule is defined as a crisp value induced by the rule's firing strength. The overall output is taken as the weighted average of each rule's output. Since each rule infers a crisp output, the Tsukamoto fuzzy model aggregate each rule's output by the method of weighted average and thus avoids the time consuming process of defuzzification. Fig. 1 illustrates a two – input single – output Tsukamoto fuzzy model.

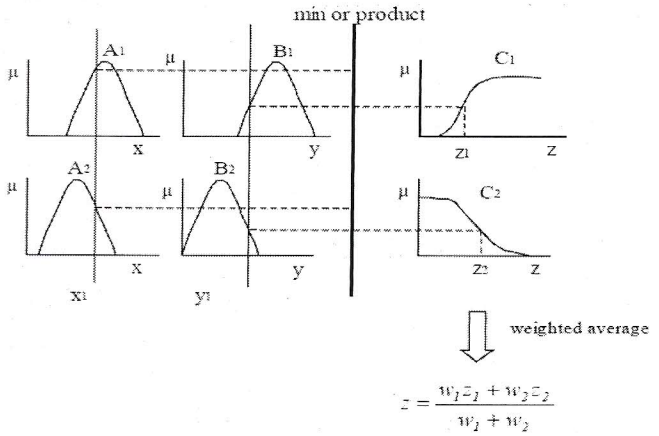


Fig. 1 Two – input single – output Tsukamoto fuzzy model

III. Tsukamoto Fuzzy Logic Controller in Antenna Array Synthesis by Space Perturbation

In antenna array synthesis by space variation using Tsukamoto fuzzy logic controller our objective is to obtain the multiple numbers of outputs that is change in spacing between the antenna arrays elements from the defuzzifier and after summing it with the old values of spacing we have implemented the new spacing in the array factor formula to synthesize the radiation pattern of the antenna array. In our work we have taken the number of output same as number of antenna array element. Thus here we have applied the parameter of the antenna array factor in the fuzzifier of the fuzzy logic controller as shown in Fig. 5. We have implemented two inputs in the fuzzifier of the fuzzy logic controller, these are namely current amplitude and the phase shift of the antenna array. In fuzzifier output we have taken three linguistic variables for both current amplitude and phase shift. For current amplitude the three linguistic variables are namely No ChangeA, LowA, and HighA shown in Fig. 2. Similarly for phase shift there are three linguistic variables namely No Change, Low and High shown in Fig. 3. After getting the fuzzified result from the fuzzifier we applied the fuzzified result into the inference engine. As the inference engine gives its decision by consulting with the Rule base [19], so in our problem we have implemented a rules base based upon some previous output from the antenna array factor formula to get the desired antenna array radiation pattern.

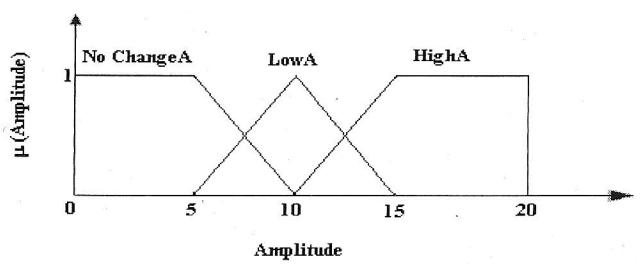


Fig. 2 Linguistic Variable for the fuzzified value of the amplitude

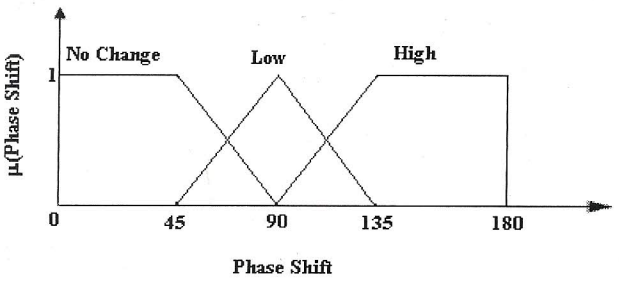


Fig. 3 Linguistic Variable for the fuzzified value of the phase shift

TABLE I
Rule Base Applied for Change in Spacing Δd (1)

Amplitude / Phase Shift	No ChangeA	LowA	HighA
No Change	Output Change in Spacing *Low1*	Output Change in Spacing *Low3*	Output Change in Spacing *Very Low3*
Low	Output Change in Spacing *Low2*	Output Change in Spacing *Very Low2*	Output Change in Spacing *No Change2*
High	Output Change in Spacing *Very Low1*	Output Change in Spacing *No Change1*	Output Change in Spacing *No Change3*

We have made the rule base for each of the output of Change in Spacing denoted by *Δd*, for this work we have made the number of rule base same as the number of antenna array element that equal to 20. Table. I is an example of such a rule base which is applied in our work to calculate the fuzzified output for change in spacing Δd(1). Similarly we have applied different rule base to generate the fuzzified output for change in spacing for Δd (2) to Δd (19). Now by consulting with the rule base the inference engine give the multiple numbers of fuzzified outputs of the change in spacing between each of the antenna array elements. In our work we have taken the number of output same as number of antenna array element. In this fuzzified output of the change in spacing between the array elements we have also taken three linguistic variables these are namely No Change, Very low and Low which is shown Fig. 4. Now this fuzzified value of the output spacing is applied into the defuzzifier from the inference engine to obtain the output spacing in terms of the crisp value. For the process of defuzzification in Tsukamoto fuzzy model we have made a program in MATLAB m.file. According to the program we have generated line which starts from the point of membership function μ (Spacing) along y – axis and cut the portion of the curve which is selected according to the rule – base. After this corresponding value of the Spacing along x – axis is being calculated from the graph. Now by

using the method of weighted average defuzzified value or the crisp value of the output change in spacing is calculated.

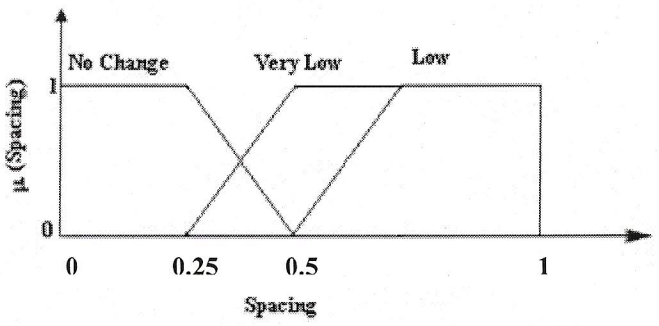


Fig. 4 Linguistic Variable for the fuzzified value of the output spacing

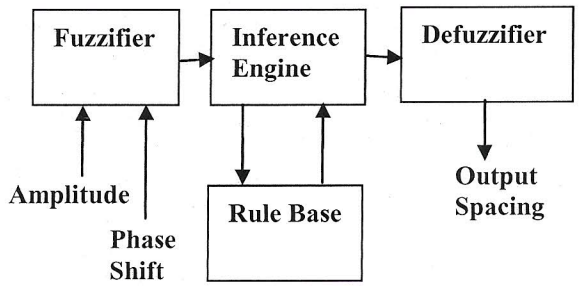


Fig. 5 Block diagram of our experimental setup

Now each output change in spacing is first added with the old value of spacing between the antenna array elements denoted by d and as a result new value of spacing between the antenna array elements is generated which is denoted by d_{new} shown in equation (6), which is put into the formula of the array factor and as a result we obtain the synthesised radiation pattern of the antenna array.

IV. Calculation Used

Among the Piecewise Linear membership function trapezoidal function and triangular function are easy to implement thus we have applied the trapezoidal and triangular function in our work as given in equation (1) and equation (2) and the corresponding formula used are as follows:

Triangular membership function is defined as:

$$\mu_A(x) = \max [\min \{\min(x-a/b-a, c-x/c-b), 1\}, 0] \quad (1)$$

Trapezoidal membership function is defined as:

$$\mu_A(x) = \max [\min \{\min(x-a/b-a, d-x/d-c), 1\}, 0] \quad (2)$$

Now in the process of defuzzification since we have used Tsukamoto fuzzy model and the Weighted Average methods for calculating the defuzzified value as written in equation (3), thus the formula we have used in this case is the formula for calculating the Weighted Average which as follows:

$$Defuzzified\ output\ (Z) = (W_1 Z_1 + W_2 Z_2)/(W_1 + W_2) \quad (3)$$

Where:

W_1 and W_2 = represent the value of the membership function obtained at the output graph for spacing.

and

Z_1 and Z_2 = represent the corresponding spacing value obtained from the graph of output spacing for W_1 and W_2 respectively.

Now in our main work that is to synthesise the antenna array radiation pattern we at first have implemented the array factor formula for the linear antenna array. The elements are consider as isotropic point source with initial spaced $d = \lambda/2$ and mutual coupling is not taken into account [20]. The array factor of a linear array of N elements placed along the z – axis is given by:

$$S(\theta) = \sum_{i=0}^{N-1} I_i e^{j(i\beta d \cos\theta + \Delta\varphi_i)} \quad (4)$$

Where:

I_i = Amplitude of each element

$\Delta\varphi_i$ = Phase of each element

Now to get the new synthesise the antenna array radiation pattern we have modified the array factor of N element placed along the z – axis as given in equation (5)

$$S(\theta) = \sum_{i=0}^{N-1} I_i e^{j(i\beta d_{new_i} \cos\theta + \Delta\varphi_i)} \quad (5)$$

where

d_{new_i} = represent the new spacing between i- number of elements of the antenna array formed by the sum of change in spacing obtained from the fuzzy logic controller and the old spacing value.

Mathematically,

$$d_{new_i} = d + \Delta d_i \quad (6)$$

where

Δd_i = represent the change in spacing between i- number of elements of the antenna array

V. Result Analysis

Our work has been conducted using the MATLAB software. We have done our work by written program in

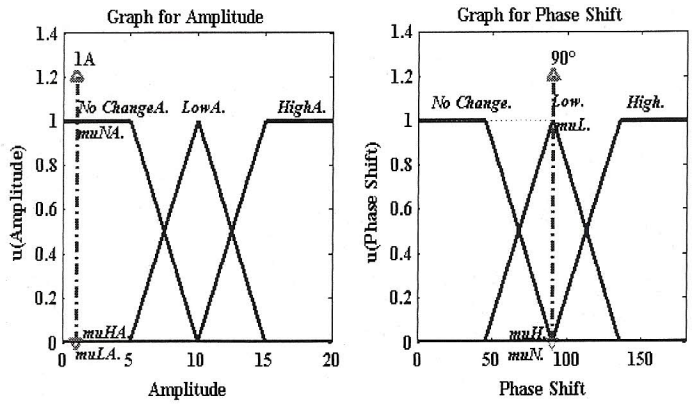


Fig. 6 Output graph for the fuzzy logic controller input simulated from the MATLAB Software

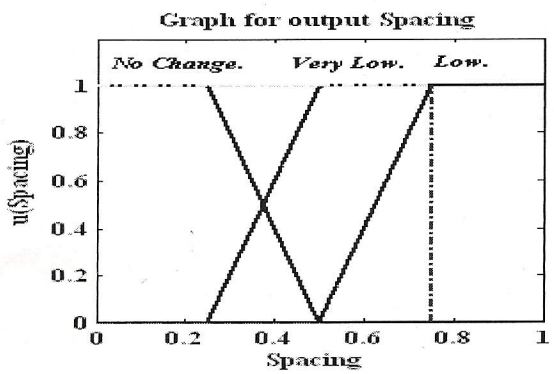


Fig. 7 Output graph for defuzzifier output for change in spacing $\Delta d(1)$ simulated from the MATLAB Software

MATLAB m.file. In the graph shown in Fig. 6, for this we have applied the input amplitude equal to 1Amp and phase shift (α) equal to 90° and as a result we are getting the output from the fuzzifier as μ (No ChangeA) at 1 and μ (Low) at 1 from the graph of linguistic variable of amplitude and phase shift respectively shown in Fig. 8. Now based on this result the inference engine by consulting with the rule base for change in spacing $\Delta d(1)$ generates the result for the fuzzified output change in spacing for $\Delta d(1)$ at μ (Low) equal to 1. Now according to the method of calculation of the defuzzified value in Tsukamoto fuzzy model, it will select the corresponding spacing value at 0.75 as illustrated in Fig. 7.

INPUT.
The Amplitude A=1Amp
The Phase Shift α=90°

FUZZIFIED INPUT.

muNA=1 **muN=0**
muLA=0 **muL=1**
muHA=0 **muH=0**

Fig. 8 Output result of the fuzzifier simulated from the MATLAB Software

Now by using the weighted average method in the program made by us generates the result output change in

W1=1
W2=0

d1=0.75
d2=0.25

CALCULATIONS CARRIED OUT.
I=(W1*d1+W2*d2)/(W1+W2).

Output Spacing =(1*0.75+0*0.25)/(1+0)

The Output Change in Spacing Δd =0.75

Fig. 9 Output result of the Tsukamoto fuzzy logic controller simulated from the MATLAB Software

spacing Δd equal to 0.75 as shown in Fig. 9. Similarly we have calculated the output change in spacing between all the elements of the antenna array that is for $\Delta d(2)$ to $\Delta d(19)$. Different rule base also been made for different change in spacing that is for $\Delta d(2)$ to $\Delta d(19)$, and the corresponding defuzzifier output for change in spacing are taken out from the fuzzy logic controller for $\Delta d(2)$ to $\Delta d(19)$. We have previously used d equal to 3 since we have taken $d = \lambda/2$ and λ equal to 6 in the array factor formula equation (4). As a result we obtained previously the value of antenna side lobe level of the antenna radiation pattern at -13.21 dB in connection with the value of phase shift at 81.9° as shown in Fig. 11. Now after putting the value of d equal to 4.5 which is obtained by using the Mamdani [21] and Tsukamoto fuzzy logic controller and uniform between all the elements of the antenna array we obtain the value of side lobe level of antenna radiation pattern at -13.5 dB for the value of phase shift at 84.6° [22] as shown in Fig. 12. Thus it has been seen that there is a very slight reduction in the side lobe level of the antenna radiation pattern. By varying the spacing between each element of the antenna array that is implementing non – uniform spacing which is obtained using Mamdani fuzzy logic controller we obtain the value of side lobe level of antenna radiation pattern at -15.2 dB for the value of phase shift at 83.34° as shown in Fig. 13.

TABLE III
RESULT ANALYSIS OF ANTENNA ARRAY SYNTHESIS USING DIFFERENT FUZZY MODEL

Side Lobe Level obtained	Type of Fuzzy Logic Controller Use to Obtain New Spacing	
	Mamdani	Tsukamoto
By Uniform Spacing	**-13.5 dB**	**-13.5 dB**
By Non – Uniform Spacing	**-15.2 dB**	**-16 dB**

Now varying the spacing between each element of the antenna array that is implementing non – uniform spacing we obtain the value of side lobe level of antenna radiation pattern

at -16 dB for the value of phase shift at 81.9° as shown in Fig. 14, so from the figure it has been seen that there is more reduction of the side lobe level of the antenna radiation pattern as compared to side lobe level reduction using uniform spacing whose value is also obtained using Mamdani and Tsukamoto fuzzy logic controller. Fig. 10 shows the new spacing value between each element of the antenna array that is obtained using implementation of the Tsukamoto fuzzy logic controller and implemented in the antenna array factor.

The New Spacing dnew=(d+Δ d) are

d(1) =3.75	d(6) =3.25	d(11) =3.5	d(16) =3.5
d(2) =3.5	d(7) =3.5	d(12) =3.75	d(17) =4.5
d(3) =3.25	d(8) =3.75	d(13) =3.75	d(18) =5.25
d(4) =3.75	d(9) =3.25	d(14) =5.5	d(19) =5.5
d(5) =3.25	d(10) =4	d(15) =5.5	

Fig. 10 Output result for new spacing dnew are simulated from the MATLAB Software

VI. CONCLUSION

By taking into account the data given in the Table. II it has been seen that there has been a very slight reduction of the side lobe level of the antenna array radiation pattern that is 0.29 dB which is obtained by implementing uniform spacing that is obtained using the implementation of Mamdani and Tsukamoto fuzzy logic controller. Now after implementing antenna array synthesis by varying the spacing between each element of the antenna array using fuzzy logic [23] where Mamdani fuzzy model is used it has been seen that side lobe level of antenna array radiation pattern reduces to 1.99 dB, but while implementing space perturbation between the element

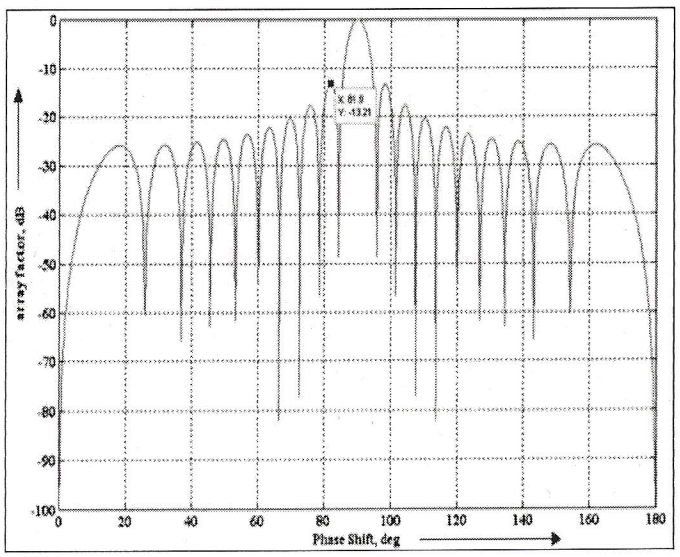


Fig. 11 Output graph for the antenna radiation pattern for 20 elements antenna array and for value of $d = 3$ simulated from the MATLAB Software

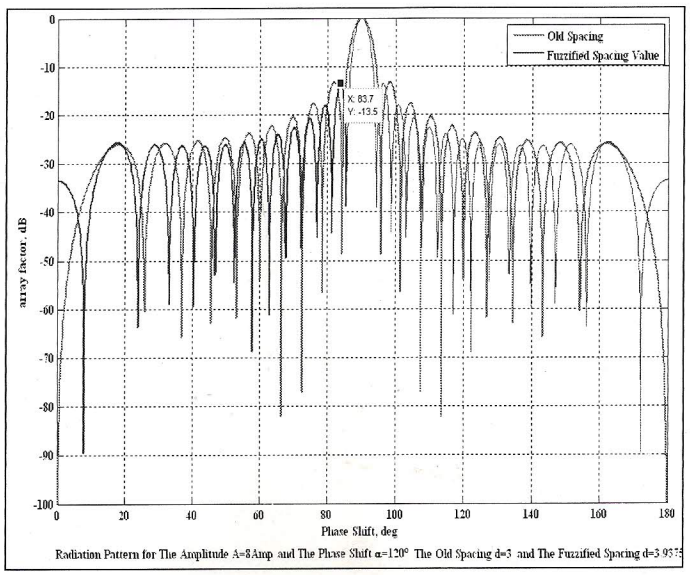


Fig. 12 Output graph for the antenna radiation pattern for 20 elements antenna array and both for value of $d = 3$ and the fuzzified output value of $d = 4.5$ simulated from the MATLAB Software. The red line in the graph showing the antenna array radiation pattern for the value of $d = 3$ and the blue line in the graph showing the antenna array radiation pattern for the value of $d = 4.5$ obtained using the fuzzy logic.

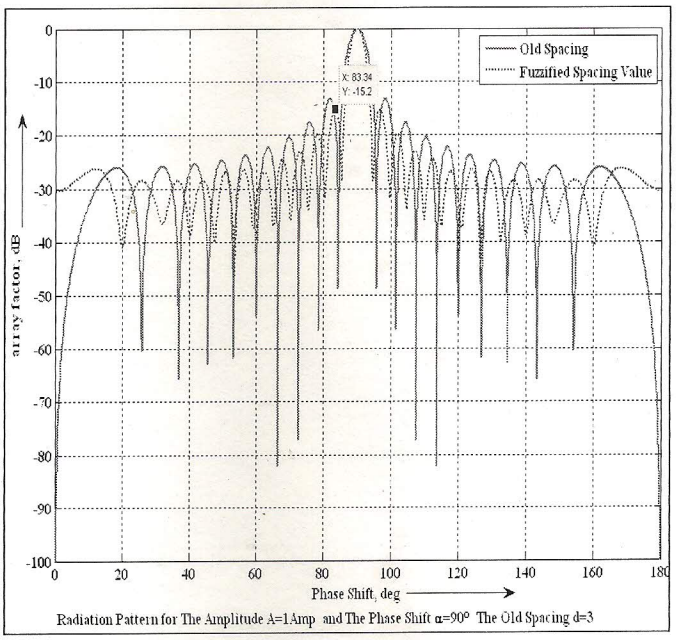


Fig. 13 Output graph for the antenna radiation pattern for 20 elements antenna array and both for value of spacing between the antenna array $d = 3$ and for the value of new spacing formed by combination of old value of spacing d and the fuzzified output value of change in spacing between each element of the antenna array simulated from the MATLAB Software. The red line in the graph showing the antenna array radiation pattern for the value of $d = 3$ and the blue line in the graph showing the antenna array radiation pattern for the value of new spacing formed by combination of old value of spacing d and the fuzzified output value of change in spacing between each element of the antenna array obtained using the Mamdani fuzzy logic model.

of the antenna array which is generated by combination of old spacing and the change in spacing value, obtained using

Tsukamoto fuzzy logic controller there is more reduction in the side lobe level of the antenna radiation pattern, and side lobe gets reduced to 2.79 dB as compared to the side lobe level obtained by putting the value of *d* equal to 6. Thus it can be stated that this antenna array synthesis by varying the spacing between each element of the antenna array using Tsukamoto fuzzy logic controller can be implemented in reducing the side lobe level of the antenna array radiation pattern, possessing null in certain direction, narrowing beam width of the antenna array radiation pattern in the future work. By proper implementation of the fuzzy logic linguistic variables, strong rule base we may further reduce the side lobe level of the antenna array radiation pattern. By proper implementation of the fuzzy logic linguistic variables it is meant that to tune the range of the linguistic it is meant that to tune the range of the linguistic variables that is to modify the range of the linguistic variables which we have implemented in our work. If we tune the range of linguistic variables it can generate the better value for the membership function and as result we can have the better output from the defuzzifier. By strong rule base it is meant that to generate the rule base with more number of data from the experts on the relative field.

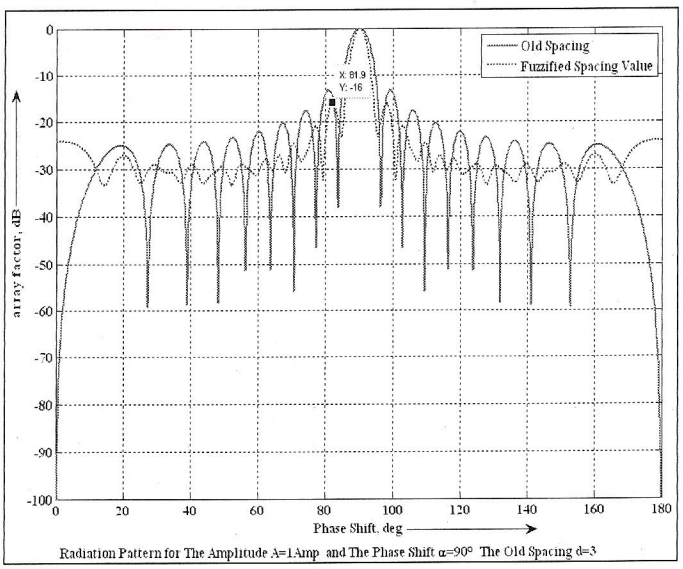


Fig. 14 Output graph for the antenna radiation pattern for 20 elements antenna array and both for value of spacing between the antenna array $d = 3$ and for the value of new spacing formed by combination of old value of spacing *d* and the fuzzified output value of change in spacing between each element of the antenna array simulated from the MATLAB Software. The red line in the graph showing the antenna array radiation pattern for the value of $d = 3$ and the blue line in the graph showing the antenna array radiation pattern for the value of new spacing formed by combination of old value of spacing *d* and the fuzzified output value of change in spacing between each element of the antenna array obtained using the Tsukamoto fuzzy logic model.

ACKNOWLEDGMENT

For our work we wish to acknowledge the Department of Electronic and Communication Engineering of National Institute of Technology, Durgapur for helping us through out the project by giving us the good laboratory facility to continue our work in this field of Antenna Array Synthesis by Implementing Non – Uniform Spacing Using Tsukamoto Fuzzy Logic Controller.

REFERENCES

[1] C. Balanis, *Antenna Theory—Analysis and Design*. New York: Wiley,2011

[2] Aritra Chowdhury, Ritwik Giri, Arnob Ghosh, Swagatam Das, Ajith Abraham and Vaclav Snasel, *"Linear antenna array synthesis using fitness-adaptive differential evolution algorithm"* , IEEE Congress on Evolutionary Computation (CEC), 2010, pp.1-8, July 2010.

[3] J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Harbor, 1975.

[4] T. Bäck, D. Fogel, Z. Michalewicz, *Handbook of Evolutionary Computation*, Oxford Univ. Press.

[5] A.E. Eiben and J.E. Smith, Introduction to Evolutionary Computing, Springer, 2003.

[6] S. Kirkpatrik, C. Gelatt, and M. Vecchi, *Optinization by Simulated Annealing*, Science, 220: 671-680, 1983.

[7] F. Glover and M. Laguna, *Tabu Search*, Kluwer, Norwell, MA, 1997.

[8] G. Taguchi, S. Chowdhury, and Y. Wu, *Taguchi's Quality Engineering Handbook,* New York: Wiley, 2005.

[9] Y.-S. Ong, and A.J. Keane, *"Meta-lamarckian learning in memetic algorithms,"* IEEE Transactions on Evolutionary Computation, vol. 8, no. 2, pp. 99-110, 2004.

[10] J. Kennedy and R. Eberhart,*"Particle swarm optimization,"* Proc. IEEE Int. conf. Neural Networks, pp. 1942-1948, 1995.

[11] J. Kennedy, R.C. Eberhart, and Y. Shi, *Swarm Intelligence,* Morgan Kaufmann, San Francisco, CA, 2001.

[12] Kadri, B., Bendimered, "Linear *Array Synthesis with a Fuzzy Genetic Algorithm"*, The International Conference on EUROCON, 2007, pp. 942-947, Sept, 2007.

[13] Ahmidi, N. Neyestanak, A.A.L. Dawes, "*Elliptical Array Antenna Design Based on Particle Swarm Method Using Fuzzy Decision Rules"*, 24th Biennial Symposium on Communication, 2008, pp. 352-355, June 2008.

[14] Om Prakash Acharya, Amalendu Patnaik and Balamati Choudhury, "Fault *Finding in Antenna Arrays Using Bacteria Foraging Optimization Technique*", National Conference on Communication (NCC) 2011, pp. 16-31, Jan. 2011.

[15] L.A. Zadeh, "*Fuzzy sets*", Information and Control, 8, pp.338 – 353, 1965.

[16] Sanmoy Bandyopadhyay, B. Maji, "*Antenna Array Synthesis Using Tsukamoto Fuzzy Logic Controller*", National Conference on Emerging Trends in Electronics & Telecommunication Engineering (ETET - 2012), pp. 1, January 27, 2012.

[17] Surajit Kundu, Sanmoy Bandyopadhyay, and N.R. Das, "*Design of a Controller Using Fuzzy Logic in Temperature Control System"*, 2nd National Conference on Engineering Education in the New Century (E2NC) 2011, pp. 34-37, January 2011.

[18] Y. Tsukamoto, "*An approach to fuzzy reasoning method*", pp 137-149, North-Holland, Amsterdam, 1979.

[19] Lee C.C., "Fuzzy *Logic in Control Systems: Fuzzy Logic Controller* ", IEEE Transactions on Systems, Man and Cybernetics, 20 (2), pp. 404-435, 1990.

[20] D. Marcano, F. Duran," *Synthesis of antenna arrays using genetic algorithms"*, IEEE Antennas and Propagation Magazine, Volume (42): Issue (3), pp. 12-20, June, 2000.

[21] E. H. Mamdani, S. Assilian, "*An experiment in linguistic synthesis with a fuzzy logic controller*", International Journal of Man – Machine Studies, 7, pp. 1 – 13, 1975.

[22] Sanmoy Bandyopadhyay, B. Maji, "*Antenna Array Synthesis Using Fuzzy Logic*", Second National Conference on Computing and Systems (NaCCS - 2012), pp. 93 - 99, March 15 - 16, 2012.

[23] Sanmoy Bandyopadhyay, B. Maji, "*Antenna Array Synthesis by Varying the Spacing Between Each Element of the Antenna Array Using Fuzzy Logic*", International Conference on Information Technology, Electronics and Communications (ICITEC - 2012), pp. 185 - 191, March 3 - 4, 2012.

Two Level Data Hiding Scheme: Using Shape Based Cryptography And MIB Steganography Technique

Smt .M.SreeLatha[1],
HOD, Department of CSE,
RVR&JC College of Engineering,
Guntur, India
E-mail: varsha@rvrjcce.ac.in

Smt.K.Venkata Ramana[2],
Assistant Professor,,
RVR&JC College of Engineering,
Guntur, India
E-mail: vinoliamanohar@gmail.com

Sushma Vemulapalli[3],
Computer Science and Engineering,
RVR&JC College of Engineering,
Guntur, India
E-mail: sushmavemulapalli.2k9@gmail.com

Mangisetty Sunitha[4],
Computer Science and Engineering,
RVR&JC College of Engineering,
Guntur, India

Abstract— **Data hiding in the internet world is more vulnerable to attacks. In this paper, we are proposing a new technique which implements data hiding using a two level hierarchy scheme. In the first level a new cryptographic technique named Shape Based technique is used for encrypting the data. In this method secret key letters and a 5x5 matrix are used to encrypt the secret message. The secret key is known only to the sender and the receiver. In the second level the encrypted message is embedded in a cover image using a new steganography technique named MIB technique. The stego key used in the above process is transmitted in a secure channel to the receiver. The main advantage of these techniques over existing techniques is that it provides double security for the sender's message in two levels. The analyst must know the alphabets and the size of the alphabets shape and the stego key for finding the actual message.**

Keywords- Cryptography; Steganography; Stego key; Cryptanalyst; Stegaanalysis

I. INTRODUCTION

Since the rise of the Internet one of the most important factors of information technology and communication has been the security of information. Cryptography [1] is the practice and study of techniques for secure communication in the presence of adversaries. Two main categories of cryptography are symmetric cryptography and asymmetric cryptography. In this paper we have proposed a new symmetric cryptographic technique for encrypting the sender message. Cryptanalysis refers to the study of ciphers, cipher text, or cryptosystems (that is, to secret code systems) with a view to finding weaknesses in them that will permit retrieval of the plaintext from the cipher text, without necessarily knowing the key or the algorithm. A Cryptanalyst is a decoder skilled in the analysis of codes and cryptograms.

Our proposed technique provides good strength against cryptanalysis by using the properties of key letters like shape and size in encrypting the secret message. This method produces a variable length cipher texts for the plain text depending upon the agreed properties of the key between the sender and the receiver. This property helps in creating confusion to the cryptanalyst.

It is sometimes not enough to keep the contents of a message secret, it may also be necessary to keep the existence of the message secret. The technique used to implement this, is called steganography. Steganography [1] is the art of writing hidden messages in such a way that no one apart from sender and intended recipient, suspects the existence of message. While cryptography is about protecting the content of messages, steganography is about concealing their very existence [2] .Secret Key Steganography takes a cover message and embeds the secret message inside of it by using a secret key (stego-key) [3]. The benefit to Secret Key Steganography is even if it is intercepted; only parties who know the secret key can extract the secret message [4]. A successful attack on a steganographic system consists of an adversary observing that there is information hidden inside a file. In our proposed technique as a second level of security we are using a new technique called MIB steganography technique. This technique hides the secret message in a cover image. This method considers the intensities of the pixel values when embedding the secret data. To reduce the distortion of the cover image changing high intensity values is the optimal choice. We have used this principle in our proposed method for embedding the secret message.

Steganalysis [5] is the art and science of detecting messages hidden using steganography. The goal of steganalysis [6] is to identify suspected packages, determine whether or not they have a payload encoded into them, and, if possible, recover that payload. The steganalysis for our method is difficult because the embedding process is random and uses high intensity pixels for embedding. The proposed technique provides double security for hiding and transferring of data. This method focuses on developing a

new system with combining security techniques like cryptography and steganography [7].

II. Related Work

In this section we are giving a brief overview of already existing steganography techniques for embedding the secret data in the images.

A. Least Significant Bit

This method is the basic technique for hiding data in the image. Least Significant Bit Hiding method is probably the easiest way of hiding information in an image and yet it is surprisingly effective. It works by using the least significant bits of each pixel in one image to hide the most significant bits of another [8]. So in a JPEG image for example, the following process would need to be taken. First load up both the host image and the data you need to hide. Next chose the pixel positions and hide the data in the least significant position of the pixel value. To get the original image back you just need to know the pixel positions and retrieve the data from the LSB positions.

Hiding depends on the settings you choose - but as an example if we hide in the 2 least significant bits then, we can hide:

MaxBytes = (image.height () * image.width () * 3* 2) /8

i.e. the number of pixels, times the number of colors (3), times the number of bits to hide in, all divided by 8 to get the number of bytes.

Following are few other algorithms currently implemented, each use least significant bit steganography and some filter the image first.

B. Blind Hide

This is the simplest way to hide information in an image. It blindly hides because it just starts at the top left corner of the image and works it's way across the image (then down - in scan lines) pixel by pixel [9]. As it goes along it changes the least significant bits of the pixel colors to match the message. To decode the process the least significant bits starting at the top left are read off. This is not very secure - it's really easy to read off the least significant bits. It also isn't very smart - if the message doesn't completely fill up the possible space then just the top part of the image is degraded but the bottom is left unchanged - making it easy to tell what's been changed.

C. Filter First

This algorithm filters the image using one of the inbuilt filters and then hides in the highest filter values first[10]. It is essentially a fancier version of Blind Hide as it doesn't require a password to retrieve the message. Because we are changing the pixels we need to be careful about filtering the picture because we don't want to use information for filtering that might change. If we do, then it may be difficult (if not impossible) to retrieve the message again. So this algorithm filters the most significant bits, and leaves the least significant bits to be changed. It is less noticeable on an image because using the filter ensures we are hiding in the parts of the image that are the least noticeable.

III. Two Level Data Hiding Scheme

This section we have proposed a new technique called two level data hiding scheme. The first level is applying cryptography to covert the secret message into unintelligent form using a novel cryptography method called shape based cryptography. The second level is applying steganography method. A new steganography technique named MIB steganography technique is proposed to embed the first level output into a cover image. The proposed steganography technique hides data in the high intensity pixel values. By using high intensity values for hiding data produces a cover image with less distortion which makes the work of the steganalyst difficult.

The process is divided into sender side algorithms and receiver's side algorithms. The detailed explanation of senders side methods are given in section A. The methods used at receiver's side are discussed in section B.

A. Sender Side

As discussed earlier two levels of security is applied to the data at sender side. They are

Level1: Shape based cryptography

Level2: MIB steganography technique.

The complete view of sender side process is shown below in Figure 1.

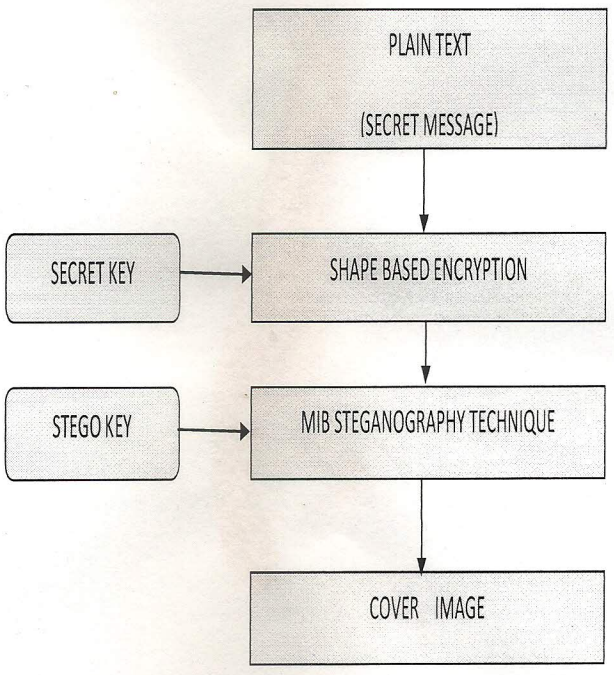


Figure 1. Sender's Side Process

The secret message is encrypted using the shape based encryption technique which produces cipher text as output. The cipher text is embedded in the LSB of some high

intensity pixels in cover image using MIB embedding technique.

Shape based cryptography

A 5x5 matrix is used for encrypting the secret message, which contains only binary values. The secret message is converted to binary form and placed in the matrix using the shape of the alphabets that form the key which is known only to both the sender and receiver. The key may contain N number of alphabets. The remaining positions in the matrix are filled by binary values 1 and 0 randomly. The cipher text is the data in the 5x5 matrix. If the size of the secret message does not fit in one matrix then more than one 5x5 matrices are considered with the same key. That's why the size of the cipher text is always in multiples of 25(the elements in the 5x5 matrix).

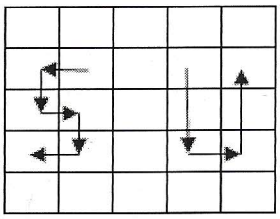


Figure 2. Key matrix with key alphabets 'S' and 'U'

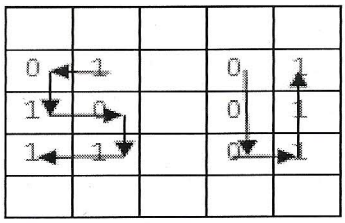


Figure 3. Secret message is arranged in the above Key matrix

1	0	1	1	0
0	1	0	0	1
1	0	1	0	1
1	1	1	0	1
0	1	1	0	1

Figure 4. Remaining positions in the key matrix are filled with random bit atream

This technique is explained by an example as follows: Let us assume a key "SU" and the secret message as "101011000111" .The key shape and size are taken as shown in Figure 2.

Now the secret message is placed in the positions of the matrix so that it exactly matches with the shape of the key as shown in the Figure 3. The remaining positions in the matrix are filled with binary values 1's and 0's randomly as shown in Figure 4. The resultant encrypted text is the 5x5 matrix read row wise i.e "10110010011010111110101101". Now this encrypted text is embedded in cover image using next level of security.

Embedding Process using MIB Steganographic Technique

MIB stands for "Mean Intensities of Blocks". This technique divides the cover image into blocks. For each block the mean intensity of the pixels is calculated. Then the blocks with the highest intensities values are considered for embedding.[10].Then stego key 'K' is added to the mean intensity of each block. Stego key is known only to sender and the receiver.

Since the size of encrypted message from first module is multiples of 25, the image is divided into 32 blocks. 25 blocks which contains the highest mean intensity values are selected for placing the message. If the encrypted message size is Mx25 then M bits are inserted in each block. The pixel in which the data is embedded is determined by the formula Mean_intensiy of the block + K- I where K is the stego key and I range from 0 to M-1.

The message is embedded in the LSB of each pixel byte. The text is placed in LSB of each pixel byte i.e. RGB. The distortion in the image is less as the pixels selected are high intensity pixels.

B. Receiver's side

The complete overview of the receiver's side process is shown in figure 3. The following levels are applied to recover the original message at receiver's side.

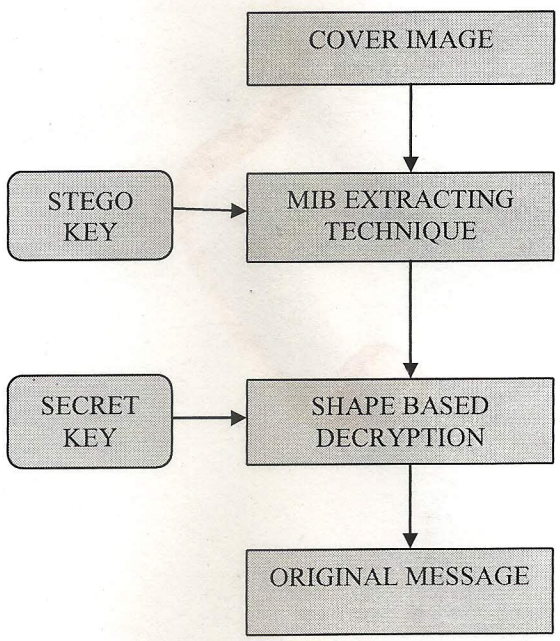


Figure 5. Receiver's Side Process

Level1: MIB extracting technique

Level2: Shape based decryption

MIB Extraction Technique

The extracting technique is done in reverse manner of embedding process. The image in which the message is embedded is sent to the receiver's side along with the factor K and the size of the message. To extract the encrypted message from the image first the image is divided into 32 blocks and the mean intensity of blocks is calculated. The 25 blocks with high intensity are determined. By applying the same formula using K and M values, the pixels used in each block for embedding the message are determined. Then the data is retrieved from the LSB of the selected pixels. The result of this level is encrypted data.

Shape Based Decryption

The data retrieved from the previous level is placed in arrays of 5x5 matrices using the same key used at sender side. The binary numbers are placed in the matrix row wise and the data at the keys are retrieved. This level takes encrypted data and decrypts it and produces the original text.

IV. RESULTS

A. *Experimental Results*

We have implemented our algorithm in java. To see the performance of our algorithm on different message sizes we have varied the size of the plaintext embedded in the image. Our results show that increasing the text size has increased the size of the cipher text. Our algorithm performs better for all sizes of text. The simulation results for two examples are shown below. Examples key matrices and the results after level1 are shown one after the other. In the example 1 - secret message size is 12 bits and the key alphabets are S and U. The key matrix with secret message arranged at key positions is shown below in the figure 6. The encrypted data after level 1 for various plaintexts are shown in the table 1.In the example 2 – secret message size is 9 bits and the key alphabets are K and L. The key matrix and results after level1 for various plaintexts are shown below in the figure7 and table 2 respectively.

N=12; key=SU;

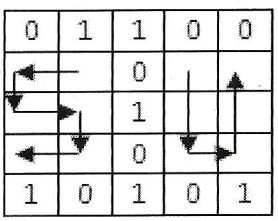


Figure 6. Key matrix with key alphabets 'S' and 'U'

TABLE I. RESULTS AFTER LEVEL 1 FOR EXAMPLE 1 WITH MESSAGE SIZE 12 AND KEY ALPHABETS 'S' AND ' U'

DATA IN THE FORM OF PLAIN TEXT	ENCRYPTED DATA AT FIRST LEVEL
110110110110	011001101001111010011 0101
011011011011	011001000110111111010101
100100100100	011000101001100000011 0101

N=9; key=KL;

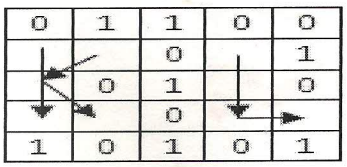


Figure 7. Key matrix with key alphabets 'K' and 'L'

TABLE II. RESULTS AFTER LEVEL1 FOR EXAMPLE 2 WITH MESSAGE SIZE 12 AND KEY ALPHABETS 'K' AND 'L'

DATA IN THE FORM OF PLAIN TEXT	ENCRYPTED DATA AT FIRST LEVEL
110110110	011001100110110010101 0101
011011011	011000001110100110111 0101
100100100	011001100100110000001 0101

After implementing MIB steganographic technique on the cover image or the original image the transmitted stego image is obtained. Both the original image and the stego image are shown below in the following figure 8.

Figure 8. The image before and after MIB steganography technique.

B. *Analysis and Comparison*

Our proposed system provides hierarchical security in various phases. It helps in creating mystification to the analyst about which part of the image actually contains secret data. If the analyst succeeded in identifying the block in the image also, it will be very difficult to retrieve the secret data from the block because the data is embedded using key which is not known to the analyst. The encryption process adds padding to the secret message by mapping the cipher text as multiples of 25, which makes the analyst difficult to predict size of the message.

V. CONCLUSION

Since our proposed scheme uses two levels it provides double security to the data. This scheme provides data confidentiality and authentication. There are many real time applications that are based on this. This is used for transferring of secret documents in military applications etc. Steganography, especially combined with cryptography, is a powerful tool which enables people to communicate without possible eavesdroppers. The proposed method provides acceptable image quality with very little distortion in the image.

REFERENCES

[1] Domenico Daniele Bloisi , Luca Iocchi, "Image based Steganography and cryptography", Computer Vision theory and applications volume 1 , pp. 127-134, 2007.

[2] Dipti Kapoor Sarmah , Neha Bajpai, "Proposed System for Data Hiding Using Cryptography and Steganography", International Journal of Computer Applications (0975 – 8887) Volume 8– No.9, October, 2010.

[3] Kharrazi.M, Sencar.H.T and Memon.N, "Image Steganography: Concepts and practice", in WSPC Lecture Notes Series, 2004.

[4] Dunbar.B, "Steganographic techniques and their use in an Open-Systems environment", SANS Institute, January, 2002.

[5] N. F. Johnsonand S. Jajodia, "Steganalysis: The investigation of hidden information," in IEEE Information Technology Conference, New York, September,1998.

[6] Wang. H & Wang. S, "Cyber warfare: Steganography vs. steganalysis", Communications of the ACM, 47:10, October 2004.

[7] Artz.D, "Digital Steganography: Hiding Data within Data", IEEE Internet Computing Journal, June, 2001.

[8] M. Wu and B. Liu, "Multimedia Data Hiding", New York:Springer-Verlag, 2003.

[9] Chun-Hsiang Huang, Shang-Chih Chuang, and Ja-Ling Wu, "Digital-Invisible-Ink Data Hiding Based on Spread- Spectrum and Quantization Techniques", IEEE Transactions on Multimedia, VOL.10, NO. 4, June, 2008.

[10] Khosravi Sara, Abbasi Dezfouli, Mashallah, Yektaie Mohammadi Hossein, Khosravi,moien, "A New Method to Steganography Whit Processing Picture in Three Colors (RGB)", International Journal of Computer Technology and Applications, Vol 2 , pp. 274-279, March-April, 2011.

Ranking and Suggesting Popular Itemsets in Mobile Stores Using Modified Apriori Algorithm

Phani Kishore Rompicharla, Ragini Mokkapati & Surapaneni Praneetha

Department of Computer Science and Engineering, Dhanekula Institute of Engineering & Technology, Ganguru, Vijayawada, India
E-mail : phani.rompicharla@gmail.com, raginidhanekula@yahoo.in, surapaneni.praneetha@gmail.com

Abstract - We considered the problem of ranking the popularity of items and suggesting popular items based on user feedback. User feedback is obtained by iteratively presenting a set of suggested items, and users selecting items based on their own preferences either the true popularity ranking of items, and suggest true popular items the difficulty is that making suggestions to users can reinforce popularity of some items and distort the resulting item ranking. The described problem of ranking and suggesting items arises in diverse applications including search query suggestions and tag suggestions for social tagging systems. We propose and study several algorithms for ranking and suggesting popular items, provide analytical results on their performance, and present numerical results obtained using the inferred popularity of tags from a month-long crawl of a popular social bookmarking service. Our results suggest that lightweight, randomized update rules that require no special configuration parameters provide good performance.

I. INTRODUCTION

1.1 TERMINOLOGY:

In this section we first want to introduce the different terms that we were going to use in our paper as fallows.

1.1.1 Ranking: Ranking is giving rank scores to the most popular item by taking user feedback. The most frequently occurring item is given the highest rank score.

1.1.2 Selection: We focus on the ranking of items where the only available information is the observed selection of items. In learning of the users preference over items, one may leverage some side information about items, but this is out of the scope of this paper.

1.1.3 Imitate: The user study was conducted in very famous mobile stores and which has been used to set of mobiles. The user may check the list and select the set of mobiles which they like most and depending on those like results the new suggestion list has been developed by the algorithm.

1.1.4 Popular: In practice, one may use prior information about item popularity. For example, in the survey the user may select the suggested mobile or they may also select the others. If they selected the already suggested items they will become more popular and if he don't they may get out of the popular list.

1.1.5 Association Rule: Association Rules are if/then statements that help uncover relationships between seemingly unrelated data in the relational database or other information repository. An example of an association rule would be **if a customer buys a nokia mobile, he is 70% interested in also purchasing nokia accessories**.

II. THEORETICAL STUDY

We consider the mobile phone selection and suggesting the best sold mobile and their combinations that were most liked by most of the users. Consider a set of mobiles M: (m1, m2, m3, m4, ….mn) where n > 1. Now we were calculating the set of items in C where were mostly sold and mostly liked by the users, as S

S: (s1, s2, s3, s4, …. sg) where g > 1.

We need to consider an item I, we interpret si as the portion of users that would select item i if suggestions were not made. We assume that the popularity rank scores s as follows

a) Items of set S were estimated to is as $s1 \geq s2 \geq s3 \geq$ …. sc,

b) s is completely normalized such that it is a probability distribution, i.e., s1 + s2 + s3 + …. +sc = 1.

c) si is always positive for all items i.

III. PROPOSED ALGORITHM AND STUDY

We have some of the systems already existing in the same field and we have also identified some of the disadvantages in them as follows:

- The popularity for any item is given based on the production of that item. This may not give good result because customers may not have the knowledge of true popularity they needed and depend on the results given by the producer.
- The updates are performed regardless of the true popularity by virtual analysis.
- Producer have to analyse things manually and complexity involves in this. Due to this time consumption may be high.
- The algorithms used in this system may fail to achieve true popularity.

We consider the problem learning of the popularity of items that is assumed to be apriori unknown but has to be learned from the observed user's selection of items. We have selected an mobile market and mobile distribution outlets as our data set and examined them completely in all areas where we can give the list of items suggested by the users and we have made an web-application to make an survey at real-time and considered the data given by more that 1000 members of different categories of people and applied our proposed apriori algorithm on the set of data and started suggesting the item in the mobile outlets for the actual users, which had helped the mobile phone companies and also the outlet in-charges. We have implemented the same in some of the mobile outlets in INDIA wehre we got very good results. The actual goal of the system is to efficiently learn the popularity of items and suggest the popular items to users. This was done to the user to suggest them the mostly used mobiles and their accessories, such that they also buy the best and at the same time the outlet owner will also get benefited. The most important feature in our project is suggesting the users by refreshing the latest results every time the user gives the input and changes his like list.

Now we have overcome many of the disadvantages of the existing systems and achieved many advantages with the proposed algorithm and method as follows:

- In our approach, we consider the problem of ranking the popularity of items and suggesting popular items based on user feedback.
- User feedback is obtained by iteratively presenting a set of suggested items, and users selecting items based on their own preferences either from this suggestion set or from the set of all possible items.
- The goal is to quickly learn the true popularity ranking of items and suggest true popular items.
- In this system better algorithms are used. The algorithms use ranking rules and suggestion rules in order to achieve true popularity.

IV. PROPOSED ALGORITHM

APRIORI ALGORITHM:

This is to find frequent item-sets using candidate generation. Apriori employs an iterative approach known as level-wise search, where k-itemsets are used to explore (k+1) itemsets. First, the set of frequent 1-itemsets is found by scanning database to accumulate the count for each item and collecting those items that satisfy minimum support. The resulting set is denoted by L1. Next, L1 is used to find L2,the set of frequent 2-itemsets,which is used to find L3 and so on, until no more frequent k-itemsets can be found. The finding of each Lk requires one full scan of database.

To improve the efficiency of the level-wise generation of frequent itemsets, an important property is called apriori property, which is used to reduce search space.

Apriori property:

- All non empty subsets of a frequent itemset must also be frequent.
- Antimonotone property-if a set cannot pass a test, all of its supersets will fail the same test as well.

Apriori algorithm is a two step process, the two steps are

- Join step
- Prune step

Join step:

To find Lk, a set of candidate k-itemsets is generated by joining Lk-1 with itself. This set of candidates is denoted by Ck. Let l1 and l2 be itemsets in Lk-1.The notation li[j] refers to jth item in li. By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. For the (k-1) itemset, li, this means that the items are sorted such that li[1]<l1[2]<……..<li[k-1].The join Lk-1∞Lk-1, is performed, where members of Lk-1 are joinable if their first (k-2) items are in common.

Prune step:

Ck is a superset of Lk, that is, its members may or may not be frequent but all of frequent k-itemsets are included in Ck. A scan of database to determine the count of each candidate in Ck would result in the

determination of Lk. Ck however can be huge. so this could involve heavy computation. To reduce the size of Ck, Apriori property is used as follows:

Any (k-1) itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (k-1) subset of a candidate k-itemset is not in Lk-1, then the candidate cannot be frequent either and so can be removed from Ck. This subset testing can be quickly done by maintaining a hash tree of all frequent itemsets.

Improving the efficiency of apriori by proposing it with some variations. Several variations are summarized as follows:

- Hash-based technique (hashing itemsets into corresponding buckets):A hash-based technique can be used to reduce the size of candidate k-itemsets, Ck, for k > 1.
- Transaction reduction (reducing the number of transactions scanned In future iterations):A transaction that doesn't contain any frequent k-itemsets cannot contain any frequent (k+1) itemsets. Therefore, such a transaction can be marked or removed from further consideration because subsequent scans of database for j-itemsets where j > k, will not require it.
- Partitioning(Partitioning the data to find candidate itemsets):A partitioning technique can be used that requires just two database scans to mine the frequent itemsets. It consists of two phases. In phase I ,the algorithm sub divides the transaction of D into n overlapping partitions. If the minimum support threshold for transactions in D is min-sup, then the minimum support count for a partition is min-sup x the number of transactions in that partition. For each partition, all frequent itemsets within the partition are found. These are referred to as local frequent itemsets. In phase II, a second scan of D is conducted in which the actual support of each candidate is accessed in order to determine the global frequent itemsets.
- Sampling: The basic idea of the sampling is to pick a random sample S of given data D and then search for frequent itemsets in S instead of D.
- Dynamic itemset counting: In a dynamic itemset counting technique, new candidate itemsets can be partitioned into blocks by start points. The technique is dynamic in that it estimates the support of all of the itemsets that have been counted so far, adding new candidate itemsets if all of their subsets are estimated to be frequent.

Algorithm:

Apriori :

Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

D, a database of transactions;

min-sup, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

```
(1)  L1=find_frequent_1_itemsets(D);
(2)      for(k=2;Lk-1≠Ø;k++)
(3)      {
(4)          Ck = apriori_gen(Lk-1);
(5)          for each transaction t Є D
(6)          {
(7)           C1=subset(Ck,t);
(8)           for each candidate c Є Ct
(9)                c.count++;
(10)         }
(11) Lk= {c Є Ck | c.count ≥ min_sup}
(12)   }
(13) return L=UkLk;
```

```
Procedure apriori_gen (Lk-1 : frequent (k-1)-itemsets)
(1)      for each itemset l1 Є Lk-1
(2)          for each itemset l2 Є Lk-1
(3)   if(l1[1]=l2[1])^(l1[2]=l2[2])^…….^(l1[k-2]=l2[k-2])^(l1[k-1]<l2[k-1]) then
(4)   {
(5)          c=l1∞l2;
(6)          if (has_infrequent_subsets(c,Lk-1) then
(7)               delete c;
(8)          else add c to Ck;
(9)   }
(10)     return Ck;
```

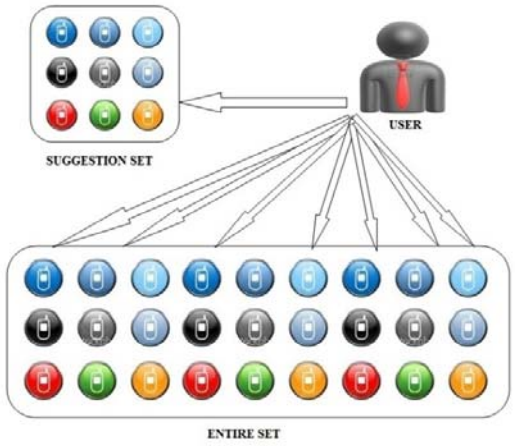
```
Procedure has_infrequent_subset (c:candidate k-itemset;
Lk-1:frequent(k-1) –itemsets);
(1)      for each (k-1) –subset s of c
(2)          If s Є Lk-1 then
```

(3) return TRUE;

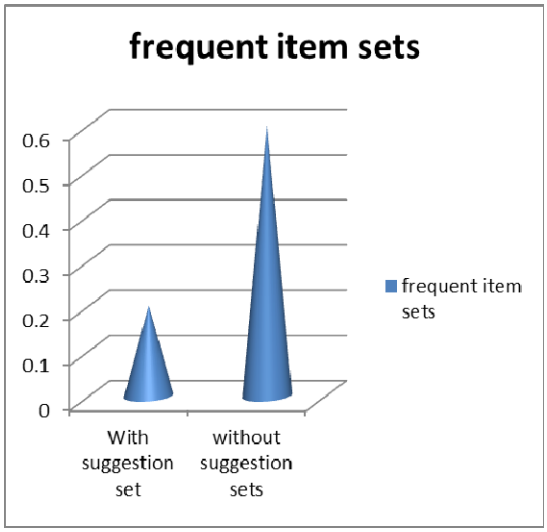
(4) return FALSE;

V. RESULTS

The above is the best method of ranking and suggesting the best methods in the scenario of mobile phone outlets in INDIA, which is shown in the following diagram



As it was shown in the above diagram we were going to take the most liked items from the users and suggesting the best mobiles or the best set of suggestions that the most of the users liked or ordered.



The confidence of the suggestions were also proved by an traditional confidence calculations as follows

In this section we are going to discuss about algorithms. Till now we have discussed some ranking rules , suggestion rules and Frequency move2set algorithm. We have some problems with these, so we go for an algorithm which suits our requirements well. The algorithm is Apriori algorithm. In order to know these algorithms we need to know some concepts of data mining.

Frequent itemsets:

Let I={I1, I2, I3,…., Im} be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that T is a subset of I. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if A is a subset of T. An association rule is an implication of the form $A \geq B$, where A is subset of I, B is subset of I and $A \cap B = \emptyset$. The rule $A \geq B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain AUB. This is taken to be the probability ,P(AUB).The rule $A \geq B$ has confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P(B/A). That is,

$$\text{Support}(A \Rightarrow B) = P(AUB)$$

$$\text{Confidence}(A \Rightarrow B) = P(B/A)$$

Rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are called strong. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply as the frequency, support count,or count of the itemset. The set of frequent k-itemset is commonly denoted by Lk.

$$\text{confidence}(A \geq B) = P(A / B) = \text{support}(AUB) / \text{support}(A) = \text{support}_{\text{count}(AUB)} / \text{support}_{\text{count}(A)}.$$

Mining frequent itemsets:

In general, association rule mining can be viewed as a two-step process:

1. Finding all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min-sup.
2. Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence

VI. CONCLUSION

All the previous process already proposed were very complex and contains very complicated computations which made the ranking and suggesting the best and popular items have been more and more complex and not getting to the actual end users. Now we have proposed as very simple randomized algorithm for

ranking and suggesting popular items designed to account for popularity bias. This was utilized by many of the mobile outlets in the country successfully.

REFERENCES

[1] Huidrom Romesh Chandra Singh, T. kalaikumaran, Dr. S. Karthik, Suggestion of True Popular Items, IJCSE, 2010.

[2] Y.Maanasa, V.Kumar, P.Satish Babu, Framework for suggesting POPULAR ITEMS to users by Analyzing Randomized Algorithms, IJCTA, 2011.

[3] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays—Part i: i.i.d. Rewards," IEEE Trans. Automatic Control, vol. 32, no. 11, pp. 968-976, Nov. 1987.

[4] J.R. Anderson, "The Adaptive Nature of Human Categorization" Psychological Rev., vol. 98, no. 3, pp. 409-429, 1991.

[5] Yanbin Ye, Chia-Chu Chiang, A Parallel Apriori Algorithm for Frequent Itemsets Mining, IEEE, 2006.

[6] Cong-Rui Ji, Zhi-Hong Deng, Mining Frequent Ordered Patterns without Candidate Generation.

[7] Huang Chiung-Fen, Tsai Wen-Chih, Chen An-Pin, Application of new Apriori Algorithm MDNC to Exchange Traded Fund, International Conference on Computational Science and Engineering, 2009.

[8] Milan Vojnovi_c, James Cruise, Dinan Gunawardena, and Peter Marbach, Ranking and Suggesting Popular Items, IEEE, 2009.

❑❑❑

A Comparative Study of TinyOS Scheduling Strategies and Future Scope

Vijay Borges, Omkar Raikar, Vinit Desai & Priya Dalvi

Dept. of Information Technology, Goa College of Engineering, Farmagudi Ponda, Goa, India
E-mail : vb@gec.ac.in, omkarraikar17@gmail.com, desaivinit9@gmail.com, dalvipriya09@gmail.com

Abstract **-** In the present era Wireless Sensor Networks (WSN) has gained a great value and importance due to its flexibility, cheaper implementation cost, mobility etc. The sensor networks are expected to play increasingly important role in future especially in monitoring and military applications on large scales. Designing an ideal operating system for a WSN has become difficult and a challenging task. TinyOS is one of the mostly used operating systems in this area. TinyOS features multithreading architecture, a very flexible networking stack, and virtual machine implementations. The paper first outlines the ideal characteristics of an ideal WSN OS followed by various scheduling strategies proposed in TinyOS. A thorough and comparative study of the algorithm has been carried out, listing each ones pros and cons. Finally we point out open research issue with regard to the TinyOS scheduling policies.

Keywords **-** *WSN; TinyOS; Scheduling;FIFO;SJP,EDF;FPS;*

I. INTRODUCTION

A wireless sensor network (WSN) consists of spatially distributed autonomous sensors to monitor physical or environmental conditions, such as temperature, pressure humidity, pollutants, etc and to cooperatively pass their data through the network to a main location.

Power is the scarcest resource of WSN nodes, and it determines the lifetime of WSNs. For this reason, algorithms and protocols need to address the issues like: lifetime maximization, robustness and fault tolerance, Self-configuration.

Operating *s*ystems for wireless sensor network nodes are typically less complex than general-purpose operating systems. The basic functionality of an operating system is to hide the low-level details of the sensor node by providing a clear interface to the external world. Hence a suitable operating system is required for WSN to provide functionalities like process management, multithreading etc.

TinyOS[1] is perhaps the first operating system specifically designed for wireless sensor networks. It is a free and open source component-based operating system and platform targeting wireless sensor Networks (WSN). TinyOS is an embedded operating system written in NesC [2] programming language as a set of cooperating tasks and processes. The section II, III, IV, V and VI outlines the various TinyOS scheduling policies.

II. FIRST IN FIRST OUT (FIFO) SCHEDULER

TinyOS adopts simple FIFO, i.e. first come first serve scheduling strategy. Tasks are scheduled by a FIFO queue. This approach helps to reduce the system requirement for storage space. However, it is not able to respond to real time emergency requests and it is likely to be overloaded resulting in the job loss and the information throughput getting lower etc.

A. Task scheduling

TinyOS supports the two-level concurrent models based on the combination of tasks and event-driven. The mechanism occurs as follows:

(i) There is no concept of priority and all tasks are equal. Also no preemption between tasks is allowed. All tasks share one execution space.

(ii) Tasks are posted in a circular task queue .The incoming tasks are posted at the end of circular queue and task to be scheduled are popped out from the front of the queue. Tasks are scheduled in a simple FIFO manner. Resources are distributed beforehand. Currently there can only be seven waiting tasks in the queue. The task-processing model is shown in Fig 1. The size of the task list in

the figure is eight. There are three tasks in the queue.

(iii) If the task queue is null and there is no events occurring, then the processor will enter into SLEEP mode automatically, and will be woken up by hardware interruption event subsequently. This is done to save energy of the system.

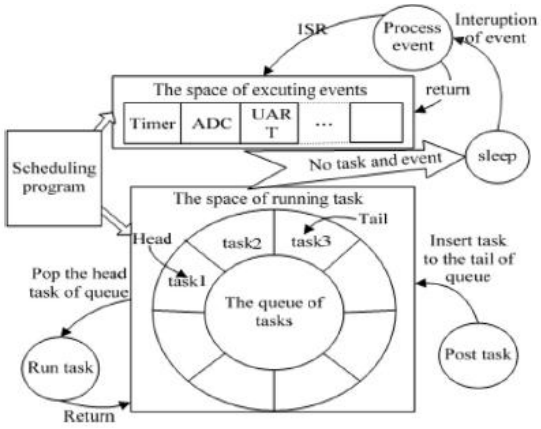


Fig. 1 TinyOS scheduler

A. *DRAWBACKS*

TinyOS is interrupt driven and typically there are two kinds of interrupts in TinyOS - clock and radio. Typically interrupts are propagated up as events and there will be appropriate event handlers to process these events. These event handlers typically post tasks to a task queue, which is a FIFO queue, and the TinyOS scheduler schedules these tasks on a FIFO [3] basis. There is a limit on the number of tasks which can be queued which is determined by the queue size. Currently, the number of tasks that can be queued is set to 7.

B. *OVERLOAD*

When the occurrence rate of the interrupts is very high, at some point the CPU is not able to execute any tasks other than the interrupt handlers. This situation is referred to as Overload .In such a situation the rate at which the CPU can complete its task is less than the arrival rate. Also it is not able to process any tasks which have been queued.

III. SHORTEST JOB PRIORITY (SJP)

SJP [4] implements the concept of time limit and based on the value (time limit) assigns priority to the jobs. As mentioned in section I, the inbuilt FIFO scheduler in TinyOS undergoes overloading resulting in a decreased throughput. Thus to overcome the pitfalls SJP divides the jobs into hard real time and soft real time jobs. Hard real time jobs are the jobs that must operate within the confines of a stringent deadline. Soft real time jobs are one in which usefulness of a result degrades after its deadline.

A. *SJP Data Structure*

The SJP data structure is defined as:

```
typedef struct
{
void (*tp) ();
uint16_t time limit;
uint16_t etime;
uint16_t runtime;
uint16_t SP;
bool preempted;
} TOSH_sched_entry_T;
```

The tp field is the entry address of the job function, and the job running is by running the entry address to complete the function. Time limit field holds a nonempty value when there is a hard real time job the etime field is empty, in case job is a soft real-time job etime field is a newly added field, which shows the user estimated run time, and the unit is also the clock tick. SP represents the current stack pointer location of the job and mainly be used for the context recovery when interruption occurs. Preempted field is an most important field, which identifies whether current job is interrupted. The field holds „1 if so.

B. *SJP Queue and Mechanism*

The SDP algorithm Judges the time limit value of the new job, if nonempty, then it inserts the job to the corresponding location of the head of the queue, in ascending order. Otherwise the job is inserted to the corresponding location of the tail of the queue in the order smallest to the largest by the etime value. The SJP queue is shown in fig 2. The jobs with nonempty time limit value are placed in front of jobs having no (empty)time limit value .Thus the former are ordered by time limit value from smallest to largest, while the latter are in the ascend order by the etime value.

	D(1)	D(4)	D(8)	D( )	D( )	D( )	
...	E(3)	E(1)	E(2)	E(2)	E(4)	E(5)	...

Fig. 2 SJP Job Queue

IV. DEADLINE SCHEDULER

Deadline scheduler [5] is an extension of FIFO scheduler which uses the concept of deadline for scheduling. Threads are executed based on their deadlines. In this scheme, another component responsible of queuing threads is created. This component uses deadline scheduling paradigm. The scheduler receives the threads with the deadline as a parameter. It then compares the deadlines of already queued tasks with newly arrived threads and places the thread in a suitable place in the scheduler. In order to avoid starvation of threads with long deadline, the deadlines of all other tasks, except the one which has finished its execution in the scheduler are decremented by (1).

V. EARLIEST DEADLINE FIRST(EDF) SCHEDULER

EDF [6] algorithm takes into consideration that every real-time task has deadline executing time,. And the priority of tasks is decided upon deadline time values assign to them. Thus lesser the absolute deadline time was, higher the priority of the task. Otherwise the further the absolute deadline time, the lower the priority of task was. The priority of tasks may be needed to adjust accordingly when a new task was ready. Thus the EDF algorithm has been proved to be dynamic priority scheduling mechanism compared to simple priority scheduling mechanism.

A. EDF Data Structure

The EDF data structure is defined as:

```
Typedef struct
 {
    void (*tp) ();
    uint16_t deadline;
    uint16_t atime;
    uint16_t runtime;
 } TOSH_sched_entry_T;
```

The fields here very well resembles to that of SJP job structure as discussed in section III. The tp points to entry address of a task, t atime represents the arrival time of a task and t runtime is the executing time of a task.

VI. FLEXIBLE POWER SCHEDULING (FPS)

Power is one of the dominant problems in wireless sensor networks. Flexible Power Scheduling (FPS) [7, 8] introduces a approach of scheduling that reduces the radio power consumption while supporting fluctuating demand in the network in sensor networks. Thus taking care of power savings. FPS protocol has been evaluated in two real world sensor network applications namely GDI (Great Duck Island) [9, 10] and TinyDB[11], on three mote platforms, mica, mica2dot, and mica2. A large class of TinyOS Applications fit this model, including equipment tracking, building-wide energy monitoring, habitat monitoring, conference room reservations etc.

A. FPS Structure

The FPS approach exploits the structure of a tree to build the schedule, which makes it useful primarily for data collection applications. A schedule tells every node when to listen and when to transmit. As the bandwidth needs are low, most nodes are idle most of the time, and the radio can be turned off during these periods along with the tree topology, it adopts an adaptive slotted communication schedule to route packets, synchronize with neighbors, and schedule radio on/off times.

B. FPS Mechanism

FPS scheduling is receiver initiated. In particular, the schedule spreads from the root of the tree down to the leaves based on the required bandwidth. Although this schedule ensures that parents and their children are contention free, there may still be contention due to other nodes in the network or poor time synchronization In brief the overall mechanism can be summarized as follows: Firstly the Time is divided into cycles and cycles are divided into slots. Each slot *s* corresponds to a length of time *T*, Slot numbering is periodic modulo *m*, i.e. slot $s+m$ is called slot *s* .Each node maintains a local schedule that indicates in what slot it transmits, receives, or idles.

The operations occur as:

1. Parent selects an idle slot *S* and advertises the slot
2. Child hears the advertisement and sends a request for slot S.
3. Parent receives the request and sends an acknowledgement.

VII. OPEN ISSUE AND CONCLUSION

The Table I gives a brief idea about TinyOS scheduling strategies describing each ones concept ,pros and cons**.**

The FIFO scheduler is simple, but it is not widely used in WSN applications due to overload in packet processing over wireless medium and decreased throughput.

So in order to improve the emergency jobs quick response, SJP introduced a new concept of dividing jobs

into hard and soft real-time jobs. Experimental evaluation depicts that SJP works fair at real time system, but the higher priority jobs need not be always critical thus resulting is the starvation of emergency jobs again.

The FPS protocol is a decentralized approach based on tree topology and works on FPS reservations. In FPS collisions can occur when two children out of radio range from each other respond to the same reservation advertisement .Some collisions may be due to one-off advertisement messages, the slots for which are randomly selected.

EDF can guarantee that all deadlines are met provided that the total CPU utilization is not more than 100%.So, compared to fixed priority scheduling techniques. EDF can guarantee all the deadlines in the system at higher loading. But evaluation shows that it does not ensure real- time scheduling. Finally the Deadline scheduler is based on threads. First, the scheduler receives the threads and knows their deadlines through receiving deadlines as arguments. It then compares these deadlines with the deadlines of already queued tasks in order to put this thread in the suitable place in the scheduler, thus avoiding starvation .Thus all the scheduler has its own benefits and pitfalls. TinyOS therefore requires a scheduler that combines all the features like preemption, low power consumption, starvation free, and, ensuring real time guarantees in one. Thus letting a wide scope for researchers in WSN operating system.

Table I

COMPARISON OF TINYOS SCHEDULING POLICIES

Scheduler	Characteristics			
	Type	Concept	Advantages	Disadvantages
FIFO	Non-preemptive	First Come First Serve	Simple	Overload, starvation of emergency jobs, decreased throughput
SJP	Preemptive	Priority to the shortest job: based on time limit	Increased throughput, resolves overload	High priority job need not be critical
FPS	--	Decentralized approach based on time division	Significant power saving, reduces contention, increased muti-hop throughput and fairness	Collision and message loss
EDF	Non-preemptive	Priority: Based on executing deadline of task	Priority decision : Dynamic	Does not ensure real- time scheduling
Deadline	Preemptive	Threads are executed based on deadlines	Prevents Starvation of threads having longer deadlines	--

VIII. ACKNOWLEDGEMENT

Our thanks to Principal, Prof. Vivek Kamat and Head of Department Prof. Nilesh Faldessai for giving us the opportunity to take up this project. We are also grateful to Prof. Seeya Gude, faculty,Dept of Information Technology, Goa college of Engineering, Farmagudi for the guidance, support and encouragement provided to us in the course of this project work.

REFERENCES

[1] www.tinyos.net

[2] P. Levis, D. Gay, “TinyOS Programming”, Cambridge University

[3] ZHAO Zhi-bin and GAO Fuxiang, “Study on Preemptive Real-Timr Scheduling Strategy for Wireless Sensor Networks”, 2009

[4] Jianhui Liu ,Information Technology College Eastern Liaoning University ,“Research on TinyOS Scheduling Strategy Based on SJP”,2010

[5] Aus dem Institut für Telematik der Universität zu Lübeck Direktor Prof. Dr. Stefan Fischer,” Management Support and CPU Scheduling Performance Enhancement in Wireless Sensor Networks”,2009

[6] Min YU and SiJi XIAHOU, XinYu LI,”A Survey of Studying on Task Scheduling Mechanism for TinyOS”,2008

[7] Barbara Hohlt and Eric Brewer Electrical Engineering and Computer Sciences Department University of California at Berkeley Berkeley, CA USA," Network Power Scheduling for TinyOS Applications"2006

[8] Barbara Hohlt, Lance Doherty, Eric Brewer," Flexible Power Scheduling for Sensor Networks",2004

[9] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, J. Anderson, "Wireless sensor networks for habitat monitoring," WSNA 2002, Atlanta, GA, USA, September 2002.

[10] R.Szewczyk, A.Mainwaring, J.Polastre, J.Anderson, D.Culler,"An Analysis of a Large Scale Habitat Monitoring Application", SenSys 2004,Baltimore, ML,USA, November 2004.

[11] S.R. Madden, M.J. Franklin, J.M. Hellerstein, and W. Hong, "TAG: a tiny aggregation service for ad-hoc sensor networks," 5th Symposium on Operating Systems Design and Implementation, Boston, MA, USA, December 2002.

❑❑❑

An Approach to Enhance CSMA/CA MAC Protocol For Wireless AD-HOC Networks

Milan Kumar Dholey[1] & Tanaya Das[2]

[1]Dept of Computer Science and Engineering, Hooghly Engineering & Technology College
[2]Dept of Computer Science and Engineering, Dr. B.C. Roy Engineering College
E-mail : milan.dholey@gmail.com[1], tanayadas.das23@gmail.com[2]

Abstract - In this paper we are going to introduce an approach for enhancement of CSMA/CA MAC protocol for ad hoc networks. Our proposal is to increase the number of unblocked nodes causes due to RTS/CTS. According to the timing diagram of CSMA/CA, for reliable communication RTS/CTS are used among nodes by setting their NAV value and are blocked during communication period. If we can reduce these number of block nodes then they can be used as another communication's source, relay or destination and will increase the network performance. Depending upon their signal power our proposal is to set an unequal NAV value of the nodes that are lies in transmission, detectable and interference zone in ad hoc network. So, blocking time gradually decreases from transmission, detectable and interference zone as the signal power of the node gradually decrease. Mathematically we shows the reduced NAV value of the nodes depending on signal power and modify the CSMA/CA MAC timing diagram which increase network performance.

Keywords - *NAV, RTS/CTS, CSMA/CA, timing Diagram.*

I. INTRODUCTION

Ad Hoc is a kind of network can be used for a specific purpose. In this there is no need of any router or any wireless base station. This network does not rely on any pre existing infrastructure and nodes are mobile. Besides this, they will send or receive data without any distortion. Since there is no pre existing infrastructure all the nodes communicates among them using radio waves, where each nodes offers a relay service and cooperate coordination using the node diversity.

These networks introduced a new art of network established and can be well suited for an environment where either the infrastructure is lost or where deploy an infrastructure is not very cost effective. The nodes in ad hoc network is surrounded by three zones, they are transmission zone, detectable zone and interference zone. Omni directional antenna is placed at each node. Omni directional antenna covers a range of 360o. Due to this Omni directional antenna these zones are at certain range having some radius.

At the time of data transmission RTS (Ready to send) is sent by the transmitter node whether it will be any relay or source. It is received by all nodes which lie in the range of transmission zone, detectable zone and interference zone and set their NAV (RTS) value and remain block during the completion of data transmission. After some time period CTS (Clear to send) is sent by the receiver node and all other nodes who receive CTS they set their NAV (CTS) value and this receiving time onwards the nodes of all three region are blocked. The signal power is gradually decrease from transmission zone to detectable and interference zone as the distance and power are inversely proportional.

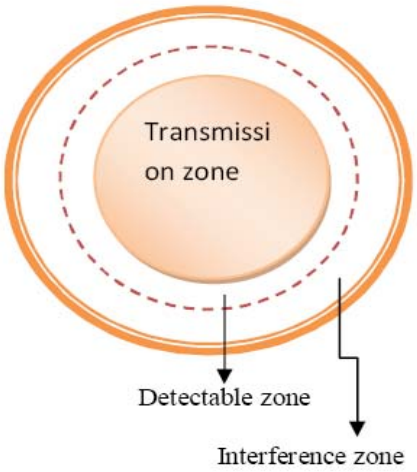


Fig. 1 : Zones of ad hoc network

As in Fig 1 all the nodes lies in three zones set same NAV value. But power of the signal is decreased gradually. In the interference zone they are not being communicate any more so there is no need to set the their NAV[1] value. In interference energy consumption is important [3] as low power signal reach to less number of nodes and consumption of energy is less. Where Energy specification [4] of CSMA/CA MAC protocol is standardized. Due to irregular distribution of sender and receiver of ad-hoc network interference averaging [5] is not effective so much. Concentrated on the idea of medium access and multimedia transmission [6] or retransmission developed a new frequency band for wireless communication. Now as it is not possible to fully avoid the interference zone so our proposal is to make less time value then original NAV (RTS) and NAV (CTS) value. Similar condition will apply in detectable and transmission zone. That is our proposal the NAV (RTS) and NAV (CTS) time value in transmission zone is greater than detectable zone's NAV (RTS) / NAV (CTS). As well as detectable zone's NAV (RTS) / NAV (CTS) than transmission zones (RTS) / NAV (CTS) i.e. make variable the NAV value. Now in interference zone's node will unblock previous to detectable zone's node and further detectable zone's node will unblock previous to transmission zone. Which nodes previously unblocked once they are not communicating now they start communication or used as a relay or destination. Hence the performance will increase.

The rest of the paper is organised as follows. Section II discusses about the existing CSMA/CA protocol with timing diagram. Section III discusses about the proposed protocol. Section IV shows the analytical representation of the proposed model with updated timing diagram of CSMA/CA. In Section V Conclusion is drawn and finally the references.

II. BASIC OF CSMA/CA MAC PROTOCOL

When data transmission occurs there is a chance of collision. Use of CSMA/CA avoid this problem of collision ,but at this time period the nodes whether it is source relay or destination remain blocked and is unable to participate in any other transmission. All the neighbouring nodes set a NAV (Network Allocation Vector) to the period during which a node cannot use the channel to transmit packets. Fig 2 shows the timing diagram of CSMA/CA.

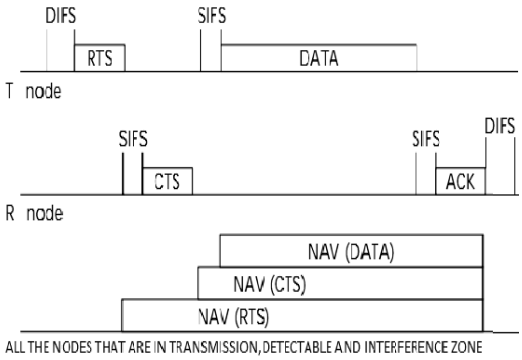


Fig. 2 : All Nodes that is present in three zones

In the above Fig. 2 we observe that a node want to send some data. For this transmission the node wait for some time period randomly. Then it sends RTS (Ready To Send) to all the nodes which are participating in the transmission. After some time the receiver nodes send CTS (Clear to send) indicates that they are ready to receive data. When a packet is received without any error the node at the destination responds with an ACK (Acknowledgement). If ACK is not received then it is assumed that the packet is lost and transmitted again.

Here in the above Fig. 2 shows NAV (RTS) and NAV (CTS) are irrespective the zones. So the nodes which lie in zones of transmission, detectable and interference of sender side have same NAV (RTS) and nodes lie in receiver side have same NAV (CTS) value. In this situation all the nodes having the same NAV values remain blocked during time of communication. These nodes remain blocked till the NAV value becomes 0.

III. PROPOSED PROTOCOL WITH TIMING DIAGRAM OF CSMA/CA

The directional antenna are use to unblock the node due RTS-CTS of CSMA/CA protocol in [1]. The authors in [2] first studied the performance of the RTS/CTS mechanism in IEEE 802.11 WLANs through simulations.

Now if we consider the data transmission between source or relay and destination where, Rs_t is radius of source node to its transmission zone; Rs_d is radius of source node to its detectable zone and Rs_i is radius of source node to its interference zone. Similarly, Rr_t is the radius of relay node to the transmission zone. Rr_d is the radius of relay node to detectable zone and Rr_i is the radius of relay node to interference zone and for

destination these are Rd_t, is the radius of destination node to transmission zone Rd_d is the radius of destination node to detectable zone and Rd_i is the radius of destination node to interference zone. Then Consider the first case where source(S) or relay (R for cooperative Ad-hoc network) and destination (D) are all in their transmission range according to Fig.1 it will look like fig 3.

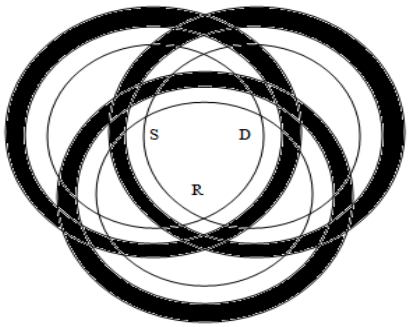


Fig.3 : The black colour shows the interference zone of source, relay and destination.

The timing diagram of the above figure is as follows in fig 4.If we want to apply our proposed protocol we need to segment the NAV (RTS) and NAV (CTS) of Fig. 2 and we will get the modifications of CSMA/CA as on Fig. 4.

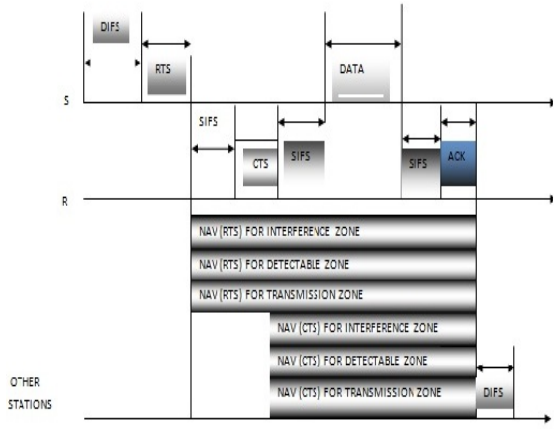


Fig. 4 **:** NAV (RTS) and NAV (CTS) for all zones

NAV (RTS) for source or relay transmission zone.

NAV (RTS) for source or relay detectable zone.

NAV (RTS) for source or relay interference zone.

NAV (CTS) for destination transmission range.

NAV (CTS) for destination detectable zone and

NAV (CTS) for destination interference zone.

When we apply our proposed protocol it will again modified and look like Figure 5 where it shows that NAV (RTS) and NAV (CTS) for interference zone is there but since this zones has less power signal i.e. the nodes in this region will not block during their data transmission to some extent.

IV. ANALYTICAL REPESENTATION OF THE PROPOSED MODEL

In the above figure we consider P as source node and Q as destination node.

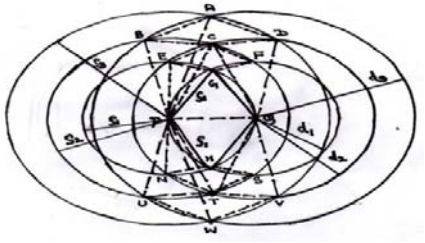


Fig 5: The intersection area of all three zones

The three zones surrounding node P has at distance s_1, s_2, s_3.Similarly node Q has distance d_1, d_2, d_3. Now According to [2] the area of intersection where the signal transmits the data is

$$A(L_{max}) = 2d_{max}^2 \cos^{-1}\left(\frac{L_{max}}{2d_{max}}\right) - \frac{1}{L_{max}}\sqrt{4d_{max}^2 - L_{max}^2} \quad (1)$$

Now,

Area of the source interference zone $A_{si}=\pi s_3^2$

Similarly,

Area of the source detectable zone $A_{sd}=\pi s_2^2$

Area of the source transmission zone $A_{st}=\pi s_1^2$

Now Area of Intersection between source interference zone and source detectable zone

$$AS_{ID} = A_{si} - A_{sd} = \pi (s_3^2 - s_2^2) \quad (2)$$

$$AS_{DT} = A_{sd} - A_{st} = \pi (s_2^2 - s_1^2) \quad (3)$$

Where, A_{DT} is the Area of Intersection between source detectable zone and source transmission zone. Similarly,

Area of the destination interference zone $A_{di}=\pi d_3^2$

Area of the destination detectable zone $A_{dd}=\pi d_2^2$

Area of the destination transmission zone $A_{dt}=\pi d_1^2$

Now Area of Intersection between destination interference zone and destination detectable zone

$$AD_{ID} = A_{di} - A_{dd} = \pi (d_3^2 - d_2^2) \quad (4)$$

$$AD_{DT} = A_{dd} - A_{dt} = \pi (d_2^2 - d_1^2) \quad (5)$$

Where, A_{DT} is the Area of Intersection between destination detectable zone and destination transmission zone.

Now if we remove the Area of intersection then the updated Fig 5 will look like Fig 6 and the area of the blocked region

A_{BA}= {(AS_{ID} + AS_{DT}) + (AD_{ID} + AD_{DT})} – (A(L_{max}) of outer intersection + inner intersection) (6)

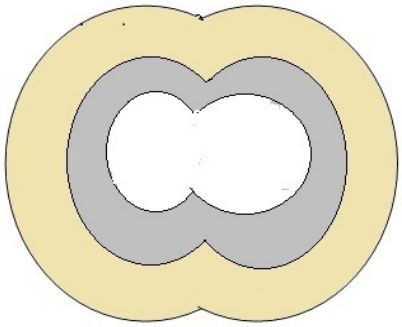


Fig. 6 : Removal of the intersection area between detectable, transmission zones, interference zone

Now according to [1] there is area of detectable zone and transmitted zone but not interference zone. But in my proposal the NAV value of detectable zone and transmission zone will not be same as [1] that means only the transmitted the data will set exact NAV value for RTS and CTS and in detectable zone it will less and in interference zone it will more less. So we consider variable NAV value for transmission, detectable and interference zone.

Now let us consider,

Ps_t is the power transmitted by the souce node in transmission zone

Ps_d is the power transmitted by the souce node in detectable zone

Ps_i is the power transmitted by the souce node in interference zone

$Ps_t > Ps_d > Ps_i$ (7)

Similarly the $SINR_t > SINR_d > SINR_i$

$SINR_t$ is the signal to noise ratio

So if we set a NAV value for all zones then according to Relation (7)

NAV_t=NAV value

NAV_d= NAV_t -Δt

Where Δt is the less amount transmission power for detectable zone.Similarly

NAV_i=NAV_d-Δt

Hence $NAV_t > NAV_d > NAV_i$

Hence the variable NAV values for the blocked area as per Equation 6 are shown in updated timing diagram in Fig 7.

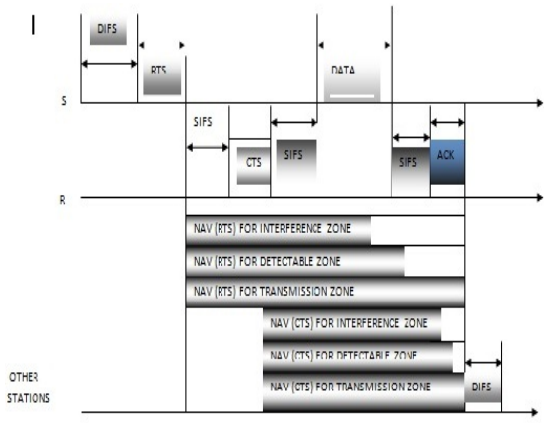


Fig.7 : Variable NAV values for three zones

Here, NAV (RTS) is now

a) NAV (RTS) for interference

b) NAV (RTS) for detectable and

c) NAV (RTS) for transmission.

Whereas, NAV (CTS) is now

a) NAV (CTS) for interference zone

b) NAV (CTS) for detectable and

c) NAV (CTS) for transmission zone.

V. CONCLUSION

In this paper we propose an approach to enhance CSMA/CA protocol for wireless ad hoc networks to reduce the blockage time with the help of variable NAV values. Since the NAV at the interference zone is less then it releases the blocked nodes faster than the nodes of other two zones and hence the blocked time gets reduced so they can take a part for other communication's source relay or destination.

To show this modified approach we have done some mathematical calculations and draw an updated timing diagram. Our future plan is to consider those nodes will transmit the signal using directional manner. As well as power of the signal is very important issues for wireless ad-hoc network so considering this in near future we can identify the interference zone and transmitting signal power by which we can reduced the power consumption and step ahead the work in these domain.

REFERENCES

[1] M.K.Dholey G.P.Biswas, Node- Blockage Reduction and Removal of Hidden-TerminalProblem of CSMA/CA MAC Protocol, IEEE RAIT-2012

[2] A. Basyouni, W. Hamouda, Amr Youssef "On Reducing Blocking Probability in Cooperative Ad-hoc Networks," IEEE " GLOBECOM" 2009 Proceedings. 978-1- 4244-4148-8/09 c2009

[3] A.Zabian,A.Ibrahim "Power Saving Mechanism in Clustered Ad-Hoc Networks" Journal of Computer Science 4 (5): 366-371, 2008 ISSN 1549-3636 © 2008 Science Publications

[4] ANSI /IEEE Std 802.11, Part 11: 'Wi reless LAN medium access control (MAC) and physical layer (PHY) specifications', (1999 edn.).

[5] Jeffrey G. Andrews, Steven Weber, Martin Haenggi , "Ad hoc networks: to spread or not to spread?" Submitted to IEEE Communications Magazine ¤ September 14, 2006

[6] G. Wu et al., "A Wireless ATM-Oriented MAC Protocol for High-Speed Wireless LAN," Proc. PIMRC '97, vol. 1, pp. 199–203.

❑❑❑

Generating Compact Rough Clusters Using an Evolutionary Algorithm

Venkata Rao J[1], Anusha K[2] & Y Navya Sree[3]

Department of Computer Science and Engineering, K L University, Vaddeswaram, Guntur(dt)-522502
E-mail : Venkat2all@gmail.com, mail4koneru@gmail.com navya.yarramsetti@gmail.com

Abstract **-** Cluster analysis is a key technique in the conventional data analysis and in knowledge discovery where it is more efficient or better suited to a particular type of data, cluster, or application. Many clustering methods have been identified, including the most widely used K-means approach, which requires the number of clusters to be specified in advance and is dependent on initial starting points. We present an evolutionary-based rough clustering algorithm, which is designed to overcome these restrictions. Rough clusters are defined in a similar manner to Pawlak's rough set concept, with a lower and upper approximation, producing different solutions to K-means analysis because of the possibility of multiple cluster membership of objects in the data set. The paper describes the templates, the data structure used to describe rough cluster. It also provides an overview of the evolutionary algorithm used to develop viable cluster solutions, consisting of an optimal number of templates providing descriptions of the clusters and extending the template descriptions to include generalized templates.

Keywords **-** *Clusters, rough sets, template, K-means.*

I. INTRODUCTION

Cluster analysis is a key technique in the traditional data analysis of data mining and in knowledge discovery. Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both. Clusters should capture the natural structure of the data if merely meaningful. In some cases, however, cluster analysis is a useful starting point for a purpose of data summarization. Cluster analysis groups data objects such that it describes the objects and their relationships. The objective is that the objects within a group be similar to one another and different from the objects in other groups. The goal is maximizing the inter-cluster similarity and minimizing intra-cluster similarity i.e., to obtain maximal homogeneity within the subgroups or clusters, and maximal heterogeneity between clusters. Clustering techniques can be broadly divided into three main types: overlapping (so-called non-exclusive), partitional (unnested), and hierarchical (nested). An object can simultaneously belong to more than one group is an overlapping or non-exclusive clustering, where as in exclusive each object is assigned to a single cluster only. In a fuzzy clustering, every object belongs to every cluster with a membership weight that is between 0 and 1 i.e., clusters are treated as fuzzy sets. Several clustering methods have been identified and among them the most prominent nonhierarchical methods is the K-means approach [8].

K-Means, a prototype-based, partitional clustering technique that attempts to find a user-specified number of clusters (K), to represent their centroids. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster, repeats until the centroids remain the same. K-medoid [7] defines a prototype in terms of a medoid as like a centroid in K-means, which is the most representative point for a group of points, and can be applied to a wide range of data since it requires only a proximity measure for a pair of objects. The data set is partitioned into clusters and proximity measure of data is Euclidean distance where an error of each data point is calculated, i.e., its Euclidean distance to the closest centroid, and then compute the total sum of the squared errors. The clusters obtained with the smallest squared error is preferred for the better representation of the points in their cluster. To find the lowest squared error is an expensive task and local optimization has conventionally been used. A major limitation of this technique is the number of clusters in each partition is prior to the analysis. A *k*-Modes approach has been developed as an extension of the *k*-Means algorithm, and has been applied to categorical data clustering by

replacing means with modes [4], a simple matching dissimilarity measure for categorical objects and a frequency based method to update modes in the K-means fashion to minimize the clustering cost function. Also it preserves the efficiency of the K-means algorithm. This technique also involves the specification of number of clusters in advance is a trade-off.

For the large volumes of data, from the past few decades as the field of data mining is advanced, impact of size and complexity of data sets are grown. Many new techniques based on advancements in computational intelligence have started to be more widely used as clustering algorithms. For example, the theory of fuzzy sets was subsequently applied to cluster analysis.

The emphasis is on the theory of rough sets, a technique from the field of computational intelligence. It offers mathematical tools to discover patterns hidden in data. The recent applications of rough set theory[9] are usually expressed as initially, a distance matrix for all object pairs is calculated. All object pairs at interobject distance D, where D steps from 0 to a determined maximum, are then identified for cluster membership. This paper describes a rough clustering technique, based on a simple extension of rough sets theory, applicable where prior group membership is not known. The section 2 describes the concept of rough sets and their properties. The section 3 is about the related work. The section 4 describes the rough clustering algorithm and finally the conclusion and the future enhancement.

II. ROUGH SETS AND PROPERTIES

Rough sets were originally proposed using equivalence relations with properties as specified by Pawlak[10]. It offers mathematical tools to discover patterns hidden in data. Recent extensions of rough set theory (rough mereology) have developed new methods for decomposition of large data sets, data mining in distributed and multi-agent systems, and granular computing. The core idea is to separate discernible from indiscernible objects and is based on the assumption that, with every object of an information system (or data matrix), there is associated a certain amount of the information. This information is expressed by means of attributes used as descriptions of the objects. The data is treated from the perspective of set theory.

The properties of rough sets are defined by Pawlak. The Fig.1 shows the Set Approximations: the lower approximation, upper approximation and a boundary area Lingras et al. [9] do not verify all the properties of rough set theory

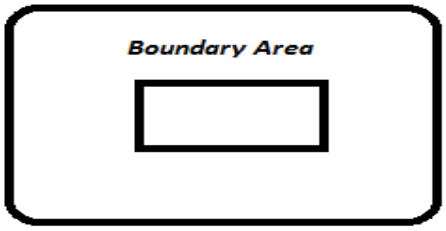


Fig. 1 : Lower, Upper approximation and Boundary Area.

-Upper Approximation

- Lower Approximation

but use to the following basic ones for their algorithm:

- Property 1: a data object can be a member of one lower approximation at most.
- Property 2: a data object that is a member of the lower approximation of a cluster is also member of the upper approximation of the same cluster.
- Property 3: a data object that does not belong to any lower approximation is member of at least two upper approximations.

Therefore, strictly speaking, the presented clustering algorithm is not part of classical rough set theory but belongs to the (reduced) interpretation of rough sets as lower and upper approximations of data constellation. So the algorithm can also be interpreted as two layer interval clustering approach with lower and upper approximations.

However, the family of upper and lower approximations are required to follow some of the basic rough set properties such as:

1. An object **v** can be part of at most one lower approximation. This implies that any two lower approximations do not overlap.
2. An object **v** that is member of a lower approximation of a set is also part of its upper approximation (**v** □ *A(**xi**)* → **v** □ *A(**xi**)*). This implies that a lower approximation of a set is a subset of its corresponding upper approximation (*A(Xi)* □ *A(Xi)*).
3. If an object **v** is not part of any lower approximation it belongs to two or more upper approximations. This implies that an object cannot belong to only a single boundary region. Note that these basic properties are not necessarily independent or

complete. However, enumerating them will be helpful in understanding the rough set adaptation of the k-means algorithm.

The complete information system expresses all the knowledge available about the objects being studied [4]. More formally, the information system is a pair, S = (*U*, *A*), where *U* is a non-empty finite set of objects called the universe and $A = \{ a_1, a_2, \ldots, a_j \}$ is a non-empty finite set of attributes on *U*. With every attribute a □ *A* we associate a set Va such that a : *U* → Va. The set Va is called the domain or value set of a. In statistical terms, this value set equates to the range of values associated with a specific variable. The initial detailed data contained in the information system is used as the basis for the development of subsets of the data that are "coarser" or "rougher" than the original set. As with any data analysis technique, detail is lost, but the removal of detail is controlled to uncover the underlying characteristics of the data. The technique works by 'lowering the degree of precision in data, based on a rigorous mathematical theory.

A core concept of RS is that of equivalence between objects (indiscernibility). Objects about which we have the same knowledge form an equivalence relation. Let S = (*U*, *A*) be an information system, then with any B □ *A* there is associated an equivalence relation, IND*A* (B), the B-indiscernibility relation. It is defined as:

$$IND_A(B) = \{ (x, x') \in U^2 \mid \forall a \in B \;\; a(x) = a(x') \} \quad (1)$$

If (x, x') □ IND*A* (B), then the objects x and x' are indiscernible from each other when considering the subset B of attributes. Equivalence relations lead to the universe being divided into partitions, which can then be used to build new subsets of the universe.

The discernibility formula :
means that in order to discern object *x1* and *x2*, at least one of the following cuts must be set, a cut between *a*(0.8) and *a*(1), a cut between *b*(0.5) and *b*(1) and a cut between *b*(1) and *b*(2).
A set is said to be *rough* if its boundary region is non-empty, otherwise the set is crisp.

Let S = (*U*, *A*) be an information system, and let B □ *A* and X □ *U* . We can describe the subset X using only the information contained in the attribute values from the subset B by constructing two subsets, referred to as the B-lower and B-upper approximations of X, and denoted as B*(X) and B*(X) respectively, where:

$$B_*(X) = \{x \mid [x]B \subseteq X\} \text{ and } B^*(X) = \{x \mid [x]B \cap X \neq \emptyset\} \quad (2,3)$$

The lower approximation (LA), defined in (2), contains objects that are definitely in the subset X and the upper approximation (UA), defined in (3), contains objects that may or may not be in X. A third subset is also useful in analysis, the boundary region, which is the difference between the upper and lower approximations. This definition of a rough set in terms of two other crisp sets is the simple but powerful insight contributed by Pawlak, which has led to numerous publications exploring the implications (e.g. [10]). Rough sets theory has developed an extensive literature well beyond the brief introduction provided here, and the interested reader is referred to [2], [8], [9] and [26] for comprehensive overviews of developments in the field. For example, one major extension of relevance to rough clustering is the development of the concept of *similarity*, a relaxing of the strict requirement of indiscernibility in canonical RS theory, to include objects that are similar rather than identical. A number of ways of defining similarity have been proposed, and [18] provides an introduction to the issues involved. Dissimilarity measures have also been proposed.

III. RELATED WORK

3.1 Rough Clustering

Rough clustering is an extension of the theory of rough or approximation sets, introduced by Pawlak. Rough clustering produces different solutions to k-means analysis because of the possibility of multiple cluster membership of objects. Traditional clustering methods generate extensional descriptions of groups, that show which objects are members of each cluster. Clustering techniques based on rough sets theory generate intentional descriptions, which outline the main characteristics of each cluster for set approximation: with a lower approximation (LA) and upper approximation (UA).

The LA of a rough cluster contains objects that only belong to that cluster, and by definition, the objects belong to the UA as well. The UA of a rough cluster contains objects that may belong to more than one cluster. The clustering algorithm described in [8] used a distance measure to construct a similarity matrix, and each object-object pair in this similarity matrix was assigned to existing or new clusters depending on whether none, one or both objects in the pair were currently assigned. Problems with this approach were the large number of clusters generated and uncertainty as to whether the lower approximations of each cluster provide the most efficient coverage of the data set.

Rough clustering is a simple extension of the notion of rough sets, involving two additional requirements – an ordered value set of attributes and a distance measure. The value set is ordered to allow a meaningful distance measure, and clusters of objects are formed on

the basis of their distance from each other, in a similar manner to standard clustering techniques. In addition,. Clusters are formed in a similar manner to agglomerative hierarchical clustering. However, an object can belong to more than one cluster. Clusters can then be defined by a lower approximation (objects exclusive to that cluster) and an upper approximation (all objects in the cluster which are also members of other clusters), in a similar manner to rough sets. An introduction to rough clustering, and a detailed comparison between rough clustering and k-means clustering, can be found in Voges, Pope and Brown (2002).

A different approach was followed in [6], who used reducts to develop clusters. Reducts are subsets of the attribute set *A*, which provide the same information as the original data set. The reducts are used as initial group centroids, which are then grouped together to form clusters. One problem with this approach is that not all information systems have reducts, and some sets of reducts overlap, which means that the cluster centroids are not necessarily well separated. The present study uses such a generalized view of rough sets. If one adopts a more restrictive view of rough set theory, the rough sets developed in [10], may have to be looked upon as interval sets.

3.2 Comparison of Rough and K-means Clusters

Comparison of rough clustering with *K*-means clustering was reported in [2], and found that the two clustering techniques resulted in some clusters that were identified by both techniques, and some clusters that were unique to the particular technique used. The rough clustering solution is necessarily different, because of the possibility of multiple cluster membership of objects. The rough clustering technique also found clusters that were "refined" subclusters of those found by *k*-means clustering, and which identified a more specific sub-segment of the data set.

With the number of clusters required to describe the data dependent on the distance measure using Rough clustering also produces more clusters than *K*-means clustering [2], More clusters means an object has a higher chance of being in more than one cluster. A solution with too few clusters does not provide a useful interpretation of the partitioning of the data. On the other hand, too many clusters make interpretation difficult. In addition, the degree of overlap between the clusters needed to be minimized to ensure that each cluster provided information to aid in interpretation. Rough clustering can be conceptualized as extracting concepts from the data, rather than strictly delineated sub-groupings. Determining a good rough cluster solution requires a trade-off between various factors. An evolutionary algorithms are a good way of conducting this trade-off which are explained in the following topic.

3.3 Evolutionary algorithms and Rough Sets

From an optimization perspective, clustering can be formally considered as a particular kind of NP-hard grouping problem [11]. This has stimulated the search for efficient approximation algorithms, including not only the use of *ad hoc* heuristics for particular classes or instances of problems, but also the use of general-purpose metaheuristics. Particularly, evolutionary algorithms are metaheuristics widely believed to be effective on NP-hard problems, being able to provide near-optimal solutions to such problems in reasonable time. Under this assumption, a large number of evolutionary algorithms for solving clustering problems have been proposed in the literature. These algorithms are based on the optimization of some objective function (i.e., the so-called fitness function) that guides the evolutionary search.

An evolutionary rough C-means clustering algorithm proposed by Mitra, to determine the relative importance of upper and lower approximations of rough sets used to model the clusters. The fitness function used in the evolutionary algorithm involved minimizing a specific measure, the Davies–Bouldin clustering validity index. Kumar [7] used an agglomerative hierarchical clustering algorithm for sequential data, where the indiscernibility relation was extended to a tolerance relation with the transitivity property being relaxed.

Kohonen self-organizing map (SOM) for pre-processing of data[3], which was then further divided into clusters using rough sets and genetic algorithms. It is an unsupervised and non-parametric neural network approach. The success of the SOM algorithm lies in its simplicity that makes it easy to understand, simulate and be used in many applications. The basic SOM consists of neurons usually arranged in a two-dimensional structure such that there are neighborhood relations among the neurons.

Lingras [3] developed a genome comprising two sections – LA membership and UA membership. The approach required some repair operators, as some randomly generated genes could be invalid. One limitation of this approach was that the number of clusters needed to be specified in advance, and this preliminary knowledge is not always available for larger data sets. For e.g., a hybrid system to develop linguistic-based technical stock market indicators with rough sets theory used to extract linguistic rules and a genetic algorithm to refine these extracted rules. The effectiveness of the proposed model was verified for both forecasting accuracy and stock returns, and showed

that the proposed model was superior to rough sets and genetic algorithms applied independently.

In the following section we present an extension of rough clustering that attempts to overcome the limitations of these previous attempts to apply RS theory to cluster analysis. The approach suggested uses an EA to maximize the coverage of the data set, without pre-specifying the number of clusters required, without relying on structural characteristics of the cluster such as reducts, and using a straight forward fitness function.

IV. ROUGH CLUSTERING ALGORITHM

In the previous section some of the shortcomings of previous attempts to apply RS theory to the clustering problem has briefly mentioned. To overcome these shortcomings, an EA based approach is proposed that attempts to find the set of lower approximations of the rough clusters, which provides the most comprehensive coverage of the data set with the minimum number of clusters [4]. The number of rough clusters is not specified in advance.

4.1 Data Structure

A data structure is a dynamic set where a set can grow or shrink or an implementation of a dynamic set.

The basic data structure used for describing a rough cluster is the *template* as described in [6]. Let S = (*U*, *A*) be an information system. Any clause of form D =(a □ Va) is called a *descriptor*, with the value set Va called the range of D. A *template* is a conjunction of unique descriptors defined over attributes from *B* □*A*. More formally, any propositional formula T = ^ a□ *B* (a □ Va) is called a *template* of S. To create a viable description of a cluster using a template, at least two attributes from *B* are chosen. This results in compact, but non-trivial, descriptions of the rough cluster.

Template T is *simple* if any descriptor of T has a range of one element. Templates with descriptors having a range of more than one element are called *generalized*. In the example presented below, only simple templates are used. However the technique could be easily extended to include generalized templates, incorporating intervals of attributes (i.e. using a similarity relation rather than an indiscernibility relation).

The data structure acted on by the EA is a cluster solution, C, which is defined as any conjunction of k unique templates,

$$C= T_1 \wedge T_2 \wedge \ldots \wedge T_k \quad (4)$$

This data structure was encoded as a simple two dimensional array with a variable length equal to the number of unique templates in the cluster solution and a fixed width equal to the number of attributes being considered. Possible values in the template were the same as the values in the data set (-2, -1, 1, 2 in Example 1 below), with 0 being used as a "don't care" value. Table 1 shows an example data structure with eight unique templates, using the data presented in Example 1 (Section 5 below).

Table 1. Data Structure example

T	Variables[1]				
	Image	Package	Price	Alcohol	Place
1	0	-2	0	0	-1
2	0	-2	0	0	1
3	0	-2	0	0	2
4	0	-1	0	1	0
5	0	-1	0	2	0
6	0	1	2	0	0
7	0	2	2	0	0
8	2	0	-2	0	0

[1] See Section 5 below

A template describes a partition of *U* and the conjunction of templates contained in a cluster solution results in some templates having both LAs (that is, objects satisfying one template only) and UAs (that is, objects satisfying more than one template). Consequently C is a *rough cluster* solution.

4.2 Pre-processing

The maximum number of templates generated depends on the number of attributes of each object, p, and the range of values for the attribute, q. In a simple case where each attribute has the same range, the maximum number of simple templates generated is p^q. Depending on the data set, many of these templates are of little value in describing the data, either because they cover too small a percentage of the data, or because they are trivial. For example, a cluster solution could be developed using one attribute with four values by simply having four templates : value 1, value 2, value 3 and value 4. This would cover 100% of the data set, but would only provide a simple frequency distribution.

To overcome this problem, objects in the complete set of possible templates were individually checked against the data set, and only those templates containing two or more attributes and matching more than a specified percentage of the data set were considered valid. As will be seen in the following example, this can considerably reduce the number of templates that the EA needs to sample in order to generate useful rough cluster descriptions. This has the effect of reducing the processing time required, as the EA is not dealing with

templates that add little to the overall coverage or that are invalid.

For example, all of values shown in the templates presented in Table 1 occur in the original data set, and all have at least two attribute values.

4.3 Fitness measure

A number of objectives need to be considered when developing a fitness measure for rough clustering: (i) maximize the data set coverage *c*, defined as the fraction of the universe of objects that match the set of templates in the cluster solution, C; (ii) minimize k, the number of templates in the cluster solution, C; (iii) maximize the accuracy *a*, of each template [15].

More formally, for any X □ *U*, the set of objects { x □ X : □ a □ B a (x) □ Va } from X satisfying any template T_i is denoted by $[T_i]X$. $[T_i]$ *X is a lower approximation if x is unique to that set. $[T_i]$ *X is an upper approximation if x is contained in $[T_i]X$ and at least one other set $[T_j]X$. We therefore define the following values:

$$c = (\Sigma \mid [T_j]_* X \mid) / \mid U \mid, \text{where } \{1 \le j \le k\} \qquad (5)$$

That is, the *coverage c*, is the sum of the cardinal values of the LAs of each template in the cluster solution, C, divided by the cardinal value of *U*, the full data set.

$$a = \Sigma (\mid [T_i]_* X \mid / \mid [T_i] X \mid), \text{where } \{1 \le j \le k\} \qquad (6)$$

That is, the *accuracy a*, is the sum of the cardinal value of the LA divided by the cardinal value of the UA for each template in the cluster solution, C.

The *fitness* value, *f*, of each cluster solution, C, is defined as the coverage multiplied by accuracy divided by the number of templates in C.

$$f = (c \times a) / k \qquad (7)$$

4.4 Recombination and mutation operators

In the current study, a multi-point operator was used to simplify the task of ensuring that only valid rough cluster solutions were generated. The size of the offspring was determined by randomly selecting a value between the sizes of both parents. Templates were then randomly selected from each parent, and then added to the offspring after checking that they were not already present in the cluster solution. In this way, a unique solution, containing material from both parents, was generated. An example of this recombination operator is presented in Table 2. To provide another source of diversity in the population, two mutation operators were developed.

The first operator (MutateAdd) randomly sampled a new template from the list of valid templates and, after checking to ensure that the template was not already in the cluster solution, added that template to the cluster solution. The second operator (MutateDelete) randomly removed a template from the cluster solution. In this current work, the probability of mutation has been set quite high, at 50%, to ensure that all of the valid templates have been sampled for possible inclusion in a cluster solution.

Repair operators were not required as infeasible solutions were not generated by either the recombination or mutation operators. The only constraint, ensuring each cluster solution contains only unique templates, was easily handled by checking the current set of templates in the cluster solution at the time the offspring were created.

Table 2.
Recombination operator example

Parent 1					
1-1	0	-2	0	0	-1
1-2	0	-2	0	0	1
1-3	0	-2	0	0	2
1-4	0	-1	0	1	0
Parent 2					
2-1	0	-1	0	2	0
2-2	0	1	2	0	0
2-3	0	2	2	0	0
2-4	2	0	-2	0	0
		↓			
Offspring					
1-1	0	-2	0	0	-1
1-2	0	-2	0	0	1
1-3	0	-2	0	0	2
2-1	0	-1	0	2	0

V. CONCLUSIONS AND FUTURE RESEARCH

This paper has presented an extension of work in rough sets theory and rough clustering. Clusters obtained from conventional techniques usually have crisp boundaries, that is, each object belongs to only one cluster, but many real-world data sets do not lend themselves to such a neat solution. Rough clusters allow an object to belong to multiple clusters. Previous studies using evolutionary algorithms have required that the number of clusters be specified in advance, a major limitation with large or complex data sets. The research presented in this paper uses templates (conjunctions of attribute-value descriptors) to describe the cluster solution. An evolutionary algorithm was used to find a rough cluster solution that covers the largest percentage of the data set with the smallest number of accurate lower approximations. Also involve extending the template descriptions to include generalized templates. Further research can also be conducted on the decision-making processes used by consumers, which would help to validate the reasoning processes suggested by the rough clustering findings and comparing this technique with other approaches.

REFERENCES:

[1] Alpigini, J. J., J. F. Peters, A. Skowron, Rough Sets and Current Trends in Computing, Third International Conference, RSCTC 2002

[2] Bezdek, J. C., "Numerical Taxonomy with Fuzzy Sets", Journal of Mathematical Biology.

[3] Bouyer, A., Hatamlou, "An Optimized Clustering Algorithm Using Genetic Algorithm and Rough Set Theory Based on Kohonen Self Organizing Map",International Journal of Computer Science and Information Security, 2010

[4] Cao F., J. Liang, D. Li, L. Bai, "A Dissimilarity Measure for the k-Modes Clustering Algorithm", Knowledge-Based Systems, 2011.

[5] Cheng, C-H., T-L. Chen, "A Hybrid Model Based on Rough Sets Theory and Genetic Algorithms for Stock Price Forecasting", Information Science, 2010

[6] do Prado, H. A., P. M. Engel, "Rough Clustering: An Alternative to Find Meaningful Clusters by Using the Reducts From a Dataset", in Alpigini, 2002

[7] A Survey of Evolutionary Algorithms for Clustering**,** Eduardo R. Hruschka, IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews

[8] Voges, K. E., N. K. Ll. Pope, "Cluster Analysis of Marketing Data Examining On-line Shopping Orientation: A Comparison of K-means andRough Clustering Approaches", in Abbass,

[9] Pawlak, Z., "Rough Sets", International Journal of Information and Computer Sciences, 11(5), 1982,

[10] Pawlak, Z. Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer, Boston, 1991.

❑❑❑

Service-Oriented Architecture (SOA) and virtualization Access Control in Cloud Computing

[1]K.GOVARDHAN BABU & [2]V. GANESH DUTT

Dept of Information Technology & Dept of Computer Science and Engineering
Sri Sunflower college of Engineering and Technology, Lankapalli
Email: kgovardhanbabu@yahoo.co.in, ganeshduttv@gmail.com

Abstract - Cloud computing is an emerging computing paradigm in which resources of the computing infrastructure are provided as services over the Internet. As promising as it is,this paradigm also brings forth many new challenges for datasecurity and access control when users outsource sensitive data for sharing on cloud servers, which are not within the same trusted domain as data owners. To keep sensitive user data confidential against untrusted servers, existing solutions usually apply cryptographic methods by disclosing data decryption keys only to authorized users. However, in doing so, these solutions inevitably introduce a heavy computation overhead on the data owner for key distribution and data management when finegrained data access control is desired, and thus do not scale well. Other hand, allowing the data owner to delegate most of the computation tasks involved in finegrained data access control to untrusted cloud servers without disclosing the underlying data contents. We achieve this goal by exploiting and uniquely combining techniques of attribute-based encryption (ABE), proxy reencryption, and lazy re-encryption. Our proposed scheme also has salient properties of user access privilege confidentiality and user secret key accountability. Extensive analysis shows that our proposed scheme is highly efficient and provably secure under existing security models.

I. INTRODUCTION

Cloud computing is a promising computing paradigm which recently has drawn extensive attention from both academia and industry. By combining a set of existing and new techniques from research areas such as Service-Oriented Architectures (SOA) and virtualization, cloud computing is regarded as such a computing paradigm in which resources in the computing infrastructure are provided as services over the Internet. Along with this new paradigm, various business models are developed, which can be described by terminology of "X as a service (XaaS)" where X could be software, hardware, data storage, and etc. Successful examples are Amazon's EC2 and S3 ,Google App Engine ,and Microsoft Azure which provide users with scalable resources in the pay-as-youuse fashion at relatively low prices. For example, Amazon's S3 data storage service just charges $0.12 to $0.15 per giga byte month.

As compared to building their own infrastructures,users are able to save their investments significantly by migrating businesses into the cloud. With the increasing development of cloud computing technologies, it is not hard to imagine that in the near future more and more businesses will be moved into the cloud. As promising as it is, cloud computing is also facing many challenges that, if not well resolved, may impede its fast growth.

Data security, as it exists in many other appliations, is among these challenges that would raise great concerns from users when they store sensitive information on cloud servers. These concerns originate from the fact that cloud servers are usually operated by commercial providers which are very likely to be outside of the trusted domain of the users.

Data confidential against cloud servers is hence frequently desired when users outsource data for storage in the cloud. In some practical application systems, data confidentiality is not only a security/privacy issue, but also of juristic concerns. For example, in healthcare application scenarios use and disclosure of protected health information (PHI) should meet the requirements of Health Insurance Portability and Accountability Act (HIPAA) ,and keeping user data confidential against the storage servers is not just an option, but a requirement. It would allow data consumers such as doctors, patients, researchers and etc, to access various types of healthcare records under policies admitted by HIPAA. To enforce these access policies, the data owners on one hand would like to take advantage of the abundant resources that the cloud provides for efficiency and economy; on the other hand, they may want to keep the data contents confidential against cloud servers. As a significant research area for system protection, data access control has been evolving in the past thirty years and various techniques have been developed to effectively

implement fine-grained access control, which allows flexibility in specifying differential access rights of individual users. Traditional access control architectures usually assume the data owner and the servers storing the data are in the same trusted domain, where the servers are fully entrusted as an omniscient reference monitor responsible for defining and enforcing access control policies. This assumption however no longer holds in cloud computing since the data owner and cloud servers are very likely to be in two different domains. On one hand, cloud servers are not entitled to access the outsourced data content for data confidentiality; on the other hand, the data resources are not physically under the full control of the owner.

II. RELATED WORK

Existing work close to ours can be found in the areas of "shared cryptographic file systems" and "access control ofoutsourced data".In Kallahalla et al proposed Plutus as a cryptographic file system to secure file storage on untrusted servers. Plutus groups a set of files with similar sharing attributes as a file-group and associates each file-group with a symmetric lockbox-key. Each file is encrypted using a unique file-blcok key which is further encrypted with the lockbox-key of the filegroup to which the file belongs. If the owner wants to share a file-group, he just delivers the corresponding lockbox-key to users. As the complexity of key management is proportional to the total number of file-groups, Plutus is not suitable for the case of fine-grained access control in which the number of possible "file-groups" could be huge.

Goh et al proposed SiRiUS which is layered over existing file systems such as NFS but provides end-to-end security. For the purpose of access control, SiRiUS attaches each file with a meta data file that contains the file's access control list (ACL), each entry of which is the encryption of the file's file encryption key (FEK) using the public key of an authorized user. The extension version of SiRiUS uses NNL broadcast encryption algorithm to encrypt the FEK of each file instead of encrypting it with each individual user's public key. As the complexity of the user revocation solution in NNL is proportional to the number of revoked users, SiRiUS has the same complexity in terms of each meta data file's size and the encryption overhead, and thus is not scalable. Ateniese et al proposed a secure distributed storage scheme based on proxy re-encryption. Specifically, the data owner encrypts blocks of content with symmetric content keys. The content keys are all encrypted with a master public key, which can only be decrypted by the master private key kept by the data owner. The data owner uses his master private key and user's public key to generate proxy re-encryption keys, with which the semi-trusted server can then convert the ciphertext into that for a specific granted user and fulfill the task of access control enforcement. The main issue with this scheme is that collusion between a malicious server and any single malicious user would expose decryption keys of all the encrypted data and compromise data security of the system completely. In addition, user access privilege is not protected from the proxy server. User secret key accountability is neither supported.

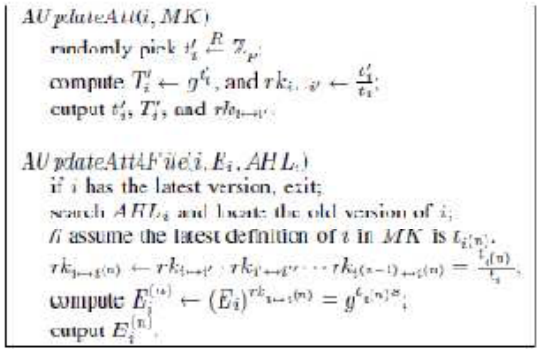
2.2.1The algorithm in this paper :

Key Policy Attribute-Based Encryption (KPABE)

KP-ABE is a public key cryptography primitive for one-to-many communications. In KP-ABE, data are associated with attributes for each of which a public key component is defined. The encryptor associates the set of attributes to the message by encrypting it with the corresponding public key components. Each user is assigned an access structure which is usually defined as an access tree over data attributes, i.e., interior nodes of the access tree are threshold gates and leaf nodes are associated with attributes. User secret key is defined to reflect the access structure so that the user is able to decrypt a cipher text if and only if the data attributes satisfy his access structure.

Algorithm level operations: Algorithm level operations include eight algorithms: *ASetup*, *AEncrypt*, *AKeyGen*,*ADecrypt*, *AUpdateAtt*, *AUpdateSK*, *AUpdateAtt*4*File* and *AMinimalSet*. As the first four algorithms are just the same as *Setup*, *Encryption*, *Key Generation*, and *Decryption* of the standard KP-ABE respectively. we focus on our implementation of the last four algorithms,depicts two of the four algorithms.

Pseudo-code of algorithm level algorithms



III. PROPOSEDWORK:

Proposed scheme is able to realize the desired security goals, i.e., fine-grained access control, data confidentiality, user access privilege confidentiality, and user secret key accountability. The goal of scalability is

also achieved since the complexity for each operation of our proposed scheme, as is shown in , is no longer dependent to the number of users in the system. Therefore, our proposed scheme can serve as an ideal candidate for data access control in the emerging cloud computing environment. On the contrary, existing access control schemes in related areas either lack scalability, and fine-grainedness , or do not provide adequate proof of data confidentiality

IV. EXPERIMENTAL RESULTS

Upload files

Details of the User

V. CONCLUSION AND FUTURE WORK

This paper aims at fine-grained data access control in cloud computing. One challenge in this context is to achieve finegrainedness,data confidentiality, and scalability simultaneously, which is not provided by current work. In this paper we propose a scheme to achieve this goal by exploiting KPABE and uniquely combining it with techniques of proxy re-encryption and lazy re-encryption. Moreover, our proposed scheme can enable the data owner to delegate most of computation overhead to powerful cloud servers. Confidentiality of user access privilege and user secret key accountability can be achieved. Formal security proofs show that our proposed scheme is secure under standard cryptographic models.

VI. REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, “Above the clouds: A berkeley view of cloud computing,” University of California, Berkeley, Tech. Rep. USB-EECS- 2009-28, Feb 2009.

[2] Amazon Web Services (AWS), Online at http://aws. amazon.com. [3] Google App Engine, Online at http://code.google.com/appengine/.

[4] Microsoft Azure, http://www.microsoft.com /azure/.

[5] 104th United States Congress, “Health Insurance Portability and Accountability Act of 1996 (HIPPA),” Online at http://aspe.hhs.gov/ admnsimp/pl104191.htm, 1996.

[6] H. Harney, A. Colgrove, and P. D. McDaniel, “Principles of policy in secure groups,” in *Proc. of NDSS’01*, 2001.

[7] P. D. McDaniel and A. Prakash, “Methods and limitations of security.

❑❑❑

Fuzzy Keyword Search Over Encrypted Data

Nikijahan Tamboli, Bharat Savani, Ruchita Choudhari, Nikesh Shah & ApekshaGadkar

Department of Computer Engg.,
Smt. KashibaiNavale College Of Engineering,Vadgaon(Bk.) Pune.
E-mail : nikkijahan.5@gmail.com, savanibharat@gmail.com, ruchitachoudhari4@gmail.com, nikesh749@gmail.com, apekshagadkar@gmail.com

Abstract - Fuzzy matching is a technique used in computer-assisted translation and some other information technology applications such as record linkage. It works with matches that may be less than 100% perfect when finding correspondences between segments of a text and entries in a database of previous translations. It usually operates at sentence-level segments, but some translation technology allows matching at a phrasal level. It is used when the translator is working with translation memory.When an exact match cannot be found in the TM database for the text being translated, there is an option to search for a match that is less than exact; the translator sets the threshold of the fuzzy match to a percentage value less than 100%, and the database will then return any matches in its memory corresponding to that percentage. Its primary function is to assist the translator by speeding up the translation process; fuzzy matching is not designed to replace the human translator.

Keywords-components: Fuzzy keyword set, Encryption, Searching.

I. INTRODUCTION

Fuzzy matching programs usually return irrelevant hits as well as relevant ones. Superfluous results are likely to occur for terms with multiple meanings, only one of which is the meaning the user intends. If the user has only a vague or general idea of the topic, or does not know exactly what to look for, the ratio of relevant hits to irrelevant hits tends to be low. The ratio is even lower, however, when an exact matching program is used in this situation.

In cryptography, encryption is the process of transforming information referred to as plaintext using an algorithm called a cipher to make it unreadable to anyone except those possessing special knowledge, usually referred to as a key. The result of the process is encrypted information in cryptography, referred to as cipher text. The reverse process, i.e., to make the encrypted information readable again, is referred to as decryption i.e., to make it unencrypted.

II. SURVEY OF EXISTING SYSTEM

Due to the polymorphous and dynamic nature of language, particularly English which accounts for 90% of all source texts undergoing translation in the localization industry, methods are always being sought to make the translation process easier and faster. Since the late 1980s, translation memory tools have been developed to increase productivity and make the whole translation process faster for the translator.In the 1990s, fuzzy matching began to take off as a prominent feature of TM(translation memory) tools, and despite some issues concerning the extra work involved in editing a fuzzy match "proposal", it is still a popular subset of TM. It is currently a feature of most popular TM tools.

Encryption, by itself, can protect the confidentiality of messages, but other techniques are still needed to protect the integrity and authenticity of a message; for example, verification of a message authentication code (MAC) or a digital signature. Standards and cryptographic software and hardware to perform encryption are widely available, but successfully using encryption to ensure security may be a challenging problem. A single slip-up in system design or execution can allow successful attacks. Sometimes an adversary can obtain unencrypted information without directly undoing the encryption.

III. PROPOSED SYSTEM

Fuzzy keyword search greatly enhances system usability by returning the matching files when users' searching inputs exactly match the predefined keywords or the closest possible matching files based on keyword similarity semantics, when exact match fails. More specifically, we use edit distance to quantify keywords similarity.

In this particular system the encryption of various file formats such as text file, image file and video file can be done. This function is used to maintain the security in stable storage and access mechanisms. Security is maintained by encrypting the data, user want to search and access.

IV. METHODOLOGY

The input given to the system is a keyword to search. Therefore a fuzzy set is to be maintained in a database. The database returns the files containing the particular keyword. The database should contain the filenames and the fuzzy set. The fuzzy set is to be created by using the algorithm which supports the edit distance n. Searching can be done according to the degree of membership of characters.The encryption can be done at server site by using various encryption algorithms such as Data Encryption Standard (DES)Advanced Encryption Standard (AES).The user will search the encrypted files and the searching can be done in the degree of membership of filename characters like dictionary search. The user can view the file in the encrypted format. For decrypting the file the private key id can be used. As to enhance more security the fuzzy key id can also be provided. Combined by both the keys user can successfully retrieve the files.

Various methods which can be used for creating fuzzy set are like:-

A] Edit Distance Matching Method:-

D(ai,bj)=

MIN(

(D(ai,bj-1)+CI(bi)),

(D(ai-1,bj-1)+CM(ai,bi)),

(D(ai-1,bj)+CD(ai))

)

where,

CI -cost of insertion.

CM -cost of match/substitution.

CD -cost of deletion.

D(ai,x) first i letters of the word a.

D(x,bj) first j letters of the word b.

B] Probablity Matching Method:-

D(ai,bj)=

MAX(

(D(ai,bj-1) . PI(bi)),

(D(ai-1,bj-1) . PM(ai,bi)),

(D(ai-1,bj) . PD(ai))

)

where,

PI -frequency of insertion.

PM - frequency of match/substitution.

PD - frequency of deletion.

D(ai,x) first i letters of the word a.

D(x,bj) first j letters of the word b.

C] Bayesian Probability Matching Method

P(t|o)=(P(o|t)P(t))/P(o)

Where,

P(t|o) is the probability that the true word is t given that the word is o.

P(o|t) is the probability that the fuzzy will output the word o when presented with the word t. (This is the probability of match function above.)

P(t) is the frequency of the word t in the dictionary.

P(o) is the frequency of the word o in the output space

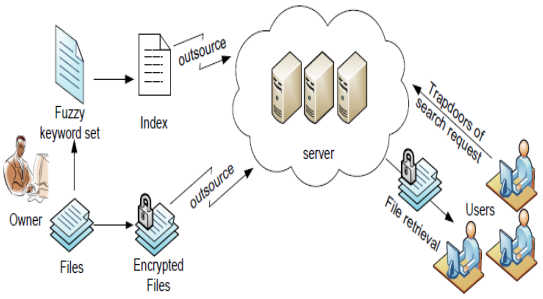
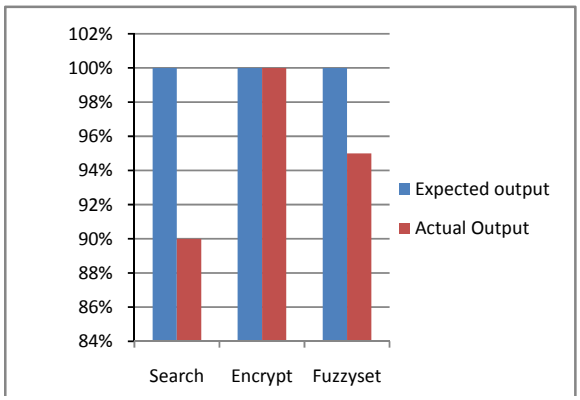


Figure 1: SysteM Architecture

V. EXPERIMENTAL RESULTS

Experiments are performed to verify the effectiveness of the proposed system. The expected output and actual output of each module in the system is compared by giving various inputs to the system.



VI. ADVANTAGES AND LIMITATIONS

A. Advantages:

1) Everytime decryption of various file is not needed.

2) The fuzzy set is used for the faster access to search the string in file and for file name.

3) No combinational search provided so it does not lead to fake results.

4) Fuzzy searching is when used for research and investigation.

5) Fuzzy searching can also be used to locate individuals based on incomplete or partially inaccurate identifying information.

B. Limitations:

1) Extreme large database is to be handled due to large number of words in thesaurus.

2) Thesaurus word meaning changes due to fuzzy sets.

VII. FUTURE SCOPE

In the future scope of this system we are willing to do the indexing of the mapped words and fuzzy sets so as to increase the functionality of the search procedure. Encryption of more file formats can be done. Also decryption of image files and media files can be done.

VIII. CONCLUSION

This paper presents aim to make search quality more efficient as it eliminates the combinational search and introduces the concept of the exact search. These technologies are being improved & enhanced as well as being made more efficient. Another thing which we can do is encryption of the stored data like text files, image files and various file formats of media files too.Through rigorous security analysis, we show that our proposed solution is secure and privacy- preserving, while correctly realizing the goal of fuzzy keyword search. Extensive experimental results demonstrate the efficiency of our solution.

REFERENCES

[1] "IEEE paper ''Enabling Efficient Fuzzy Keyword Search over Encrypted Data in Cloud Computing by Jin Li, Qian Wang, Cong Wang, Ning Cao, KuiRen, and Wen jing Lou.''

[2] "N-Gram-Based Text Categorization" by William B. Cavnar and John M. Trenkle.

[3] http://en.wikipedia.org/wiki/Fuzzy_logic

[4] M. Bellare, A. Boldyreva, and A. O'Neill, "Deterministic and efficiently searchable encryption," in Proceedings of Crypto 2007, volume 4622 of LNCS. Springer-Verlag, 2007.

[5] D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. of IEEE Symposium on Security and Privacy'00, 2000.

[6] E.-J. Goh, "Secure indexes," Cryptology ePrint Archive, Report 2003/216, 2003, http://eprint.iacr.org/.

❑❑❑

Analysis Of The Packet Scheduling Algorithms For WiMAX

G.Sateesh, Prasanthi Bheri, P. Rajesh, A.Rama Rao
Department of Computer Science & Engineering
Lendi Institute of Engineering &Technology

Abstract*—The use of Broadband Wireless Access (BWA) is rapidly increasing due to user mobility and data access facilities. IEEE 802.16e WiMAX networks guarantees the best available QoS for mobile data services. The WiMAX packet scheduling algorithms allocate the resource bandwidth for the transfer of packets. In this paper, we have done the analysis of WiMAX packet scheduling algorithms, Round Robin(RR), Deficit Round Robin(DRR), Weighted Deficit Round Robin(WDRR) and Weighted Fair Queuing(WFQ), calculated the throughput, delay and fairness for all the algorithms.

Key Terms*—*IEEE 802.16e, QoS, Round Robin, Deficit Round Robin, Weighted Deficit Round Robin, Weighted Fair Queuing, WiMAX.

I. INTRODUCTION

IEEE 802.16 is a set of telecommunications technology standards aimed at providing wireless access over long distances in a variety of ways - from point-to-point links to full mobile cellular type access. IEEE 802.16 standards group has been developing a set of standards for broadband (high-speed) wireless access (BWA)[1][2][3]. IEEE 802.16 standards based WiMAX (Worldwide Interoperability for Microwave Access) is the technology aimed at providing architecture is designed to achieve goals like easy deployment, high-speed data rate, large spanning area, and large frequency spectrum[1][11]. The IEEE 802.16 standard is capable of providing QoS to all different kinds of application including real time traffic in the form of flow type associated with each application.

II. IEEE 802.16 PHYS: SINGLE CARRIER (SC), OFDM AND OFDMA

IEEE 802.16 supports a variety of physical layers. Each of these has its own distinct characteristics. First, Wireless MANSC (Single Carrier) PHY is designed for 10 to 60 GHz spectrum. While IEEE has standardized this PHY, there are not many products implementing it because this PHY requires line of sight (LOS) communication[1]. To allow non-line of sight (NLOS) communication, IEEE 802.16 designed the Orthogonal Frequency Division Multiplexing (OFDM) PHY using spectrum below 11 GHz. Multiple users are allowed to transmit using different subcarriers in the same time slot[6]. The scheduling decision then is to decide which subcarriers and what time slots should be allocated to which user. This combination of time division and frequency division multiple access in conjunction with OFDM is called Orthogonal Frequency Division Multiple Access (OFDMA).

III. WiMAX FRAME STRUCTURE

The IEEE 802.16 defines the frame format of the WiMAX. Each frame consists of downlink (DL) and uplink (UL) sub-frames. A preamble is used for time synchronization. The downlink map (DL-MAP) and uplink map (UL-MAP) define the burst-start time and burst-end time, modulation types and forward error control (FEC) for each MS[1]. Frame Control Header (FCH) defines these MAP's lengths and usable subcarriers. Basically each MPDU is a MAC frame with MAC header (6bytes), other sub-headers such as fragmentation and packing sub-headers, grant management (GM) sub-header (2 bytes) if needed and finally a variable length of payload.

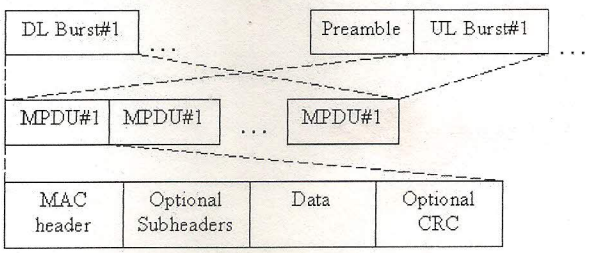


Fig. 1: MPDU Frame Format

Therefore, WiMAX supports adaptive modulation and coding, i.e., the modulation and coding can be changed adaptively depending on the channel condition. Either MS or BS can do the estimation and then BS decides the most efficient modulation and coding scheme. Channel Quality

Indicator (CQI) is used to pass the channel state condition information. We discuss five main issues related to the frame structure below; namely, number of bursts, two dimensional rectangular mapping for downlink sub frame, MPDU size, fragmentation and packing considerations[1].

IV. WiMAX QoS SERVICE CLASSES

IEEE 802.16 defines five QoS service classes[7][8]: Unsolicited Grant Scheme (UGS), Extended Real Time Polling Service (ertPS), Real Time Polling Service (rtPS), Non Real Time Polling Service (nrtPS) and Best Effort Service (BE). Each of these has its own QoS parameters such as minimum throughput requirement and delay/jitter constraints.

Unsolicited Grant Service(UGS): This service class provides a fixed periodic bandwidth allocation. Once the connection is setup, there is no need to send any other requests. The main QoS parameters are maximum sustained rate (MST), maximum latency and tolerated jitter (the maximum delay variation).

Real time Packet Service(rtPS): This service class is for variable bit rate (VBR) realtime traffic such as MPEG compressed video. The QoS parameters are similar to the UGS but minimum reserved traffic rate and maximum sustained traffic rate need to be specified separately.

Extended Real Time Packet Service(ertPS): This service is designed to support VoIP with silence suppression. No traffic is sent during silent periods. ertPS service is similar to UGS in that the BS allocates the maximum sustained rate in active mode, but no bandwidth is allocated during the silent period.

Non Real Time Packet Service(nrtPS): This service class is for non-real-time VBR traffic with no delay guarantee. Only minimum rate is guaranteed. File Transfer Protocol (FTP) traffic is an example.

Best Effort(BE): Most of data traffic falls into this category. This service class guarantees neither delay nor throughput. The
bandwidth will be granted to the MS if and only if there is a left-over bandwidth from other classes.

V. SCHEDULING ALGORITHMS IN WiMAX

Packet scheduling is the process of resolving contention for bandwidth. A scheduling algorithm has to determine the allocation of bandwidth among the users and their transmission order. One of the most important objectives of a scheduling scheme is to satisfy the Quality of Service (QoS) requirements of its users while efficiently utilizing the available bandwidth. Some of the existing algorithms are:

A. Round Robin (RR): Round_Robin as a scheduling algorithm is considered the most basic and the least complex scheduling algorithm[2][10]. Basically the algorithm services the backlogged queues in a round robin fashion.

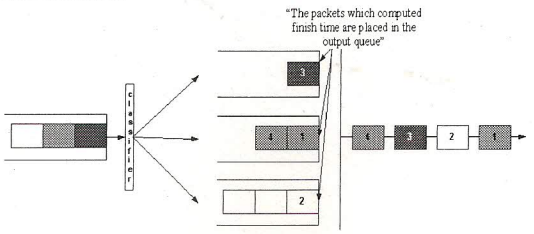


Fig. 2: Round Robin

B. Deficit Round Robin (DRR): Weighted round robin was designed to differentiate flows or queues to enable various service rates[2][4][10]. It operates on the same bases of RR scheduling.

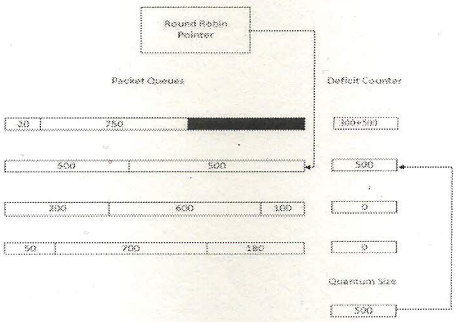


Fig. 3: Deficit Round Robin

C. Weighted Deficit Round Robin (WDRR):
Weighted deficit round robin allows a network administrator to group traffic into classes[2][5]. A "class" is an entity the network administrator defines to receive distinct treatment in the queue[4][10].

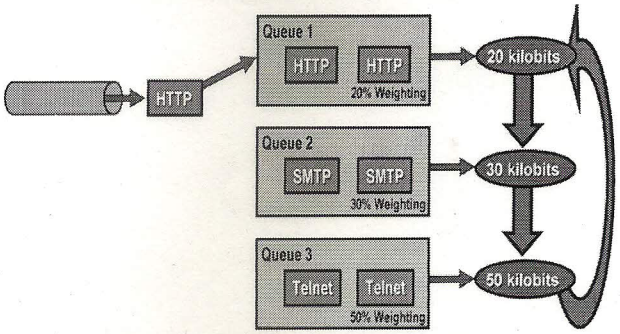


Fig. 4: Weighted Deficit Round Robin

D. Weighted Fair Queuing (WFQ): WFQ is an approximation of General Processor Sharing (GPS)[2][10]. WFQ does not make the assumption of infinitesimal packet size.

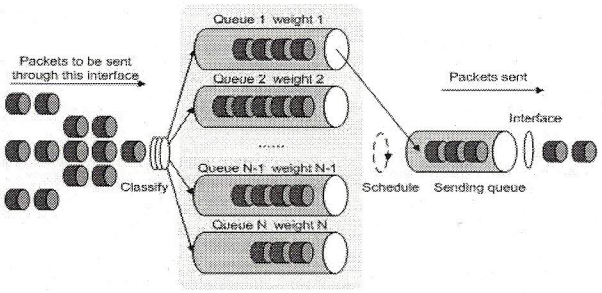


Fig. 5: Weighted Fair Queuing

VI. SIMULATION MODEL AND ANALYSIS

A. Evaluation Methodology

The overall goal of this study is to analyze the performance of different existing scheduling algorithm in WiMAX environment. We have considered the most versatile method of testing scheduling algorithms which will actually simulate the designed algorithm with real life data and conditions.

B. Simulated Scenario

The overall goal of this analysis is to analyze the performance of different existing algorithms in WiMAX. The important parameters used to configure the PHY and MAC layers are summarized in Table1.

Table 1:Simulation Parameters

Base Station parameter	ANTENNA-TYPE = OMNI
	ANTENNA-GAIN = 15 dB
	ANTENNA-HEIGHT = 25 m
Transmission parameter	TX-POWER = 15 dBm
	CH.-BANDWIDTH = 5MHz
	FFT-SIZE = 512
	CYCLIC-PREFIX = 8
	FRAME-DURATION = 20MS
	DUPLEX MODE=TDD
PHY 802.16	FREQUENCY = 2.4 GHz
	PATHLOSS = TWO-RAY
	FADING = RAYLEIGH

To avoid queuing packets of different service types into one queue several precautions have been taken. The precedence values corresponding for each queue are shown in Table 2.

Table 2: Traffic Classes Vs Precedence

MAC Layer Services	Precedence/Queue
BE	0
nrtPS	2
rtPS	3
ertPS	4
UGS	7

Throughput Optimization

In communication networks, such as Ethernet, throughput or network throughput is the average rate of successful message delivery over a communication channel[8] slot.

The data obtained for the throughput are specified in Table 3.

Table 3: Values obtained for throughput

	RR	DRR	WDRR	WFQ
0	2379	1034	943	3533
1	2366	934	978	2943
2	3058	933	818	2636
3	3043	932	970	1895
4	2828	943	856	3423
5	2266	928	1217	1753
6	2030	936	1287	1882
7	1707	970	953	1895
8	1998	915	864	2366
9	3056	729	1112	1627
10	1718	1194	780	1624
11	1457	936	1130	1943
12	3155	932	1319	1867
13	1776	1243	813	1681
14	1408	1433	1074	1501
15	1932	897	965	2334
16	2185	923	933	2251
17	1533	1202	936	2338
18	1886	934	901	2285
19	2406	914	947	1880
Mean	2209.4	993.1	989.8	2183

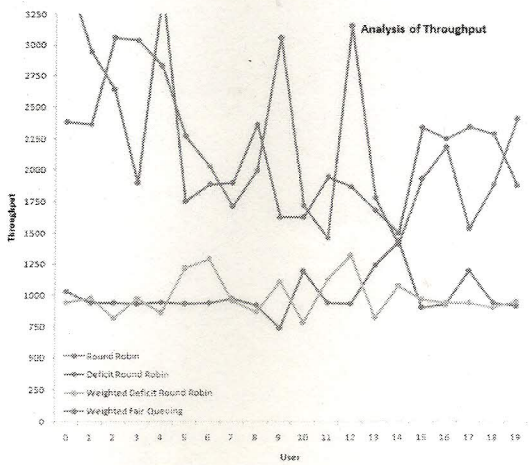


Fig. 6 : Analysis of Throughput

Based on the values obtained for the different algorithms, Weighted Deficit Round Robin has high mean value for the throughput.

Fairness

Fairness is used in network engineering to determine whether users or applications are receiving a fair share of system time.

The data obtained for the fairness are specified in Table 4.

Table 4: Values obtained for fairness

	RR	DRR	WDRR	WFQ
0	0.995798319	0.942307692	0.926315789	0.896358543
1	0.941176471	0.729166667	0.979591837	0.942567568
2	0.928571429	0.821052632	0.975609756	0.835820896
3	0.97704918	0.914893617	0.897959184	0.921465969
4	0.957746479	0.926315789	0.953488372	0.950581395
5	0.852173913	0.872340426	0.894308943	0.903954802
6	0.950980392	0.957446809	0.9	0.957671958
7	0.810344828	0.897959184	0.927083333	0.973684211
8	0.940298507	0.945652174	0.931034483	0.941176471
9	0.986928105	0.98630137	0.928571429	0.981595092
10	0.988372093	0.95	0.873417722	0.902439024
11	0.911564626	0.852631579	0.912280702	0.913265306
12	0.890282132	0.914893617	0.917293233	0.678306878
13	0.97752809	0.944	0.914634146	0.773255814
14	0.915492958	0.951388889	0.944444444	0.940397351
15	0.857142857	0.966666667	0.948453608	0.974358974
16	0.886877828	0.924731183	0.925531915	0.916299559
17	0.954545455	0.933884298	0.957446809	0.906779661
18	0.978835979	0.729166667	0.793478261	0.978165939
19	0.901234568	0.934782609	0.864583333	0.894736842
Mean	0.93014721	0.904779093	0.918276365	0.919144113

Fairness Index Ratio= (tpmax-tpmin)/tpmax

Where, tpmax is the maximum throughput

tpmin is the minimum throughput

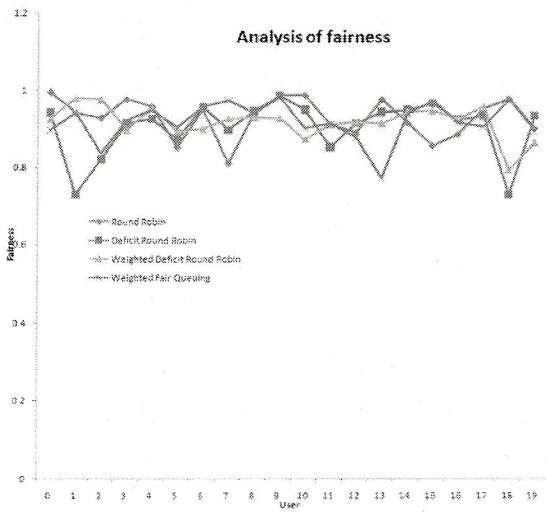


Fig. 7: Analysis of Fairness

Based on the values obtained for the different algorithms, Round Robin has high mean value for the fairness.

Delay

Delay (or queuing delay) is the time a packet waits in a queue until it can be served. For a better algorithm the delay is always less. We observed that Weighted Deficit Round Robin has the least Delay and Weighted Fair Queuing has the highest delay.

The data obtained for the delay are specified in Table 5.

Table 5: Values obtained for delay

	RR	DRR	WDRR	WFQ
0	1.00421941	1.06122449	1.07954545	1.115625
1	1.0625	1.371428571	1.02083333	1.0609319
2	1.07692308	1.217948718	1.025	1.19642857
3	1.02348993	1.093023256	1.11363636	1.08522727
4	1.04411765	1.079545455	1.04878049	1.05198777
5	1.17346939	1.146341463	1.11818182	1.10625
6	1.05154639	1.044444444	1.11111111	1.0441989
7	1.23404255	1.113636364	1.07865169	1.02702703
8	1.06349206	1.057471264	1.07407407	1.0625
9	1.01324503	1.013888889	1.07692308	1.01875
10	1.01176471	1.052631579	1.14492754	1.10810811
11	1.09701493	1.172839506	1.09615385	1.09497207
12	1.12323944	1.093023256	1.09016393	1.13855422
13	1.02298851	1.059322034	1.09333333	1.29323308
14	1.09230769	1.051094891	1.05882353	1.06338028
15	1.16666667	1.034482759	1.05434783	1.02631579
16	1.12755102	1.081395349	1.08045977	1.09134615
17	1.04761905	1.07079646	1.04444444	1.10280374
18	1.02162162	1.371428571	1.26027397	1.02232143
19	1.10958904	1.069767442	1.15662651	1.11764706
Mean	1.07817041	1.112786738	1.09131461	1.09138042

Based on the values obtained for the different algorithms, Round Robin has high mean value for the delay.

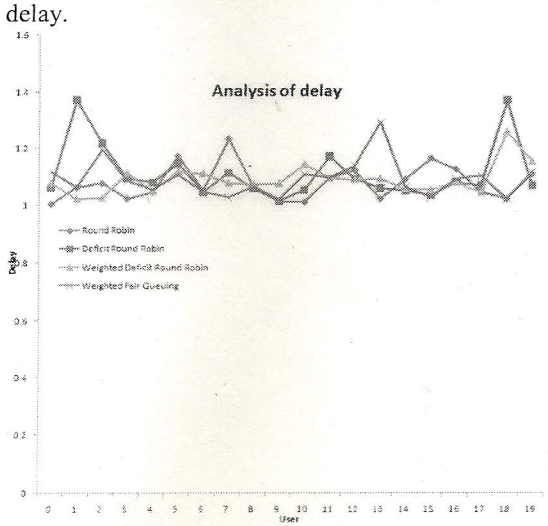


Fig. 8 : Analysis of Delay

The analysis of the algorithms are reported using the Table 6.

Table 6: Analysis of Algorithms

	RR	DRR	WDRR	WFQ
Fairness	High	Low	Medium	Medium
Throughput	Low	Medium	High	Medium
Delay	High	Low	Medium	Medium

VII.CONCLUSION

IEEE 802.16e WiMAX networks guarantees the best available quality of experience for multimedia data services. In this paper, we have done the analysis of WiMAX packet scheduling algorithms, Round Robin, Deficit Round Robin, Weighted Deficit Round Robin and Weighted Fair Queuing, for the parameters like throughput, delay and fairness. In this, Weighted Deficit Round Robin algorithm has got high average throughput, Round Robin algorithm has got high average Fairness and low average Delay. Comparing all these algorithm performances WDRR algorithm is the best in packet scheduling algorithm. These algorithms can be further extended to improve many factors such as end-to-end delay, resilience, bandwidth and packet delivery ratio.

ACKNOWLEDGEMENT

We would like to thank our principal Dr. V.V.Rama Reddy, management members our chairman Sri P.Madhusudhana Rao, Vice-chairman Sri Srinivasa Rao and Secretary Sri K.Siva Rama Krishnan for their valuable support and guidance. We would also add Sri M.Rajan Babu, Vice-principal and Professor who suggested and guided us to successfully complete this work.

REFERENCES

[1] Chakchai So-In, Raj Jain and Abdel-Karim Tamimi, "Scheduling In IEEE 802.16e Mobile WiMAX Networks: Key Issues and a Survey", IEEE Journal on Selected Areas In Communications, VOL. 27, NO. 2, FEBRUARY 2009.

[2] Ahmed H. Rashwan, Hesham M.ElBadawy, Hazem H. Ali "Comparative Assessments for Different WiMAX Scheduling Algorithms", WCECS 2009, October 20-22, 2009, San Francisco, USA.

[3] WiMAX Forum, "WiMAX System Evaluation Methodology V2.1," Jul. 2008, 230 pp. Available: http://www.wimaxforum.org/technology/documents

[4] Sari R. F., Gde D I, Mukhayaroh N, Laksmiati D, "Performance Evaluation of Weighted Round Robin", Proceedings of Quality in Research Conference (QIR) 2007, December 2007, FTUI Jakarta.

[5] Shreedhar and G.Varghese, "Weighted Deficit Round Robin (DWRR)", 1995. mhttp://en.wikipedia.org/wiki/DWRR.

[6] "Time Division Duplex (TDD) vs Frequency Division Duplex (FDD) in Wireless Backhauls", NETKROM TECHNOLOGIES.

[7] "WiMAX : E vs. D ,The Advantages of 802.16e over 802.16d", WHITE PAPER: WiMAX, Vol. 6, 2007, Motorola, Inc.

[8] "WiMAX QoS Classes Using WiMAX QoS Classes to support Voice, Video, and Data Traffic", 2010, Tranzeo wireless technologies Inc.

[9] Silberschatz, Galvin and Gagne; , A textbook for "Operating System Concepts", Sixth Edition, Chapter 3, Pg NO: 39 ,Wiley.

[10] Data packet Scheduling algorithms http://en.wikipedia.org/wiki/Weighted_round_robin, http://en.wikipedia.org/wiki/round_robin , http://en.wikipedia.org/wiki/Deficit_round_robin. , http://en.wikipedia.org/wiki/Weighted_fair_queuing

[11] "IEEE 802.16 Standard", IEEE 802.16 Working Group, 2004

❑❑❑

Effective and Scalable Outsourcing of Linear Programming in Cloud Computing

G. V. S. S. PRASADA RAJU & ABDUL VAHED

Dept of Information Technology, Sri Sunflower college of Engineering and Technology, Lankapalli
Dept of Computer Science and Engineering, Sri Sunflower college of Engineering and Technology, Lankapalli
Email: vsspraju@gmail.com, abdulvahed@hotmail.com

Abstract - Cloud Computing has great potential of providing robust computational power to the society at reduced cost. It enables customers with limited computational resources to outsource their large computation workloads to the cloud, and economically enjoy the massive computational power, bandwidth, storage, and even appropriate software that can be shared in a pay-per-use manner. Despite the tremendous benefits, security is the primary obstacle that prevents the wide adoption of this promising computing model, especially for customers when their confidential data are consumed and produced during the computation. Treating the cloud as an intrinsically insecure computing platform from the viewpoint of the cloud customers, we must design mechanisms that not only protect sensitive information by enabling computations with encrypted data, but also protect customers from malicious behaviors byenabling the validation of the computation result. Such a mechanism of general secure computation outsourcing was recently shown to be feasible. in theory, but to design mechanisms that are practically efficient remains a very challenging problem. Focusing on engineering computing and optimization tasks, this paper investigates secure outsourcing of widely applicable linear programming (LP) computations. In order to achieve practical efficiency, our mechanism design explicitly decomposes the LP computation outsourcing into public LP solvers running on the cloud and private LP parameters owned by the customer.

I. INTRODUCTION

Cloud Computing provides convenient ondemand network access to a shared pool of configurable computing resources that can be rapidly deployed with great efficiency and minimal management overhead. One fundamental advantage of the cloud paradigm is computation outsourcing, where the computational power of cloud customers is no longer limited by their resource-constraint devices. By outsourcing the workloads into the cloud, customers could enjoy the literally unlimited computing resources in a pay-per-use manner without committing any large capital outlays in the purchase of both hardware and software and/or the operational overhead therein. Despite the tremendous benefits, outsourcing computation to the commercial public cloud is also depriving customers' direct control over the systems that consume and produce their data during the computation, which inevitably brings in new security concerns and challenges towards this promising computing model. On the one hand, the outsourced computation workloads often contain sensitive information, such as the business financial records, proprietary research data, or personally identifiable health information etc. To combat against unauthorizedinformation leakage, sensitive data have to be encrypted before outsourcing so as to provide endto-end data confidentiality assurance in the cloud and beyond. However, ordinary data encryption techniques in essence prevent cloud from performing any meaningful operation of the underlying plaintext data, making the computation over encrypted data a very hard problem. On the other hand, the operational details inside the cloud are not transparent enough to customers. As a result, there do exist various motivations for cloud server to behave unfaithfully and to return incorrect results, i.e., they may behave beyond the classical semihonest model. For example, for the computations that require a large amount of computing resources, there are huge financial incentives for the cloud to be "lazy" if the customers cannot tell the correctness of the output.

II. RELATED WORK

A. Work on Secure Computation Outsourcing

General secure computation outsourcing that fulfills all aforementioned requirements, such as input/output privacy and correctness/soundness guarantee has been shown feasible in theory by Gennaro et al. However, it is currently not practical due to its huge computation complexity. Instead of outsourcing general functions, in the security community, Atallah et al. explore a list of work for securely outsourcing specific applications. The customized solutions are expected to be more efficient than the general way of constructing the circuits. it gives

the first investigation of secure outsourcing of numerical and scientific computation.

B. Basic Techniques

Before presenting the details of our proposed mechanism, we study in this subsection a few basic techniques and show that the input encryption based on these techniques along may result in an unsatisfactory mechanism.

However, the analysis will give insights on how a stronger mechanism should be designed. Note that to simplify the presentation, we assume that the cloud server honestly performs the computation, and defer the discussion on soundness to a later section.

C. Enhanced Techniques via Affine Mapping

To enhance the security strength of LP outsourcing, we must be able to change the feasible region of original LP and at the same time hide output vector x during the problem input encryption. We propose to encrypt the feasible region of _ by applying an affine mapping on the decision variables x.

D. Result Verification

Till now, we have been assuming the server is honestly performing the computation, while being interested learning information of original LP problem. However, such semihonest model is not strong enough to capture the adversary behaviors in the real world. In many cases, especially when the computation on the cloud requires a huge amount of computing resources, there exists strong financial incentives for the cloud server to be "lazy". Note that in our design, the workload required for customers on the result verification is substantially cheaper than solving the LP problem on their own, which ensures the great computation savings for secure LP outsourcing.

E. The Complete Mechanism Description

Based on the previous sections, the proposed mechanism for secure outsourcing of linear programming in the cloud is summarized below.

- KeyGen(1k): Let K = (Q,M, r,_,). For the system initialization, the customer runs KeyGen(1k) to randomly generate a secret K, which satisfies Eq. (4).
- ProbEnc(K,_): With secret K and original LP problem _, the customer runs ProbEnc(K,_) to compute the encrypted LP problem _K = (A′,B′, b′, c′) from Eq. (3).
- ProofGen(_K): The cloud server attempts to solve the LP problem _K in Eq. (5) to obtain the optimal solution y. If the LP problem _K has an optimal solution, should indicate so and include the dual optimal solution (s, t). If the LP problem _K is infeasible, _ should indicate so and include the primal and the dual optimal solutions of the auxiliary problem in Eq. (8). If the LP problem _K is unbounded, y should be a feasible solution of it, and _ should indicate so and include the primal and the dual optimal solutions of Eq. (9), i.e. the auxiliary problem of the dual problem of _K. • ResultDec(K,_, y, _): First, the customer verifies y and _ according to the various cases. If they are correct, the customer computes x =My − r if there is an optimal solution or reports _ to be infeasible or unbounded accordingly; otherwise the customer outputs _, indicating the cloud server was not performing the computation faithfully.

IV. SECURITY ANALYSIS

A. Analysis on Correctness and Soundness Guarantee

We give the analysis on correctness and soundness guarantee via the following two theorems.

Theorem *1: Our scheme is a correct verifiable linear programming outsourcing scheme.*

Proof: The proof consists of two steps. First, we show that for any problem _ and its encrypted version _K, solution y computed by honest cloud server will always be verified successfully. This follows directly from the duality theorem of linear programming. Namely, all conditions derived from duality theorem and auxiliary LP problem construction for result verification are necessary and sufficient. Next, we show that correctly verified solution y always corresponds to the optimal solution x of original problem _. For space limit, we only focus on the normal case. The reasoning for infeasible/unbounded cases follows similarly. By way of contraction, suppose x = My − r is not the optimized solution for _. Then, there exists x_ such that cT x_ < T x, where Ax_ = b and Bx_ ≥ 0. Since x_ = My_ − r, it is straightforward that cTMy_ − Ct r = cT x_ < cT x = cTMy−cT r, where A′y_ = b′ and B′y_ ≥ 0. Thus, y_ is a better solution than y for problem _K, which contradicts the fact that the optimality of y has been correctly verified. This completes the proof of **theorem** 1.

Theorem *2:* Our scheme is a sound verifiable linear programming outsourcing scheme. *Proof:* Similar to correctness argument, the soundness of the proposed mechanism follows from the facts that the LP problem _ and _K areequivalent to each other through affine mapping, and all the conditions thereafter for result verification are necessary and sufficient.

B. *Analysis on Input and Output Privacy Guarantee*

We now analyze the input and output privacy guarantee. Note that the only information that the cloud server obtains is _K = (A′,B′, b′, c′). We start from the relationship between the primal problem _ and its encrypted one _K. First of all, the matrix A and the vector b are protected perfectly. Because for _ m × n matrix A′ that has the full row rank and _ n × 1 vector b′, _ a tuple (Q,M, r) that transforms (A, b) into (A′, b′). This is straightforward since we can always find invertible matrices Q,M for equivalent matrices A and A′ such that A′ = QAM, and then solve r from b′ = Q(b + Ar).

Thus from (A′, b′), cloud can only derive the rank and size information of original equality constraints A, but nothing else. Secondly, the information of matrix B is protected by B′ = (B − _QA)M. Recall that the n × m matrix _ in the condition _b′ = Br is largely underdetermined. Namely, for each m×1 row vector in _, there are m−1 elements that can be set freely. Thus, the abundant choices of _, which can be viewed as encryption key with large key space, ensures that B is well obfuscated. Thirdly, the vector c is protected well by scaling factor and M. By multiplication of matrix M, both the elements and the structure pattern of c are no longer exposed from c′ = MTc. As for the output, since M, r is kept as a one-time secret and drawn uniformly at random, deriving x = My − r solely from y can be hard for cloud. Given the complementary relationship of primal and dual problem, it is also worth looking into the input/output privacy guarantee from dual problems of both _ and _K. Same as eq. (6), the dual problem of _ is defined as, maximize bT_subject to AT_+BT_ = c, _ ≥ 0, (10) IEEE

TRANSACTIONS ON CLOUD COMPUTING

April 10-15, 2011 7

V. PERFORMANCE ANALYSIS

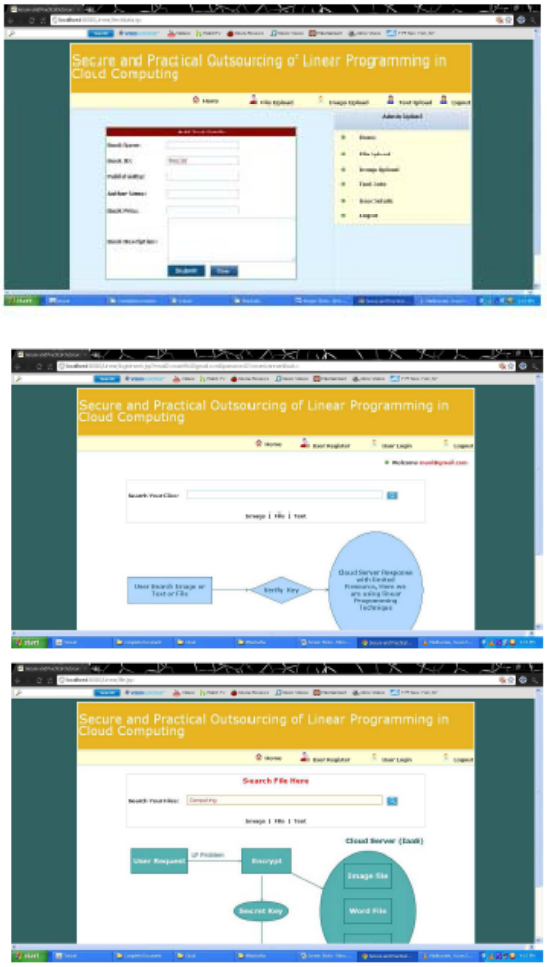
A. Theoretic Analysis

1) Customer Side Overhead: According to our mechanism, customer side computation overhead consists of key generation, problem encryption operation, and result verification, the computation complexity of these two algorithms are upper bounded via O(n2). i.e.,O(n_) for some 2 < _ ≤ 3. In our experiment, the matrix multiplication is implemented via standard cubic-time method,

2) Server Side Overhead: For cloud server, its only computation overhead is to solve the encrypted LP problem _K as well as generating the result proof _, both of which correspond to the algorithm ProofGen. If the encrypted LP problem _K belongs to normal case, cloud server just solves it with the dual optimal solution as the result proof _, which is usually readily available in the current LP solving algorithms and incurs no additional cost for cloud (see Section III-D).

B. *Experiment Results*

We now assess the practical efficiency of the proposed secure and verifiable LP outsourcing scheme with experiments.



REFERENCES

[1] P. Mell and T. Grance, “Draft nist working definition of cloud computing,” Referenced on Jan. 23rd, 2010 Online at http://csrc.nist.gov/groups/SNS/cloud-computing/index.html, 2010.

[2] Cloud Security Alliance, “Security guidance for critical areas of focus in cloud computing,” 2009, online at http://www.cloudsecurityalliance.org.

[3] C. Gentry, "Computing arbitrary functions of encrypted data," *Commun. ACM*, vol. 53, no. 3, pp. 97–105, 2010.

[4] Sun Microsystems, Inc., "Building customer trust in cloud computing with transparent security," 2009, online at https://www.sun.com/offers/ details/sun transparency.xml.

[5] D. Benjamin and M. J. Atallah, "Private and cheating-free outsourcing of algebraic computations," in *Proc. of 6th Conf. on Privacy, Security, and Trust (PST)*, 2008, pp. 240–245.

[9] R. Gennaro, C. Gentry, and B. Parno, "Noninteractive verifiable computing: Outsourcing computation to untrusted workers," in *Proc. Of CRYPTO'10*, Aug. 2010.

[10] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure ranked keyword search over encrypted cloud data," in *Proc. Of ICDCS'10*, 2010.

❑❑❑

Data Integrity Proofs

Jyoti Nandimath, Manish Ranglani, Rohit Shinde & Pushkar Wangikar

Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune - 411046
E-mail : adap jyotign@gmail.com, Manish.ranglani07@gmail.com

Abstract - Remote storage has been envisioned as the de-facto solution to the rising storage costs of IT Enterprises. With the high costs of data storage devices as well as the rapid rate at which data is being generated it proves costly for enterprises or individual users to frequently update their hardware. Apart from reduction in storage costs data outsourcing to the remote storage also helps in reducing the maintenance. Remote storage moves the user's data to large data centers, which are remotely located, on which user does not have any control. However, this unique feature of the remote storage poses many new security challenges which need to be clearly understood and resolved.

One of the important concerns that need to be addressed is to assure the customer of the integrity i.e. correctness of his data in the remote storage. As the data is physically not accessible to the user the remote storage should provide a way for the user to check if the integrity of his data is maintained or is compromised. In this paper we provide a scheme which gives a proof of data integrity in the remote storage which the customer can employ to check the correctness of his data in the storage. This proof can be agreed upon by both remote storage the and the customer and can be incorporated in the Service level agreement (SLA). This scheme ensures that the storage at the client side is minimal which will be beneficial for thin clients.

I. INTRODUCTION

In this paper we deal with the problem of implementing a protocol for obtaining a proof of data possession in the remote storage sometimes referred to as Proof of retrievability (POR).This problem tries to obtain and verify a proof that the data that is stored by a user at a remote data storage not modified by the archive and thereby the integrity of the data is assured. Such kinds of proofs are very much helpful in peer-to-peer storage systems, network file systems, long term archives, web-service object stores, and database systems. Such verification systems prevent the storage archives from misrepresenting or modifying the data stored at it without the consent of the data owner by using frequent checks on the storage archives. Such checks must allow the data owner to efficiently, frequently, quickly and securely verify that the remote storage archive is not cheating the owner. Cheating, in this context, means that the storage archive might delete some of the data or may modify some of the data. It must be noted that the storage server might not be malicious; instead, it might be simply unreliable and lose or inadvertently corrupt the hosted data. But the data integrity schemes that are to be developed need to be equally applicable for malicious as well as unreliable storage servers. Any such proofs of data possession schemes do not, by itself, protect the data from corruption by the archive. It just allows detection of tampering or deletion of a remotely located file at an unreliable remote storage server. To ensure file robustness other kind of techniques like data redundancy across multiple systems can be maintained.

II. SURVEY OF EXISTING SYSTEM

The simplest Proof of retrivability (POR) scheme can be made using a keyed hash function hk(F). In this scheme the verifier, before archiving the data file F in the storage, pre-computes the cryptographic hash of F using hk(F) and stores this hash as well as the secret key K. To check if the integrity of the file F is lost the verifier releases the secret key K to the remote storage archive and asks it to compute and return the value of hk(F). By storing multiple hash values for different keys the verifier can check for the integrity of the file F for multiple times, each one being an independent proof.

Though this scheme is very simple and easily implementable the main drawback of this scheme are the high resource costs it requires for the implementation. At the verifierside this involves storing as many keys as the number of checks it want to perform as well as the hash value of the data file F with each hash key. Also computing hash value for even a moderately large data files can be computationally

burdensome for some clients(PDAs, mobile phones, etc).

Ari Juels and Burton S. Kaliski Jr proposed a scheme called Proof of retrievability for large files using ”sentinels”. In this scheme, unlike in the key-hash approach scheme, only a single key can be used irrespective of the size of the file or the number of files whose retrievability it wants to verify. Also the archive needs to access only a small portion of the file F unlike in the key-has scheme which required the archive to process the entire file F for each protocol verification. This small portion of the file F is in fact independent of the length of F.

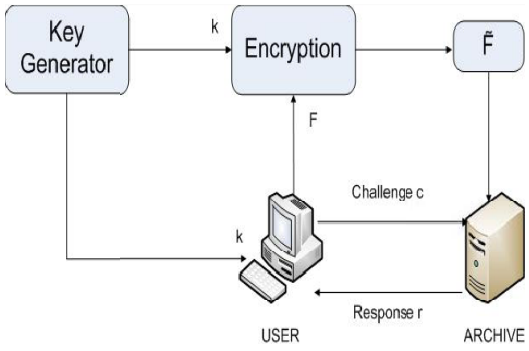
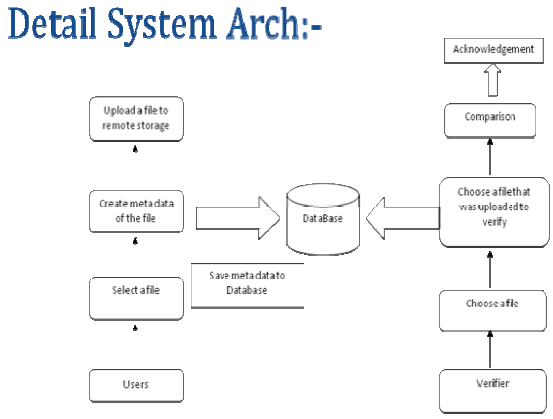


Fig. 1 : Schematic view of a proof of retriveability based on inserting random sentinels in the data file F.

In this scheme special blocks (called sentinels) are hidden among other blocks in the data file F. In the setup phase, the verifier randomly embeds these sentinels among the data blocks. During the verification phase, to check the integrity of the data file F, the verifier challenges the prove by specifying the positions of a collection of sentinels and asking the prover to return the associated sentinel values. If the prover has modified or deleted a substantial portion of F, then with high probability it will also have suppressed a number of sentinels. It is therefore unlikely to respond correctly to the verifier.To make the sentinels indistinguishable from the data blocks, the whole modified file is encrypted and stored at the archive. The use of encryption here renders the sentinels indistinguishable from other file blocks. This scheme is best suited for storing encrypted files. As this scheme involves the encryption of the file F using a secret key it becomes computationally cumbersome especially when the data to be encrypted is large. Hence, this scheme proves disadvantages to small users with limited computational power (PDAs, mobile phones etc.). There will also be a storage overhead at the server, partly due to the newly inserted sentinels and partly due to the error correcting codes that are inserted. Also the client needs to store all the sentinels with it, which may be a storage overhead to thin clients (PDAs, lowpower devices etc.).

III. PROPOSED SYSTEM

Taking into consideration the drawback of the existing work that uses the Hash function h(f) its makes user to store lots of keys n also need more memory that adds to the cost of it and also the encryption of the whole file. So, instead of creating an encryption of whole file it will be easy to use only some bits of the file. Encryption will be done on few randomly selected bits from the file rather than the whole file. The few randomly selected bits will be stored as the meta data with us. This bits will be use to check out the integrity of the file.



Now by selecting the random bits helps us to do the encryption of some part of the file rather than the whole. The client storage overhead is also minimized as it does not store any data with it. Hence our scheme suits well for thin clients.

In our data integrity protocol the verifier needs to store only a single cryptographic key - irrespective of the size of the data file F- and two functions which generate a random sequence. The verifier does not store any data with it. The verifier before storing the file at the archive, preprocesses the file and appends some meta data to the file and stores at the archive. At the time of verification the verifier uses this meta data to verify the integrity of the data.

Only comparison of meta data will not surely give the accurate result. In order to increase the probability of finding the changes we also calculate the ASCII of all the characters of each line. The ASCII total will also be stored along with the meta data at the client side.

It is important to note that our proof of data integrity protocol just checks the integrity of data i.e. if

the data has been illegally modified or deleted. It does not prevent the archive from modifying the data. In order to prevent such modifications or deletions other schemes like redundant storing etc, can be implemented which is not a scope of discussion in this paper.

IV. A DATA INTEGRITY PROOF BASED ON SELECTING RANDOM BITS IN DATA BLOCKS:

The client before storing its data file F at the client should process it and create suitable meta data which is used in the later stage of verification the data integrity at the storage. When checking for data integrity the client queries the storage for suitable replies based on which it concludes the integrity of its data stored in the client.

A. Setup phase:

Let the verifier V wishes to the store the file F with the archive. Let this file F consist of n file blocks. We initially pre process the file and create metadata to be appended to the file. Let each of the n data blocks have m bits in them. A typical data file F which the client wishes to store as shown in Figure 1. The initial setup phase can be described in the following steps:

1) **Generation of meta-data:** Generation of meta data is a unction G(f) that generates to numbers as start point and end point. The data between the start and the end point will be the meta data created for that line. G(i,j) where i- start point, j- End point. Hence g(i, j) gives the meta data from the ith to the jth bit.

   The meta data will be stored in the data base in form of table. The table will also contain the fields like length of line, staring point, ending point, meta data, Encryption of meta data and the ASCII total.

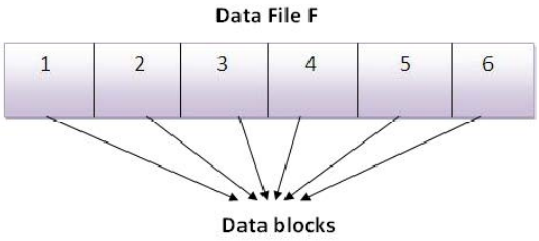


Fig. 1 : A data file F with 6 data blocks

2) **Encrypting the meta data:** Each of the meta data from the data blocks mi is encrypted by using a suitable algorithm to give a new modified meta data Mi. Without loss of generality we show this process by using a simple XOR operation. Let h be a function which generates a k bit integer _i for each i. This function is a secret and is known only to the verifier V .

   h : i ! _i, _i 2 {0..2n}

   For the meta data (mi) of each data block the number _i is added to get a new k bit number Mi.

   Mi = mi + _i

   In this way we get a set of n new meta data bit blocks. The encryption method can be improvised to provide still stronger protection for verifiers data.

3) **Appending of meta data:** All the meta data bit blocks that are generated using the above procedure are to be concatenated together. This concatenated meta data should be appended to the file F before storing it at the server. The file F along with the appended meta data e F is archived with the remote storage. Figure 4 shows the encrypted file e F after appending the meta data to the data file F.

Fig. 2 : The encrypted file eF which will be stored in remote storage.

B. Verification phase:

1) Let the verifier V want to verify the integrity of the file F. It throws a challenge to the archive and asks it to respond. The challenge and the response are compared and the verifier accepts or rejects the integrity proof. Suppose the verifier wishes to check the integrity of the file. The verifier challenges the remote storsge. A data block of the file F with random bits selected in it Fig. 1. The encrypted file eF which will be stored in the remote storage. specifying the block number i and a bit number j generated by using the function g which only the verifier knows. The verifier also specifies the position at which the meta data corresponding to the original file. This meta data will be a k-bit number. Using the number I and j i.e start and end points we will extract a new meta data from the newly arrived file. The meta data with us will be then compared to the newly created meta data of the same file. Any mismatch between the two would mean a loss of the integrity of the clients data at the remote storage.

2) In order to increase the probability of finding changes made the ASCII total of each line is calculate n compared with the ASCII total of that line already with client in the data base. Any mismatch between the two would mean a loss of the integrity of the clients data at the remote storage.

V. ADVANTAGE AND DISADVANTAGES

A. Advantages:

1) Using camparision of meta data and comparision of ASCII total increases the probability of finding the changes occurred in file.

2) It allows encryption of small part of file instead of whole file.

B. Disadvantage:

1) It does not detects if the characters of the line are swaped by the changer.

2) Size of Meta data created will depend on the size of the file.

VI. FUTURE WORK

In this project we have worked to facilitate the client in getting a proof of integrity of the data which he wishes to store in the remote storage servers with bare minimum costs and efforts. Our scheme was developed to reduce the computational and storage overhead of the client as well as to minimize the computational overhead of the remote storage server. We also minimized the size of the proof of data integrity so as to reduce the network bandwidth consumption. At the client we only store two functions, the bit generator function g, and the function h which is used for encrypting the data. Hence the storage at the client is very much minimal compared to all other schemes that were developed. Hence this scheme proves advantageous to thin clients like PDAs and mobile phones.

The operation of encryption of data generally consumes a large computational power. In our scheme the encrypting process is very much limited to only a fraction of the whole data thereby saving on the computational time of the client. Many of the schemes proposed earlier require the archive to perform tasks that need a lot of computational power to generate the proof of data integrity. But in our scheme the archive just need to fetch and send few bits of data to the client.

It should be noted that our scheme applies only to static storage of data. It cannot handle to case when the data need to be dynamically changed. Hence developing on this will be a future challenge.

In this it also doesn't provide all detection of changes in file i.e already discussed that character of same line when swapped cant be detected. Hence developing on this will be a future challenge.

REFRENCES:

1) Data Integrity Proofs in Cloud Storage.

By- Sravan Kumar R Software Engineering and Technology labs Infosys Technologies Ltd Hyderabad, India Email: sravan r@infosys.com Ashutosh Saxena Software Engineering and Technology labs Infosys Technologies Ltd Hyderabad, India Email: ashutosh saxena01@infosys.com

2) E. Mykletun, M. Narasimha, and G. Tsudik, "Authentication and integrity in outsourced databases," Trans. Storage, vol. 2, no. 2, pp. 107–138, 2006.

3) D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in SP '00: Proceedings of the 2000 IEEE Symposium on Security and Privacy. Washington, DC, USA: IEEE Computer Society, 2000, p. 44.

4) A. Juels and B. S. Kaliski, Jr., "Pors: proofs of retrievability for large files," in CCS '07: Proceedings of the 14th ACM conference on Computer and communications security. New York, NY, USA: ACM, 2007, pp. 584–597.

❑❑❑

XML Based Web Usage Mining In Server Logs

Y.S.S.R Murthy, L.Balaji & Lakshmi Tulasi.Ambati

Computer Science Engineering, ShriVushnu Engineering College For Women,
Vishnupur,Bhimavaram, West Godavari District, AndhraPradesh, India

Abstract - Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to improve web based applications. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use these information for the specific needs**.** Web log files are the primary data source for web usage mining. This usage analysis includes tasks like page aces frequency, finding the common traversal paths through a website. These log files contains information that can't be directly interpreted ,for example information like who is accessing, which pages are accessing by whom, how much time user is accessing a particular page ,can't be obtained directly these log files. Since log files are unformatted text files , complex to interpret and analyze. In this paper we propose a novel approach using universally accepted formatting language XML. In our approach text based log files are converted into XML format using parsers. Once log file is in XML format , using DOM API or other types of parser API's we can retrieve the required information in an easy manner such as user and session identification and the paths that are frequently accessed. This paper presents several data preparation techniques based on XML parsers in order to increase the usability of websites.

I. INTRODUCTION

Web usage mining tries to make sense of the data generated by the web surfer's sessions or behaviors. While the Web content and structure mining utilize the real or primary data on the web, Web usage mining mines the secondary data derived from the interactions of the users while interacting with the web. The web usage data includes the data from web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls and any other data as the results of interactions [1]. Web usage mining focuses on data collection, preprocessing and the data mining techniques.

1.1 WEB LOG DATA

In Web Mining, data can be collected at the server side, client side and proxy servers. Each type of data collection only differs not only in terms of the location of the data source, but also the kinds of data available and its methods of implementation. Logs are mostly stored simply as text files, each line corresponding to one access (i.e. one request). The most widely used log file formats are, implied by the Common Log File format (CLF) and the Extended Log File format (Ex LF)

1.1.1 Server Level Collection

A Web server log is an important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the (possibly concurrent) access of a Web site by multiple users. The Web server can also store other kinds of usage information such as cookies and query data in separate logs [5]. Limitations

- However, the site usage data recorded by server logs may not be entirely reliable due to the presence of various levels of caching within the Web environment. Cached page views are not recorded in a server log.
- In addition, any important information passed through the POST method will not be available in a server log. Packet sniffing technology is an alternative method to collecting usage data through server logs. Packet sniffers monitor network traffic coming to a Web server and extract usage data directly from TCP/IP packets.

1.1.2 Client Level Collection

Client-side data collection can be implemented by using a remote agent [3] (such as JavaScript or Java

applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities. The implementation of client-side data collection methods requires user cooperation, either in enabling the functionality of the Java scripts and Java applets, or to voluntarily use the modified browser.

Advantage:

Client-side collection has an advantage over server-side collection because it ameliorates both the caching and session identification problems. However, Java applets perform no better than server logs in terms of determining the actual view time of a page.

Limitations:

- It may incur some additional overhead especially when the Java applet is loaded for the first time.
- Java scripts, on the other hand, consume little interpretation time but cannot capture all user clicks (such as reload or back buttons). These methods will collect only single-user, single-site browsing behavior.
- A modified browser is much more versatile and will allow data collection about a single user over multiple Websites. The most difficult part of using this method is convincing the users to use the browser for their daily browsing activities.

1.1.3 Proxy Level Collection

A Web proxy [4] acts as an intermediate level of caching between client browsers and Web servers.

- Proxy caching can be used to reduce the loading time of a Web page experienced by users as well as the network traffic load at the server and client sides.
- The performance of proxy caches depends on their ability to predict future page requests correctly.
- Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data source for characterizing the browsing behavior of a group of anonymous users sharing a common proxy server.

1.2 WEB LOG FORMAT

In this paper we take W3C extended log format shown below

date time c-ip cs-username s-sitename s-computername s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status time-taken cs-version cs-host cs(User-Agent) cs(Referer)

W3C Extended Logging Field Definitions

Prefix	**Meaning**
s	Server actions.
c	Client actions.
cs	Client-to-server actions.
sc	Server-to-client actions.

Table 1 : Web log filed prefixes and their meanings

Field	**Appeared as**	**Meaning**
Date	date	The date that the activity occurred.
Time	time	The time that the activity occurred.
Client IP Address	c-ip	The IP address of the client that accessed your server.
User Name	cs-username	The name of the authenticated user who accessed your server. This does not include anonymous users, who are represented by a hyphen (-).
Service Name	s-sitename	The Internet service and instance number that was accessed by a client.
Server Name	s-computername	The name of the server on which the log entry was generated.
Server IP Address	s-ip	The IP address of the server on which the log entry was generated.
Server Port	s-port	The port number the client is connected to.
Method	cs-method	The action the client was trying to perform (for example, a GET method).
URI Stem	cs-uri-stem	The resource accessed; for example, Default.htm.

URI Query	cs-uri-query	The query, if any, the client was trying to perform.
Protocol Status	sc-status	The status of the action, in HTTP or FTP terms.
Bytes Sent	sc-bytes	The number of bytes sent by the server.
Bytes Received	cs-bytes	The number of bytes received by the server.
Time Taken	time-taken	The duration of time, in milliseconds, that the action consumed.
Protocol Version	cs-version	The protocol (HTTP, FTP) version used by the client. For HTTP this will be either HTTP 1.0 or HTTP 1.1.
Host	cs-host	Displays the content of the host header.
User Agent	cs(User-Agent)	The browser used on the client.
Cookie	cs(Cookie)	The content of the cookie sent or received, if any.
Referrer	cs(Referer)	The previous site visited by the user. This site provided a link to the current site.

Table 2 : web log fields and their meanings

Log files were designed to produce site-level performance statistics. It's thus no surprise they can't provide even the minimum information needed to effectively investigate a potential usability problem. Here are some specific ways log files provide insufficient or misleading data:

• *Who is visiting your site*. For you to know who is visiting your site, the log file must contain a person ID such as a login to the server or to the user's own computer. However, most web sites do not require users to log in, and most web servers do not make a "back door" request to learn the user's login identity on his/her own computer.

• *The path visitors take through your pages*. The path that visitors follow within your site is clear if the log file contains an entry for every page viewed. However, when browsers are set to view pages from cache (usually the default), or when corporate or ISP servers retrieve pages from a central cache, then some pages will not be logged by the web server and the log file will have gaps. For example, with caching, pages viewed using the Back button typically are not logged.

In addition, nothing appears in the log file when visitors arrived at a page by typing its URL, using a bookmark, or following an email link. In these cases one can try to infer from Referrer data.

• *How much time visitors spend on each page*. The log file records the time when a data transmission was initiated, but not the time when the transfer was completed. In addition, it is unclear when during the download process the user began viewing a page. However, by comparing the timestamps of the current request and the next request, you can calculate roughly how much time a visitor is spending on a page— unless the visitor walks away while the computer is displaying the page. Some timing details may also be obtained by analyzing the transmission of graphics files associated with a page.

• *Where visitors are leaving your site*. The log file records the last page transferred by the server for that user session, but there are two reasons why it might not be the last page viewed. First, the last page viewed may have been displayed from cache. Second, the user may have left his/her workstation for a period of time that exceeds what the log analysis software regards as a session.

1.3 Using XML format of web log file

The log files contain information that can't be directly interpreted. for example information like who is accessing, which pages are accessing by whom ,how much time user is accessing a particular page ,can't be obtained directly these log files. Since log files are unformatted text files , complex to interpret and analyze. In our approach text based log files are converted into XML format using parsers. Once log file is in XML format , using DOM API or others types of parser API's we can retrieve the required information in an easy manner.

The general log file format is shown below

date time c-ip cs-username s-sitename s-computername s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status time-taken cs-version cs-host cs(User-Agent) cs(Referer)

and it's equivalent XML formatted file is shown below.

```
<web-log>
   <record>
      <date> </date>
          <time> </time>
          <c-ip> </c-ip>
          <cs-username> </cs-username>
          <s-sitename> </s-sitename>
          <s-computername> </s-computername>
          <s-ip>   </s-ip>
          <s-port> </s-port>
          <cs-method> </cs-method>
          <cs-uri-stem> </cs-uri-stem>
          <cs-uri-query> </cs-uri-query>
          <sc-status> </sc-status>
          <sc-win32-status> </sc-win32-status>
          <sc-bytes> </sc-bytes>
          <cs-bytes> </cs-bytes>
          <time-taken> </time-taken>
          <cs-version> </cs-version>
          <cs-host> </cs-host>
          <cs-User-Agent> </cs-User-Agent>
          <cs-Cookie>        </cs-Cookie>
<cs-Referer> </cs-Referer>
   </record>
</web-log>
```

II. OUR APPROACH

Our XML based web usage mining process is shown below figure

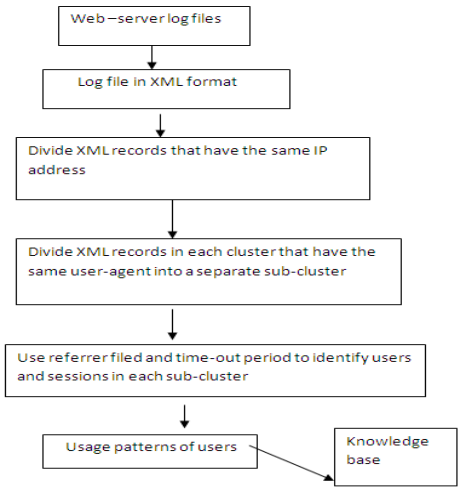


Fig.1: XML based web usage mining process

2.1 Data cleaning

The first task to do is data cleaning, which should remove entries unhelpful to data analyzing and mining. Firstly, it should remove entries that have status of "error". Secondly, some access records generated by automatic search engine agent should be identified and removed from the access log,. Primarily, it should identify log entries created by so called crawlers that are used widely in Web Information Retrieval Search engine tools. Such data offer retrieval mechanisms nothing to the analyzing of user navigation behaviors. Many crawlers voluntarily declare themselves in agent field of access log, so a simple string match during the data cleaning phase can strip off a significant amount of agent traffic. In addition, to exclude these accesses, employs several heuristic methods that are based on indicators of non-human behavior. These indicators are (1) the repeated request for the same URL from the same host; (2) a time interval between requests too short to apprehend the contents of a page; and (3) a series requests from one host all of whose referrer URLs are empty. The referrer URL of a request is empty if the URL was typed in, requested using a bookmark, or requested using a script. The last task of data cleaning, which is also disputable is whether it needs to remove log entries covering image, sound, and video files. once log file is converted into XML format we start data cleaning phase in which we remove log entries involving image files and failed requests.

Next user identification phase starts. in this phase we group log records having the same IP address. the different group suggests different IP address. Same IP address doesn't means to be a single user, because when using proxy servers, different users requests will come from same IP address. therefore in each cluster, we again group log records having the same user-agent into a sub-cluster.i,e this cluster contains log records of same IP address and same user-agent.

After this step, log records in each cluster having the same IP address are again sub-clustered , ie same IP addresses. But we can't simply judge, same IP address and same user-agent means a single user so we use the referrer filed to find users and sessions using time-out periods.If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, this indicates there is another user with the same IP Address. Using time-out periods , if the time between page requests exceeds a certain limit, we can assume that the user is starting anew session.

Data Pre-processing can be done using XML (Extended Markup Language). XML provides a structure to the records which are present in web logs. Hence, understanding of web logs becomes easier. Logs

recorded in the web log which is a text file are converted into DOM tree structure using XML parsers.

Consider the following sample log file

1. 192.168.5.234 - [10/nov/2009:03:0:01 -0500] "GET A.aspx HTTP/1.0" 200 3290 – IE5/Win2k

2. 192.168.5.234 - [10/nov//2009:03:0:09 -0500] "GET B.aspx HTTP/1.0" 200 2050 a.ASPX IE5/Win2k

3. 196.132.0.21- [10/nov/2009:03:0:10 -0500] "GET C.ASPX HTTP/1.0" 200 4130 - IE4/Win98

4. 196.132.0.21- [10/nov/2009:03:00:12 -0500] "GET B.ASPX HTTP/1.0" 200 5096 C.ASPX- IE4/Win98

5. 196.132.0.21- [10/nov/2009:03:0:15 -0500] "GET E.ASPX HTTP/1.0" 200 3290 C.ASPX IE4/Win98

6. 192.168.5.234 – [10/nov/2009:03:0:19 -0500] "GET C.ASPX HTTP/1.0" 200 2050 A.ASPX IE5/Win2K

7. 196.132.0.21- [10/nov/2009:03:00:22 -0500] "GET D.ASPX HTTP/1.0" 200 8140 B.ASPX IE4/Win98

8. 192.168.5.234 - [10/nov/2009:03:0:22 -0500] "GET A.ASPX HTTP/1.0" 200 1820 - IE4/Win98

9. 192.168.5.234 - [10/nov/2009:03:0:25 -0500] "GET E.ASPX HTTP/1.0" 200 2270 C.ASPX IE5/Win2k

10. 192.168.5.234 - [10/nov/2009:03:00:25 -0500]"GET C.ASPX HTTP/1.0" 200 7220 A.ASPX IE4/Win98

11. 192.168.5.234 - [10/nov/2009:03:10:33 -0500]"GET B.ASPX HTTP/1.0" 200 3290 C.ASPX IE4/Win98

12. 192.168.5.234 - [10/nov/2009:03:0:58 -0500]"GET D.ASPX HTTP/1.0" 200 3290B.ASPX IE4/Win98

13. 192.168.5.234 - [10/nov/2009:03:01:10-0500] "GET E.ASPX HTTP/1.0" 200 3290 D.ASPX IE4/Win98

14. 192.168.5.234 -[10/nov//2009:03:01:15-0500] "GET A.ASPX HTTP/1.0" 200 3290 - IE5/Win2k

15. 192.168.5.234 - [10/nov/2009:03:01:16-0500] "GET C.ASPX HTTP/1.0" 200 3290 A.ASPX IE5/Win2k

16. 192.168.5.234 -[10/nov/2009: 03:01:17-0500] "GET F.ASPX HTTP/1.0" 200 3290 C.ASPX IE4/Win98

17. 192.168.5.234 -[10/nov/2009: 03:01:25-0500] "GET F.ASPX HTTP/1.0" 200 3290 C.ASPX IE5/Win2k

18. 192.168.5.234 - [10/nov/2009: 03:01:30 -0500] "GET B.ASPX HTTP/1.0" 200 3290 A.ASPX IE5/Win2k

19. 192.168.5.234 - [10/nov/2009: 03:01:36 -0500] "GET D.ASPX HTTP/1.0" 200 3290 B.ASPX IE5/Win2k

Fig. 2 : a sample web log file

The above log file is converted into XML based log file using XML parsers Since all fields are not required for analysis we take data and time and client IP address and referrer and cs-uri-stem fields only. As shown below.

```
<record>
 <date> 25/Apr/2009 </date>
 <time> 03:04:41 </time>
 <c-ip> 192.168.5.234</c-ip>
 <cs-Referer> -</cs-Referer>
 <cs-uri-stem> A.aspx </cs-uri-stem>
</record>
<record>
 <date> 25/Apr/2009 </date>
 <time> 03:05:34 </time>
 <c-ip> 192.168.5.234</c-ip>
 <cs-Referer>- </cs-Referer>
 <cs-uri-stem>L.aspx </cs-uri-stem>
</record>
<record>
 <date> 25/Apr/2009 </date>
 <time> 03:05:39 </time>
 <c-ip> 192.168.5.234</c-ip>
 <cs-Referer>A.ASPX </cs-Referer>
 <cs-uri-stem>B.ASPX </cs-uri-stem>
</record>
<record>
 <date> 25/Apr/2009 </date>
 <time> 03:06:02 </time>
 <c-ip>192.168.5.234 </c-ip>
 <cs-Referer> -</cs-Referer>
 <cs-uri-stem> A.ASPX</cs-uri-stem>
</record>
……..
```

Fig. 3: a partial XML file after completion of data cleaning phase

2.2 USER AND SESSION IDENTIFIACTION*:*

After data cleaning the XML file is analyzed for user and session identification. For this task we used DOM API and XSLT to transform XML tree in any form required for our analysis. The following sample code shows our approach of extracting only required fields ,I,e time, IP address, URL, referrer fields.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0":
  - <xsl:template match="web-log/record">
    - <tr>
      - <td>
          <xsl:vlaue-of select="time" />
        </td>
      - <td>
          <xsl:vlaue-of select="c-ip" />
        </td>
      - <td>
          <xsl:vlaue-of select="cs-Referer" />
        </td>
      - <td>
          <xsl:vlaue-of select="cs-uri-stem" />
        </td>
      </tr>
    </xsl:template>
  </xsl:stylesheet>
```

Fig. 4 : An XSLT transformation applied to cleaned XML log file

Using DOM API we can access any element in the tree by means of simple methods. User –identification task is is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers .For example: If the IP address of the same, but the proxy information has changed, indicating that the user may be behind a firewall in a different users within the network, you can mark for different users; also can access information, refer to institutions with information and site topology to construct a user's browsing path, if the current page request to drop the user has viewed the page with no link between the existence of IP addresses that the same number of users . The Web Usage Mining methods that rely on user cooperation are the easiest ways to deal with this problem. However, even for the log/site based methods, there are heuristics that can be used to help identify unique users. Even if the IP address is the same, if the agent log shows a change in browser software or operating system, a reasonable assumption to make is that each different agent type for an IP address represents a different user.. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, again, the heuristic assumes that there is another user with the same IP address.

Identifying the user during a session in another issue is determining whether the access log is not an important record of the request. This requires the path to add to complete these records. If the current page the user requested the last page of the http request is no hypertext links, the user may use the browser "BACK" function call to cache the page in the machine. Check reference information to determine which page from the current request, if the user access to record the history of more than one page contains a link to the page with the current request, the request is the time closest to the source as the current request, if the reference information is not complete, you can take advantage of the topology of the site instead .For logs that span long periods of time, it is very likely that users will visit the Web site more than once. The goal of session identification is to divide the page accesses of each user into individual sessions. The simplest method of achieving this is through a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout, and established a timeout of 25.5 minutes based on empirical data. Once a site log has been analyzed and usage statistics obtained, a timeout that is appropriate for the specific Web site can be fed back into the session identification algorithm.

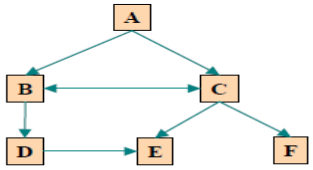
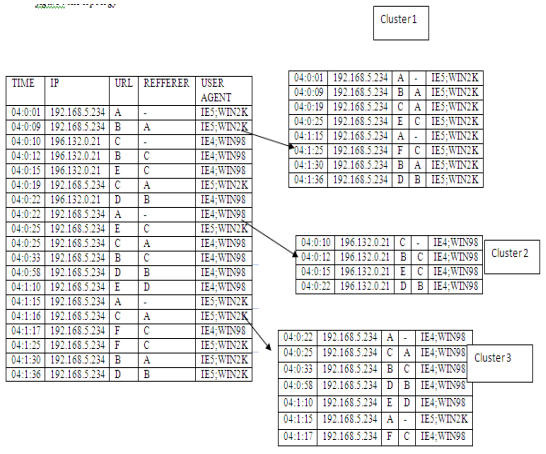


Fig. 5 : site topology



TIME	IP	URL	REFFERER	USER AGENT
04:0:01	192.168.5.234	A	-	IE5;WIN2K
04:0:09	192.168.5.234	B	A	IE5;WIN2K
04:0:10	196.132.0.21	C	-	IE4;WIN98
04:0:12	196.132.0.21	B	C	IE4;WIN98
04:0:15	196.132.0.21	E	C	IE4;WIN98
04:0:19	192.168.5.234	C	A	IE5;WIN2K
04:0:22	196.132.0.21	D	B	IE4;WIN98
04:0:22	192.168.5.234	A	-	IE4;WIN98
04:0:25	192.168.5.234	E	C	IE5;WIN2K
04:0:25	192.168.5.234	C	A	IE4;WIN98
04:0:33	192.168.5.234	B	C	IE4;WIN98
04:0:58	192.168.5.234	D	B	IE4;WIN98
04:1:10	192.168.5.234	E	D	IE4;WIN98
04:1:15	192.168.5.234	A	-	IE5;WIN2K
04:1:16	192.168.5.234	C	A	IE5;WIN2K
04:1:17	192.168.5.234	F	C	IE4;WIN98
04:1:25	192.168.5.234	F	C	IE5;WIN2K
04:1:30	192.168.5.234	B	A	IE5;WIN2K
04:1:36	192.168.5.234	D	B	IE5;WIN2K

04:0:01	192.168.5.234	A	-	IE5;WIN2K
04:0:09	192.168.5.234	B	A	IE5;WIN2K
04:0:19	192.168.5.234	C	A	IE5;WIN2K
04:0:25	192.168.5.234	E	C	IE5;WIN2K
04:1:15	192.168.5.234	A	-	IE5;WIN2K
04:1:25	192.168.5.234	F	C	IE5;WIN2K
04:1:30	192.168.5.234	B	A	IE5;WIN2K
04:1:36	192.168.5.234	D	B	IE5;WIN2K

04:0:10	196.132.0.21	C	-	IE4;WIN98
04:0:12	196.132.0.21	B	C	IE4;WIN98
04:0:15	196.132.0.21	E	C	IE4;WIN98
04:0:22	196.132.0.21	D	B	IE4;WIN98

04:0:22	192.168.5.234	A	-	IE4;WIN98
04:0:25	192.168.5.234	C	A	IE4;WIN98
04:0:33	192.168.5.234	B	C	IE4;WIN98
04:0:58	192.168.5.234	D	B	IE4;WIN98
04:1:10	192.168.5.234	E	D	IE4;WIN98
04:1:15	192.168.5.234	A	-	IE5;WIN2K
04:1:17	192.168.5.234	F	C	IE4;WIN98

Fig. 6 : XML parser converted output file is clustered sub clusters based on IP address and user agent

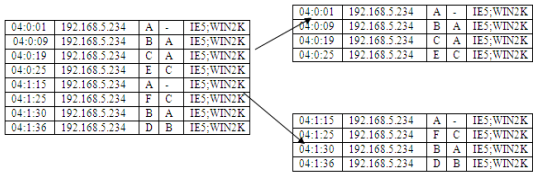


Fig. 7: sessions produced using time heuristic of 30 minutes on cluster 1

III. RESULTS

To validate the effectiveness and efficiency of our methodology mentioned above, we have made an experiment with the web server log. the data source size for our experiment is 11MB. Our experimental results are shown below. After data cleaning, the number of requests declined from 7890 to 3297. Finally, on the basis of user identification's results, we have identified 1945 sessions by a threshold of 30 minutes and path completion.

Entries in raw web	log Entries after data cleaning	Number of users	Number of sessions
7890	3297	1352	1945

The results are shown in the below bar chart

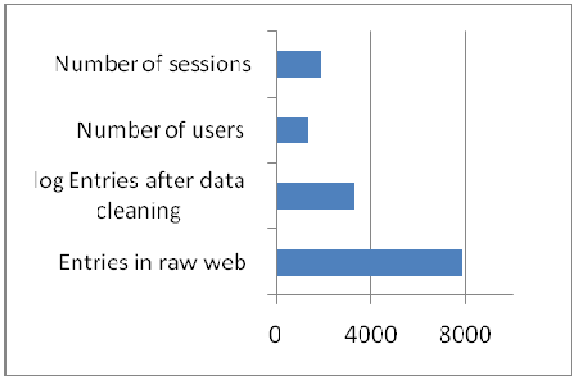


Fig. 8 : Results of our approach identifying users and sessions

IV. CONCLUSION

This paper has presented the details of data preprocessing tasks that are necessary for performing Web Usage Mining, the application of data mining and knowledge discovery techniques to web server access logs. We give some rules based on heuristics in every phase of data preprocessing in order to design and implement them easily. Our experiments have let us estimate data preprocessing importance and our methodology's effectiveness. It not only reduces the log file size but also increases the quality of the available data. As we used XML DOM technology, in the future XML technology based web services demand more, so using XML in our approach could increase effectiveness of the web usage mining process

REFERENCES

[1] Raymond Kosala, Hendrik Blockeel, "Web Mining Research: A Survey" , Katholieke Leuven, Belgium

[2] Miha Grcar, "User Profiling: Web usage Mining", Jozef Stefan Institute, Solvenia

[3] Yongjian Fu, Ming-Yi Shih, "A Framework for Personal Web Usage Mining", University of Missouri, Rolla.

[4] Jan Kerkhofts, Dr.Koen Vanhoof,"Web Usage Mining on Proxy Servers", Limburg University Centre,July 30,2001.

[5] Jaideep srivastava, Robert cooley, "University of Minnesota", Minneapolis

[6] Berendt B., Mobasher B., Nakagawa M., Spiliopoulou M.. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. Proc. WEBKDD 2002: Mining WebData for Discovery Usage Patterns and Profiles, LNCS 2703, Springer-Verlag, 2002:159-179

[7] Pirolli P., Pitkow J., Rao R.. Silk from a sow's ear: Extracting usable structures from the Web. In: Proc. 1996 Conference on Human Factors in Computing Systems (CHI-96), Vancouver, BritishColumbia, Canada, 1996.

[8] Tanasa D., Trousse B.. Advanced data preprocessing for intersites Web usage mining. IntelligentSystems, IEEE,2004(19): 59 – 65

[9] Catledge L., Pitkow J.. Characterizing browsing behaviors on the World Wide Web, ComputerNetworks and ISDN Systems ,1995,27(6):1065-1073.

[10] Chen M.S., Park J.S., Yu P.S.. Data mining for path traversal patterns in a web environment. InProceedings of the 16th International Conference on Distributed Computing Systems, 1996:385-392.

Secure Strategy for Privacy Preserving Association Rule Mining

V K S K Sai Vadapalli & G Loshma

Dept. of CSE, Sri Vasavi Engineering College, Pedatadepalli, Tadepalligudem—534101, India
E-mail : krishnasaibtech@gmail.com, loshma@gmail.com

Abstract - The advances of data mining techniques played an important role in many areas for various applications. In context of privacy and security issues, the problems caused by association rule mining technique are recently investigated. It is proved that the misuse of this technique may disclose the database owner's sensitive information to others which database owners do not want. Many of the researchers in this area have recently made effort to preserve privacy for sensitive knowledge in statistical database. We present a detailed overview and classification of approaches which have been applied to knowledge hiding in context of association rule mining. For centralized data, we propose an algorithm namely DSRRC (Decrease Support of R.H.S. item which provide privacy for sensitive association rules at certain level while ensuring data quality. Our proposed algorithm cluster the sensitive association rules based on certain criteria and hide as many as possible rules at a time by modifying fewer transactions in database. Less modifications in database helps to maintain data quality. We describe comparative Performance evaluation metrics and analyze the performance results of proposed algorithm and the existing algorithms for centralized data. Finally we conclude our work by defining some future trends.

Keywords - *Market Basket Analysis, Association rule mining, frequent item sets, Apriori algorithm, DSRRC Algorithm, Sensitive Rules.*

I. INTRODUCTION

Data mining services require accurate input data for their results to be meaningful, but privacy concerns may influence users to provide spurious information[1][2]. In order to preserve the privacy of the client in data mining process, a variety of techniques based on random perturbation of data records have been proposed recently. Randomization and Distortion are the two dominant methods provided as a means to preserve the privacy[1][3]. Randomization process modifies each transaction by replacing some of the existing items with non-existing items, And adding some fake items, thereby preserving the privacy. Distortion process operates on a transaction database by probabilistically changing some of the items in each transaction[4]. We focus on an improved distortion process that tries to enhance the accuracy by selectively modifying the list of items. The normal distortion procedure does not provide the flexibility of tuning the probability parameters for balancing privacy and accuracy parameters, and each item's presence/absence is modified with an equal probability. In improved distortion technique, frequent one item-sets, and non-frequent one item-sets are modified with a different probabilities controlled by two probability parameters *fp*, *nfp* respectively[5][6]. The owner of the data has a flexibility to tune these two probability parameters (*fp* and *nfp*) based on his/her requirement for privacy and accuracy. The experiments conducted on real time datasets confirmed that there is a significant increase in the accuracy at a very marginal cost in privacy[4][5][6].

A. Model of Data Miners

Two classes of data miners are considered in this system. One is legal data miners. These miners always act legally in that they perform regular data mining tasks and would never intentionally breach the privacy of the data. On the other hand, *illegal data miners* would purposely discover the privacy in the data being mined. Illegal data miners come in many forms. In this paper, we focus on a particular sub-class of illegal miners. That is, in our system, illegal data miners are *honest but curious*: they follow proper protocol (i.e., they are honest), but they may keep track of all intermediate communications and received transactions to perform some analysis (i.e., they are *curious*) to discover private information. Even though it is a relaxation from Byzantine behavior, this kind of honest but curious (nevertheless illegal) behavior is most common and has been widely adopted as an adversary model in the literatures. This is because, in reality, a workable system must benefit both the data miner and the data providers. For example, an online bookstore (the data miner) may use the association rules of purchase records to make recommendations to its customers (data providers). The data miner, as a long-term agent, requires large numbers of data providers to collaborate with. In other words,

even an illegal data miner desires to build a reputation for trustworthiness. Thus, honest but curious behavior is an appropriate choice for many illegal data miners.

B. *Randomization Model*

Let us consider the entire mining process as an iterative one. In each stage the data miner obtains a perturbed transaction from a different data provider. With the randomization approach, each data provider employs a randomization operator R(.) and applies it to one transaction to which the data provider holds. Upon receiving transactions from the data providers, the legal data miner must first perform an operation called *support recovery* which intends to filter out the noise injected in the data due to randomization, and then carry out the data mining tasks. At the same time, an illegal data miner may perform a particular privacy recovery algorithm in order to discover private data from that supplied by the data providers. Clearly, the system should be measured by its capability in terms of supporting the legal miner to discover accurate association rules, while preventing illegal miner from discovering private data.

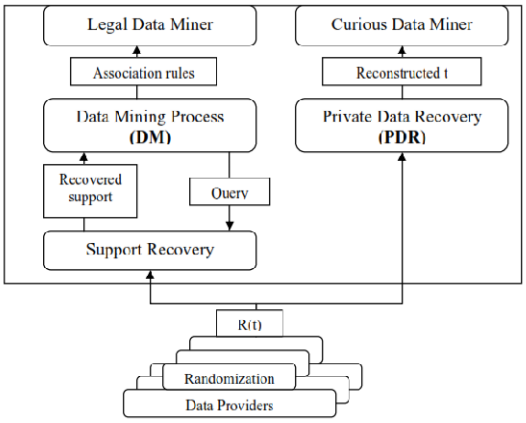


Fig 1.1 Infrastructure of a Typical Randomization System.

C. New Model

Figure 2 shows the infrastructure of the newly proposed system. The legal data miner contains two components, Data Mining process (**DM**) and Perturbation Guidance (**PG**). When a data provider *C* initializes a communication session, **PG** first dispatches a reference *V k* to *C I* . Based on the received *V* the data perturbation component of *C* transforms the transaction *t* to a perturbed one *R(t)* and transmits *R(t)* to **PG**. **PG** then updates *Vk I* based on the recently received *R(t)* and forwards *R(t)* to the Data Mining process **DM**. The key here is to properly design *V* so that correct guidance to data provider on how to distort the data transactions. In this system, *Vkk* is an algebraic quantity derived from *T (Transaction Database)* which enables us to effectively maintain the accuracy of data mining while significantly reduces the leakage of private information.

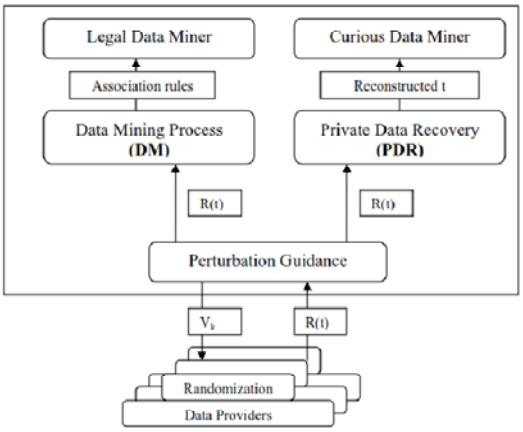


Fig 1.2 : Infrastructure of a newly Proposed Randomization System

D. Communication Protocol

The details of the communication protocol used between data providers and data miners are as follows.

On the side of the data miner there are two current threads that perform the following operations iteratively after initializing *Vk*

Thread of registering Data provider

R1: Negotiate on the truncation level *k* with a data provider

R2: Wait for a ready message from a data provider

R3: Upon receiving the ready message from a data provider

Register the data provider

Send the data provider current V_k

R4: Goto R1

Thread of Receiving data transaction

T1: Wait for a (perturbed) data transaction *R(t)* from a data provider

T2: Upon receiving the data transaction from a registered data provider,

Update V_k based on the newly received perturbed data transaction

Deregister the data provider

T3: Goto T1

For a data provider , it performs the following operations to transfer its transactions to the data miner.

P1: Send the data miner a ready message indicating that this provider is ready to contribute to the mining process.
P2: Wait for message that contains V_k from the data miner.
P3: Upon receiving the message from the data miner, compute $R(t)$ based on t and V_k
P4: Transfer $R(t)$ to the data miner.

This paper is organized as follows: Section II describes Basic Concepts of Association Rule Mining. Section III describes proposed work. Section IV describes implementation details. Section V presents experimental results and performance analysis. Section VI presents conclusion and future scope.

II. BASIC CONCEPTS OF ASSOCIATION RULE MINING

Association rule finds interesting associations and/or correlation relationships among large set of data items. Association rule shows attribute value conditions that occur frequently together in a given dataset. A typical and widely-used example of association rule mining is Market Basket Analysis. For example, data are collected using bar-code scanners in supermarket. Such 'market basket' databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together. They could use this data for adjusting store layouts (placing items optimally with respect to each other), for cross-selling, for promotions, for catalog design and to identify customer segments based on buying patterns. Association rules do not represent any sort of causality or correlation between the two item sets The problem of mining association rules can be described as below: if I = {I1, I2, I3,-------, In} is the set of items. Suppose D is database transaction set and each transaction T contains set of items, such that T⊆ I. Each transaction has identifier called as TID i.e. transaction id. Suppose A is a set of items and transaction T is said to contain A only if A ⊆T. Association rule is an implication like as A ⇒B i n which A, B ⊂ I and A ∩ B=Ø [6].Definition of support: The support is the percentage of transactions that demonstrate the rule. An item set is called frequent if its support is equal or greater than an agreed upon minimal value the support threshold. [8].Definition of Confidence: Every association rule has a support and a confidence.

An association rule is of the form: X => Y.

X => Y: if someone buys X, he also buys Y.

The confidence is the conditional probability that, given X present in a transition, Y will also be present. Confidence measure, by definition:

Confidence(X=>Y) = support(X, Y)/ support(X)

The aim of association rule is to find all association problems having support and confidence not less than given threshold value. For the given support i.e. minsupp, if the item set of D's support is not less than minsupp, then it can say that D is the frequent item set.

The most common approach to find association rules is to break up the problem into 2 parts

1. Find Large Itemsets
2. Generate rule from the frequent Itemsets

A Large (Frequent) Itemset is an Itemset whose number of occurance is above the threshold (s).

III. PROPOSED WORK (ASSOCIATION RULE MINING AND SECURITY ALGORITHMS)

A. Association Rule Mining using Apriori

One of the first algorithms to evolve for frequent itemset and Association rule mining was Apriori. Apriori is iterative approach that uses level wise search. In each level it uses k frequent item sets to explore k+1 frequent item sets. Two major steps of the Apriori algorithm are the join and prune steps.

The join step is used to construct new candidate sets. A candidate itemset is basically an itemset that could either be frequent or infrequent with respect to the support threshold. Higher level candidate itemsets (Ci) are generated by joining previous level frequent itemsets are Li-1 with itself. The prune step helps in filtering out candidate item-sets whose subsets (prior level) are not frequent. This is based on the anti-monotonic property as a result of which every subset of a frequent item set is also frequent. Thus a candidate item set which is composed of one or more infrequent item sets of a prior level is filtered(pruned) from the process of frequent itemset and association mining.

Apriori Algorithm

Input D, a database of transactions Min_sup, the minimum threshold support

Output Lk Maximal frequent itemsets in D Ck Set of Candidate k-itemsets.

Method:

1. L1 =Frequent items of length 1.
2. For(k=1;Lk!=ϕ;k++) do.
3. Ck+1=candidates generated from Lk.

4. For each transaction t in database D do.
5. Increment the count of all candidates in Ck+1 that are contained in t.
6. Lk+1 =candidates in Ck+1 with minimum support
7. end do
8. Return the set Lk as the set of all possible frequent itemsets

The main notation for association rule mining that is used in Apriori algorithm is the following. 1) A k – itemset is a set of k items. 2) The set Ck is a set of candidate k-itemsets that are potentially frequent. 3) The set Lk is a subset of Ck and is the set of k-itemsets that are frequent. 4) First we find Ck by joining Lk-1 with itself. 5) Then we apply pruning on Ck to eliminate infrequent itemsets i.e. the candidate item sets whose support is less than Minsupp is pruned.

B. DSRRC Algorithm

For this algorithm, to hide an association rule like X⇒Y, we decrease its confidence (|X∪Y|/|X|)to smaller than specified minimum confidence threshold (MCT). We decrease the support of Y (R.H.S. of the rule) in the most sensitive transactions. Therefore algorithm reduces the confidence faster than reducing the support of X. To decrease support count of an item, we delete one item from selected transaction by changing from 1 to 0.

The following shows proposed framework of DSRRC Algorithm.

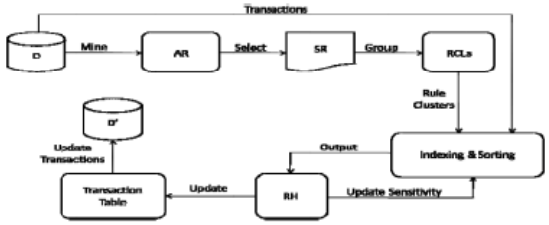


Fig 3.1 Framework for DSRRC Algorithm

The proposed framework of DSRRC algorithm is shown in Fig. 3.1. Initially association rules (AR) are mined from the source database D by using association rule mining algorithms e.g. Apriori algorithm in. Then sensitive rules (SR) are specified from mined rules. Selected rules are clustered based on common R.H.S. item of the rules. Rule-clusters are denoted as RCLs. Then for each Rule-cluster sensitive transactions are indexed. Sensitivity of each item (and each rule) in each Rule-cluster is calculated. Rule-Clusters are sorted in decreasing order of their sensitivity and sensitive transactions supporting first rule-cluster decreasing order of their sensitivity are sorted in. After sorting process, rule hiding (RH) process hides all the sensitive rules in sorted transactions for each cluster by using strategy mentioned in this section and updates the sensitivity of sensitive transactions in other cluster. Hiding process starts from highest sensitive transaction and continues until all the sensitive rules in all clusters are not hidden.

DSRRC Algorithm

INPUT: Source database D, Minimum Confidence Threshold (MCT), Minimum support threshold (MST).

OUTPUT: The sanitized database D'.

1. Begin
2. Generate association rules.
3. Select the sensitive rule set RH with single antecedent and consequent e.g. x→y.
4. Clustering-based on common item in R.H.S. of the selected rules
5. Find sensitivity of each item in each cluster.
6. Find the sensitivity of each rule in each cluster.
7. Find the sensitivity of each cluster
8. Index the sensitive transactions for each cluster.
9. Sort generated clusters in decreasing order of their sensitivity.
10. For the first cluster, sort selected transactions in decreasing order of their sensitivity
11. For each cluster c□ C
12. {
13. While(all the sensitive rules □ c are not hidden)
14. {
15. Take first transaction for cluster c.
16. Delete common R.H.S. item from the transaction.
17. Update the sensitivity of deleted item for modified transaction in other cluster and sort it.
18. For i = 1 to no. of rules Rh→c
19. {
20. Update support and confidence of the rule r→c.
21. If(support of r < MST or confidence of r < MCT)
22. Remove rule r from Rh
23. }
24. Take next transaction.
25. }
26. End while

27. }
28. End for
29. Update the modified transactions in D.
30. End

IV. IMPLEMENTATION DETAILS

In this section, we analyze the performance of our new approach for mining frequent item sets with privacy and compared with traditional algorithms. The Algorithms were implemented in java language. Swing framework is used for designing GUI. We have placed transactional data records in data sets. The SQL Server 2000 data base is used for managing the performance results.

V. EXPERIMENTAL RESULTS

In order to evaluate the performance of our proposed algorithm, we have conducted experiments on a PC (CPU: Intel(R) Core2Duo, 3.16GHz) with 4GByte of main memory running Windows XP. We used 3 data sets. One contains 72 samples (transactions), second one 39 samples and third one 120 samples.

The following shows the results of Apriori Algorithm for generating frequent item sets.

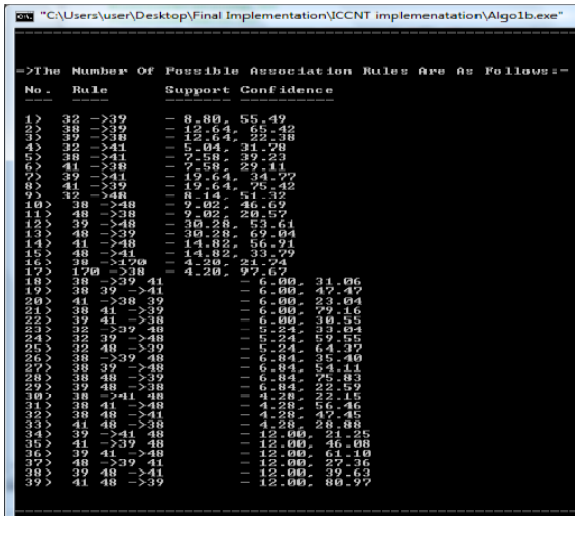


Fig 5.1 Association rules satisfying given MST and MCT in 5k transactions

The following shows sensitive rules generated by DSRRC Algorithm.

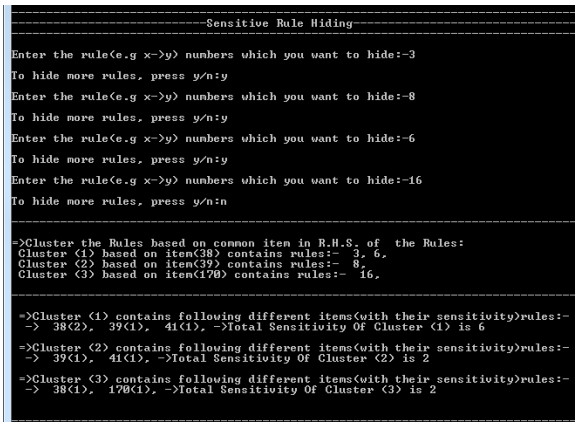


Fig 5.2 Rule clusters generated by DSRRC algorithm

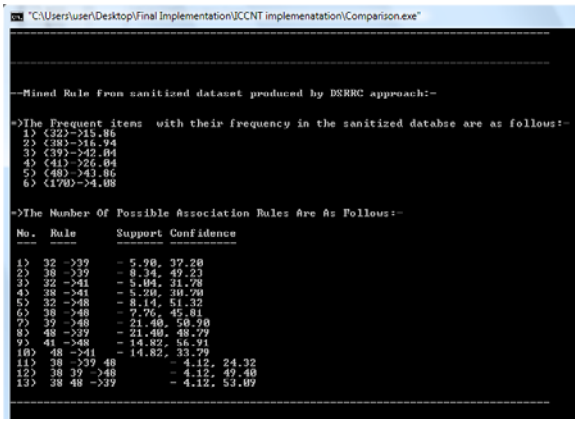


Fig 5.3 Rules generated from sanitized database produced by DSRRC algorithm

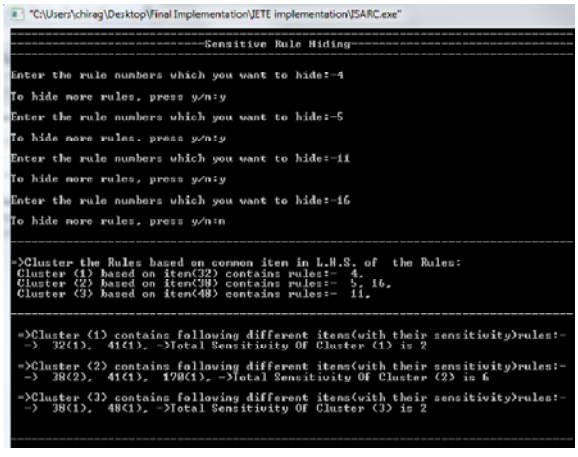


Fig 5.4 Rule clusters generated by ISARC algorithm

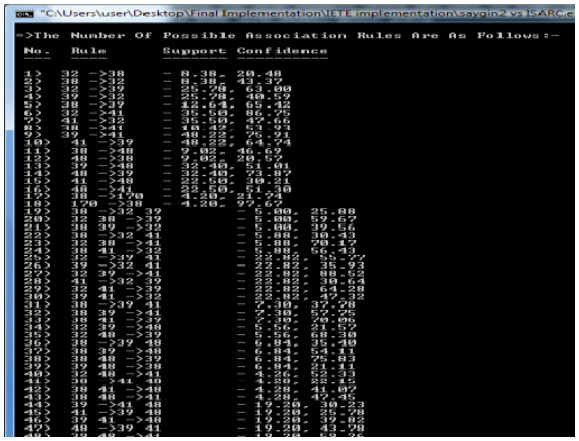


Fig 5.5 Rules mined from sanitized database produced by ISARC algorithm

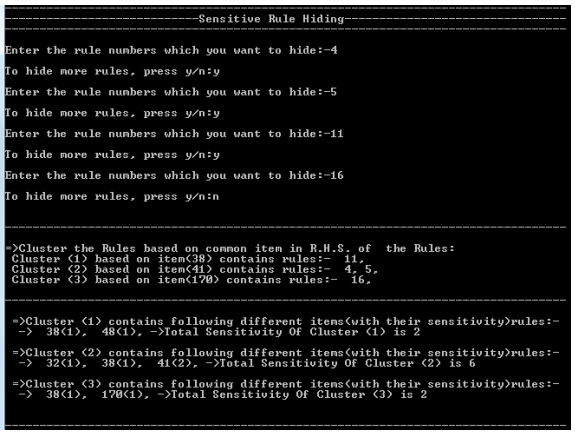


Fig 5.6 Rule clusters generated by DSCRC algorithm

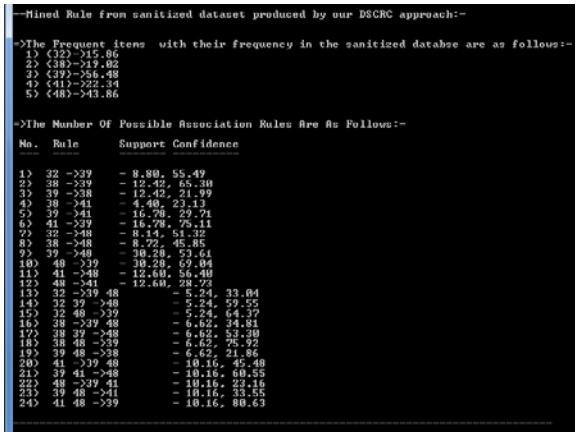


Fig 5.7 Rules mined from sanitized database produced by DSCRC algorithm

Algorithm	MCT	5K Transactions				
		(%) HF	*(%) MC*	*(%) AP*	*(%) DISS*	*SEF*
ISARC	10	0.00	68.18	0.00	2.89	68.18
	15	0.00	58.54	0.00	2.55	58.54
	20	0.00	57.14	0.00	1.90	57.14
DSRRC	10	50.00	50.00	0.00	1.43	45.45
	15	50.00	41.16	0.00	1.15	36.59
	20	0.00	48.57	0.00	0.95	48.57

Table 5.1 Performance comparison between ISARC and DSRRC for 5k transactions.

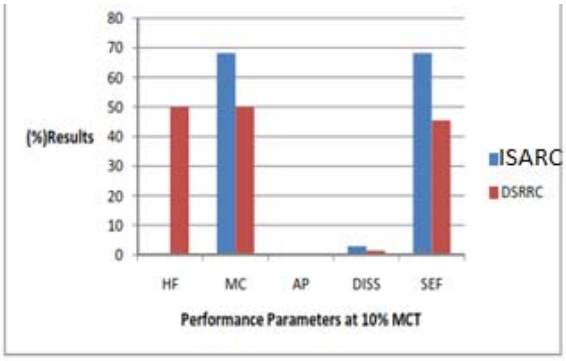


Fig 5.7 Performance comparison between ISARC and DSRRC algorithm at 10% MCT value for database of 5k transactions.

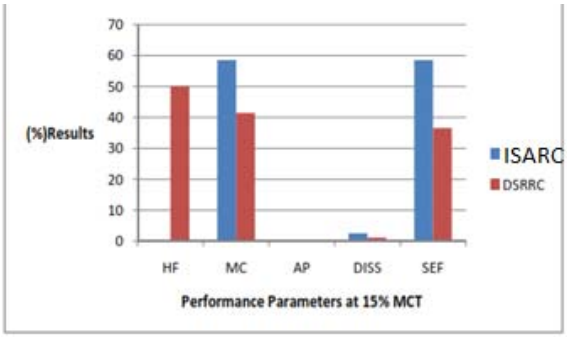


Fig 5.8 : Performance comparison between and DSRRC algorithm at 15% MCT value for database of 5k transactions.

The following shows the summary of performance of security algorithms.

	DSRRC	ISARC	DSCRC
Privacy	***	*	*
Knowledge Discovery	**	***	**
Data Quality	***	**	***
Side Effect Factor	*	**	*
Efficiency	***	***	***
Scalability	***	***	***

VI. CONCLUSION AND FUTURE WORK

Privacy preserving data mining is a new body of research focusing on the security and privacy implications originating from the applications of data mining algorithms to large public databases. In this paper, we have briefly surveyed existing approaches regarding knowledge hiding problem in context of association rule mining by their performance and limitations. We proposed heuristic algorithms named DSRRC, ISARC, DSCRC and Algorithm 1 for centralized data, which hide many sensitive association rules at a time while maintaining database quality. An example demonstrating for each proposed algorithm is discussed.. We analysed security and privacy of it against involving sites or adversary. It provides certain level of privacy and security under some other security assumption. The communication and computation cost are also reasonable for small databases which contain less number of items.

REFERENCES

[1] Chirag N.Modi, Udai Pratap Raoand Dhiren R.Patel "An Efficient Solution for Privacy Preserving Association Rule Mining".

[2] J. Han and M. Kamber, "Data mining concepts and techniques(2ed.)," Morgan Kaufmann, pp. 227-275,2006.

[3] M. Atallah, E. Bertino, A.K. Elmagarmid, brahim, V.S. Verykios. "Disclosure Limitation of Sensitive Rules". In Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX 1999).

[4] S.R.M. Oliveira, O.R. Zaïane, Y. Saygin. "Secure Association Rule Sharing". In Proceedings of the 8th Pacific-Asia Conference on PAKDD2004, Sydney, Australia, pp. 74–85, 2004.

[5] V.S. Verykios, A.K. Elmagarmid, E. Bertino, Y. Saygin, E. Dasseni. "Association Rule Hiding". IEEE Transactions on Knowledge and Data Engineering, Vol. 16, no. 4, pp. 434–447, 2004.

[6] S.L. Wang, A. Jafari. "Using Unknowns for Hiding Sensitive Predictive Association Rules". In Proceedings IEEE International Conference on Information Reuse and Integration (IRI 2005), pp.223–228, 2005.

❑❑❑

An Optimistic Approach to Spectral Kernel Learning Algorithm for Regression

Koushik Akkinapalli[1], Pradeep Aidam[2] & Muralidhar Adari[3]

[1]Tata Consultancy Services, Chennai, India, [2]WebPark.In (Info Edge Group), Hyderabad, India
[3]Tata Consultancy Services, Bangalore, India
E-mail : koushik501@gmail.com[1], pradeepaidam29@gmail.com[2], murali_adari@yahoo.com[3]

Abstract - Semi Supervised Kernel learning approach designs a kernel using label information of labeled data and unlabeled data and is also an important research area. In this paper, a schematic approach to spectral kernel learning for regression has been introduced to handle both labeled and unlabeled data by making use of kernel matrix. The semi supervised spectral kernel learning algorithm characterizes the *maximum margin algorithm* where a *non-linear function* is learned by linear learning machine in kernel induced feature space. The *Lagrangian Support Vector Machine* and *Gradient Descent Algorithm* has been introduced to train the kernel machines for regression in an optimized manner.

Keywords - *Regression, Support Vector Machines, Spectral Kernel Learning, Kernel Matrix.*

I. INTRODUCTION

Kernel learning is an important learning algorithm for semi supervised learning to build better classifiers which combines both labeled and unlabeled data.Here SVM(Support Vector Machines) has been introduced which is a part of Semi Supervised Learning. The Objective of SVM is to find an optimal function between linearly separable classes by maximizing the margin between classes into a high dimensional input space. Here Lagrangian Support Vector Machines(LSVM) has been proposed since it is more effective than SVM for improving accuracy for building classifiers.

A good kernel method should not only make use of training data but also must make use of testing data to evaluate the performance. Semi Supervised learning is an important technique for handling both labeled and unlabeled data.

A Graph based method is also one of the semi supervised learning strategy where a graph is defined as the nodes which are the examples in the data set and edges denotes the similarity measures of the data. A graph based method can be viewed as semi supervised kernel by tuning spectrum of graph laplacian on both labeled and unlabeled data. The spectra kernel learning method has been extended by learning kernel matrix with faster decay rate.

Since only labeled data are used for training the performance of kernel methods to build effective classifiers is not enough. So, some algorithms has to be merged into kernel learning algorithms. Here LSVM and kernel based algorithms has been designed to build effective classifiers. A new method has been proposed to learn a spectra of kernel matrix by making use of maximal margin algorithm.

A Kernel matrix is discussed in section 3 and SVM for regression is discussed in section 4 and approach to spectral kernel learning algorithm for regression is discussed in section 5.

II. KERNEL MATRIX

Given a labeled data set{(x1,y1),(x2,y2)….(xi,yi) } and unlabeled data set{(xi+1…xn) .Let yi denotes a class label and n denotes the size of data. Let k denotes a positive semidefinite kernel matrix that defines a kernel Hilbert space where Kij=k(xi,xj).

Supervised Kernel Matrix :

Here we consider the eigen decomposition of new kernel matrix which is given by

$$K = \sum \lambda_i v_i v_i^{\mathbb{T}}$$

Where vi are eigen vectors of K, λ are eigen values of K.

Unsupervised Kernel Matrix

A spectral transformation function r(.) that is non-negative and decreasing. The new Kernel matrix is

given as follows

$$\bar{K} = \sum_{i=1}^{n} r(\lambda_i) v_i v_i^{\top}$$

In this paper a new spectral transformation function $u_i = r(\lambda_i)$ has been developed to obtain new spectral coefficients of new kernel matrix. The new matrix can also be regarded as the linear combination of set of rank-one matrix

$$\bar{K}_i = v_i v_i^{\top}.$$

The new Kernel matrix can be rewritten as

$$\bar{K} = \sum_{i=1}^{n} \mu_i \bar{K}_i$$

III. SUPPORT VECTOR MACHINE FOR REGRESSION

SVM can be applied to regression problems by the introduction of loss function. Loss function is modified to include a distance measure. The figure shows loss function.

The loss function corresponds to least square error criterion. **Linear Regression**

Consider the problem of approximating set of data

$$D = \{(x^1, y^1), \ldots, (x^l, y^l)\},$$

With a linear function

$F(x) = (\omega, x) + b$

The optimum regression function is given by the minimum of the functional

$$\Phi(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_i (\xi_i^- + \xi_i^+),$$

Where C is a prespecified value and ξ^-, ξ^+ are slack variables representing upper and lower constraints on the output of system

Quadratic Loss Function Using this equation

$$L_{quad}(f(\mathbf{x}) - y) = (f(\mathbf{x}) - y)^2.$$

The solution is given by

$$\max_{\alpha,\alpha^*} W(\alpha, \alpha^*) = \max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle + \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)y_i - \frac{1}{2C}\sum_{i=1}^{l}(\alpha_i^2 + (\alpha_i^*)^2).$$

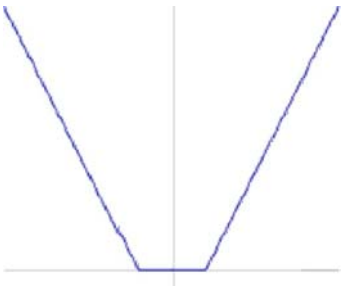
The corresponding optimization can be simplified by exploiting Karush-Kuhn Tucker(KKT) condition

$$\bar{\alpha}_i \bar{\alpha}_i^* = 0, \quad i = 1, \ldots, l.$$

Where the both variables in above equation are Lagrange multipliers and these implies $\beta_i^* = |\beta_i|$. and the resulting solution is given by

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\beta_i\beta_j\langle x_i, x_j\rangle - \sum_{i=1}^{l}\beta_i y_i + \frac{1}{2C}\sum_{i=1}^{l}\beta_i^2$$

ϵ-insensitive Loss Function



The insensitive loss function from this figure is given by

$$L_\epsilon(y) = \begin{cases} 0 & \text{for} \quad |f(\mathbf{x}) - y| < \varepsilon \\ |f(\mathbf{x}) - y| - \varepsilon & \text{otherwise} \end{cases}.$$

The optimal solution is given from the above equation

$$\max_{\alpha,\alpha^*} W(\alpha, \alpha^*) = \max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i, x_j) + \sum_{i=1}^{l}\alpha_i(y_i - \epsilon) - \alpha_i^*(y_i + \epsilon)$$

with constraints

$$0 \le \alpha_i, \alpha_i^* \le C, \quad i = 1, \ldots, l$$

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0.$$

Using the above equations determines Langrange multipliers the regression function is given by Linear function where

$$\bar{w} = \sum_{i=1}^{s}(\alpha_i - \alpha_i^*)x_i$$

$$\bar{b} = -\frac{1}{2}\langle \bar{w}, (x_r + x_s)\rangle.$$

Therefore support vectors are points where exactly one of the lagrange multipliers is greater than zero. when

$\epsilon = 0$, we will get the L1 loss function and corresponding optimal solution is obtained in a simplified way which is given by

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\beta_i\beta_j\langle x_i, x_j\rangle - \sum_{i=1}^{l}\beta_i y_i$$

With constraints

$$-C \le \beta_i \le C, \quad i = 1, \ldots, l$$
$$\sum_{i=1}^{l}\beta_i = 0,$$

and regression function is given by given by

$$\bar{w} = \sum_{i=1}^{l}\beta_i x_i$$
$$\bar{b} = -\frac{1}{2}\langle \bar{w}, (x_r + x_s)\rangle.$$

IV. AN APPROACH TO SPECTRAL KERNEL LEARNING ALGORITHM FOR REGRESSION

The SVM goal is to find the optimal margin classifier f(x) by making use of maximal margin algorithm. The standard SVM algorithm for regression is given by

$$\min \frac{1}{2}\|w\|^2 + C\sum_t(\xi_+^t + \xi_-^t)$$

subject to

$$\begin{aligned} r^t - (w^T x + w_0) &\le \epsilon + \xi_+^t \\ (w^T x + w_0) - r^t &\le \epsilon + \xi_-^t \\ \xi_+^t, \xi_-^t &\ge 0 \end{aligned}$$

where ε is a slack variable and there are two variables i.e., positive and negative and C>0 is the penalty parameter of training error.

In this paper Langrange SVM is used which is used is more effective than standard svm and LSVM is used for spectral kernel learning.

The LSVM for regression is given by

$$\frac{1}{2}\|w\|^2 + C\sum_t(\xi_+^t + \xi_-^t)$$
$$-\sum_t \alpha_+^t\left[\epsilon + \xi_+^t - r^t + (w^T x + w_0)\right]$$
$$-\sum_t \alpha_-^t\left[\epsilon + \xi_-^t + (w^T x + w_0) - r^t\right]$$
$$-\sum(\mu_+^t\xi_+^t + \mu_-^t\xi_-^t)$$

Correspondingly dual solution of above is given by

$$-\frac{1}{2}\sum_t\sum_s(\alpha_+^t - \alpha_-^t)(\alpha_+^s - \alpha_-^s)(x^t)^T x^s$$
$$-\epsilon\sum_t(\alpha_+^t + \alpha_-^t) - \sum_t r^t(\alpha_+^t - \alpha_-^t)$$

maxα+α-

subject to

$$0 \le \alpha_+^t \le C, 0 \le \alpha_-^t \le C, \sum_t(\alpha_+^t - \alpha_-^t) = 0$$

The spectral learning problem is turned into

$$-\frac{1}{2}\sum_t\sum_s(\alpha_+^t - \alpha_-^t)(\alpha_+^s - \alpha_-^s)(x^t)^T x^s$$
$$-\epsilon\sum_t(\alpha_+^t + \alpha_-^t) - \sum_t r^t(\alpha_+^t - \alpha_-^t)$$

Min μMaxα+α-

Here $(x^t)^T x^s$ ıas been replace with kernel $K(x^t, x^s)$,

Subject to

$$0 \le \alpha_+^t \le C, 0 \le \alpha_-^t \le C, \sum_t(\alpha_+^t - \alpha_-^t) = 0$$
$$\bar{K} = \sum_{i=1}^{a}\mu_i\bar{K}_i,$$
$$trace(\bar{K}) = c,$$
$$\mu_i \ge 0,$$
$$\mu_i \ge P\mu_{i+1}, i = 1 \ldots d-1.$$

The problem of above is considered in the following form:

$$\min_{\mu} \; G(\mu)$$
$$\text{subject to} \quad \sum_{i=1}^{d} \mu_i = c,$$
$$\mu_i \geq 0,$$
$$\mu_i \geq P\mu_{i+1}, i = 1 \ldots d-1$$

Where

$$G(\mu)=\max_{\alpha+} \; -\frac{1}{2}\sum_t \sum_s (\alpha_+^t - \alpha_-^t)(\alpha_+^s - \alpha_-^s)(\boldsymbol{x}^t)^T \boldsymbol{x}^s - \epsilon \sum_t (\alpha_+^t + \alpha_-^t) - \sum_t r^t(\alpha_+^t - \alpha_-^t)$$

$$0 \leq \alpha_+^t \leq C, 0 \leq \alpha_-^t \leq C, \sum_t (\alpha_+^t - \alpha_-^t) = 0$$

subject to

The gradient of G(μ) is computed as

$$G(\mu)/\mu m = -\frac{1}{2}\sum \sum (\alpha_+^t - \alpha_-^t)(\alpha_+^s - \alpha_-^s) \qquad s$$
$$k_m(x, x)$$

V. CONCLUSION AND FUTURE WORK

In this paper a new kernel matrix has been designed.The semi supervised kernel learning method is to modify the spectra of original kernel by making use of maximal margin algorithm.The LSVM and spectral kernel learning for regression can be combined together for learning spectra of new kernel matrix in an optimized manner.

In Future Work effective algorithms such as distance metric learning for regression is going to be designed.

VI. ACKNOWLEDGEMENT

We express our deep sense of gratitude to Gayatri Vidya Parishad College Of Engineering (Autonomous), Visakhapatnam for providing us an opportunity to fullfill our most cherished desire of reaching out goal. We express our sincere thanks to Prof. D.Murali Krishna, Prof. D.V.V.Sharma, P.Krishna Subba Rao Head of the Department, N.Sharmili Assoc. Prof., N.Vishali Asst. Prof., N.V.L.P. Raju Asst. Prof., Department of Compurer Science and Engineering for the imponderable guidelines. We are thankful to the staff of our department and every body who has been directly or indirectly involved for their encouragement in completing our work successfully.

REFERENCES

[1] Semi-Supervised Learning Literature Survey by Xiaojin Zhu ,Computer Sciences TR 1530 on July 19,2008.

[2] Introduction to Machine Learning by Nils J.Nilsson on November 3,1988I.

[3] Support Vector Machines and Kernel Machines by Nello Cristianini on 2001

[4] Introduction to Machine Learning by Ethem Alpaydin on 2004

[5] Supervised Machine Learning:A review of classification techniques by S.B.Kotsiantis onJuly 16,2007.

[6] Semi Supervised Support Vector Machines by P.Bennet on 1998

[7] Maximum Margin Based Semi Supervised Spectral Kernel Learning by Zenglin Xu, Jianke Zhu, Michael R. Lyu, Fellow, IEEE, and Irwin King, Member, IEEE on 2007

[8] Support Vector Machines by Abdrew W. Moore, Professor, School of Computer Science Camegie Mellon University

❑❑❑

Rule Based Intelligent Mining System - Jiffy Analyser

Atesh Kumar[1], Unique Gangwar[2], Ritika Sethi[3] & Ashish Ranjan[4]

[1,2&3]Department of Computer Science, Jaypee Institute of Information Technology, Noida, India
[4]Department of Computer Science, National Institute of Technology, Patna, India
E-mail : ateshsingh@yahoo.com[1], unique.gangwar@gmail.com[2],
ritikasethi03@gmail.com[3], ashishranjan28@gmail.com[4]

Abstract - With the rapid growth of e-commerce applications, Internet shopping is becoming part of our daily lives. Traditional Web-based product searching are based on keywords which seems insufficient and inefficient in the 'sea' of information. In this paper, we propose an innovative intelligent multi-agent based environment, namely (RBIMS) – Rule Based Intelligent Mining System - to provide an integrated and intelligent agent-based platform in the e-commerce environment. In addition to contemporary agent development platforms, which focus on the autonomy and mobility of the multi-agents, RBIMS provides an intelligent layer (known as the 'conscious layer') to implement various Artificial Intelligence functionalities in order to produce 'smart' agents.

Keywords - *Algorithm ,Data mining,NLP,Arificial Intelligence.*

I. INTRODUCTION

With the explosive growth of data available on the Internet, personalization of this information space becomes a necessity. An important component of web personalization is the automatic knowledge extraction. However, analysis of large web is a complex task not fully addressed by existing web access analyzers. Using a social networking website, we have attempted to apply data mining techniques (association rules and clustering) to analyze the public opinion on the chosen website. The presented software recognizes several reading patterns and discusses approaches for mining the data of various social networking websites

The Internet and the World Wide Web are the most important technology present today. It is being used to enhance the standard of living and providing all together different approach to the lifestyles in the big manner. Today's applications can be written to communicate among the world's hundreds of millions of the computers.

This project presents a new way of analyzing the reviews or comments of different people towards a particular topic. From the concrete point of view to analyze peoples view towards a particular topic organization have to maintain records. By this project we can reduce the burden of creating the long registers and maintain them.

RBIS holds the main objective of securing the mined data by analyzing within the hidden layer and representing the goodwill or the downfall of any entity stored in the knowledgebase. It provides the parameter on the bases of which the analyzer calculates the values and finally displays the result.

It performs the following task:

- Collection of reviews from various social networking websites
- Tagging of reviews into is Parts Of Speech (POS) format.
- Selection of the parameter as the input to analysis phases.
- Searching on the basis of parameters defined by the user.
- Analysis of the search.

In the evolution course of computing, a field called artificial intelligence was developed in the 70s and 80s with an aim to make computers reason like human beings. Rule-based programming paradigm emerged at that time as ways to implement systems that appear to think and reason like human beings. Examples of rule-based systems are expert systems that have the knowledge of a doctor or a tax advisor and can answer

complex questions people would normally ask those professionals. The idea of rule-based programming is to represent a domain expert's knowledge in a form called rules. That's why it's called rule-based. Besides rules, another important ingredient in a rule-based system is facts. Here's an example. Say John is a weather reporter who gives advice to people on TV channels based on weather conditions. Here's John's knowledge about weather: A rule-based system consists mainly of three things: facts, rules and an engine that acts on them. Rules represent knowledge and facts represent data. A rule-based system solves problems by applying rules on facts (i.e. matching facts with rules' if clauses). A rule consists of two parts: conditions (if clauses) and actions. The action part of a rule might assert new facts that fire other rules.

Among the most popular rule engines, RBIMS is probably of the most interest to Java developers. It is the reference implementation of JSR 094 Java Rule Engine API and it has plug-ins to support development of rule systems in Eclipse.

Therefore, the rule code we will see in this article will be in RBIMS's syntax. Before we leave this section, it's helpful to introduce RBIMS and take a glimpse of its programming syntax.

RBIMS is software that interprets rules and facts expressed in its programming language. Just as Java is a language for expressing objects and Java compiler is software that interprets Java code, RBIMS has a language for expressing rules and facts and a compiler to interpret the code. Here's what the rules and facts in our weather example look like in RBIMS code.

Since the web consists of very huge amount of text data and is not possible to dumb all the text in our simple pc because of the available memory boundation. The solution of this problem is to use a web crawler and then providing the facility to the user that from which URL he wants to perform the data mining. Whenever the user enters the URL our web crawler attached with our software retrieve the whole text from that website and dumb it into a form of text file .after applying filtering on the text file we perform pos tagging with the help of pos tagger. With the help of pos tagger we are able to distinguish the sentences and we are able to judge there positive and negative polarity and generate the result according to this mining.

II. METHODOLOGY

Rule Based Intelligent Mining System (RBIMS) is the stand-alone application which overcomes the problem of customer, of checking out the market status of the product on the basis of user reviews. It is the combination of NATURAL LANGUAGE PROCESSING (NLP) and INTELLIGENCE (learning factor).

A. Task Performed by RBIMS

a) Collection of Reviews :

In a social networking web application we have millions of reviews we have focuses on some of the websites for the reviews collection like so www.twitter.co.in, www.bikeadvice.in, www.Mouthshut.com, www.Comparebike.com, www.Autonagar.com .

Our software has also provided the user review collection facility in which the user can submit his/her reviews on his willingness about a particular product.

b) Tagging of Reviews :

It refers to converting of simple review text to its Parts Of Speech (POS) form with the help of Stanford university POS tagger.

c) Selection of Parameters :

On the basis of user's selection of the product, the respective parameter for its analysis is to be selected. Suppose that the user select the bike name "Discover" and the respective parameters are (saying) Speed, Look, Average, and Price

d) Searching on the Basis of Parameters:

With respect to selected parameter the search is made in the related knowledgebase. In order to check in the positive or negative dataset, on the basis of Adjective-noun rule respectively. As in the above example, the four parameters mentioned, the search is made in the positive dataset and the negative dataset.

e) Analysis the Search:

On the basis of the search in the datasets respective result is been analyzed i.e., how many persons reviews is in the positive respect and how many is in the negative respect.

B. Intelligent Factor:

Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages and with the very good use of NLP the RBIMS is holding the intelligent feature also by building the learning ability of RBIMS.

RBIMS is based on the supervised learning which is a machine learning technique for deducing a function from training data. The training data consist of pairs of input objects (typically vectors), and desired outputs.

The output of the function can be a continuous value (called regression), or can predict a class label of the input object (called classification). The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples (i.e. pairs of input and target output). To achieve this, the learner has to generalize from the presented data to unseen situations in a "reasonable" form.

In respect to our software (RBIMS) NLP performs the operation of converting the simple English sentence in its Parts Of Speech (POS) format.

We will see that with the help of following examples,

Discover has a lean look.

Pulsar has low fuel efficiency.

These two sentences are than tagged to their POS format;

Discover_ noun has _helping verb an article lean_ adjective look_ proper noun.

Pulsar_noun has_ helping verb low_ adjective fuel_proper noun efficiency _verb.

C. Data Analysis

a) Data Dictioinary

In RBIMS the knowledgebase that is used is divided into two datasets as Fig 1 shows

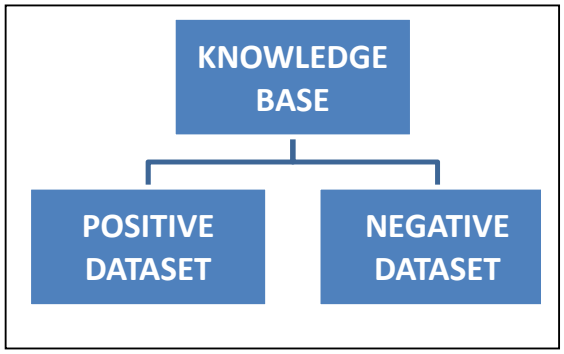


Fig. 1: Represents the Knowledgebase content

- *Positive Dataset*:

It is the collection of all the positive words for the product 'bikes' positive means the words that describes the product in its goodwill.

- *Negative Dataset*:

It is the collection of all the negative words for the product 'bikes' negative means the words that describes the product in its downfall.

Positive Dataset
Oozes out
Gorgeous
Best
Good
Nice
Perfect
Compact
Fantastic
Agile
Beautiful

Negative Dataset
Worst
Bad
Not reliable
Ugly
Disappointment
Shabby
Disbelieve
Low
Poor
Disgrace

The entire searching and analysis is performed in these two dataset of the knowledgebase.

D. Key Logic

The main logic used in our software (RBIMS) is NLP rule i.e. ADJECTIVE-NOUN rule.

An adjective is a word that modifies a noun by specifying an attribute of the noun. Examples include adjectives of color, like red, size or shape, like round or large, along with thousands of less classifiable adjectives like willing, onerous, etc.

In grammar rules, we use the symbol ADJ for the pre-terminal category of adjectives. Adjectives are also used as the complements of sentences with verbs like "be" and "seem" - "He is happy", "He seems drunk".

ADJ is a lexical grammatical category. A noun is a word describing a (real or abstract) object. Contrast verb, adjective, adverb, preposition, conjunction, and interjection. Noun is often abbreviated to N. N is a lexical grammatical category.

In respect to POS tagged text the respective adjectives are searched for the particular noun where the parameter given by the user is searched in positive dataset and then the negative dataset for the respective adjectives of the particular bike.

Like in statement,

Discover_noun has_helping verb good _adjective look_propernoun.

In this the main noun is the Discover in this respect the proper noun look is searched and with respect to it the adjective is searched ie good.

Discover ⟹ look ⟹ good

This means Discover bike have a positive review on the bases of 'look' parameter.

III. ARCHITECTURE

The architecture of RBIMS includes :

i. *Internet:* As Internet is the vast pool of data from which we have crawled reviews from limited number of websites as the default knowledge base of our software (RBIMS).

ii. *Revised Dataset:* It is the collection of crawled reviews from the internet in a simple text format. While the, USER REVIEWS is the reviews given by the user randomly or through our software (RBIMS) which is again in the simple text format.

iii. *Pos Tagger:* It is the downloaded software which converts the simple text data to its POS tagged form and generates the respective dataset. it tags the entire review dataset containing the user reviews given by using our software.

iv. *Pos Tagged Reviews Dataset:* It is the dataset been formed after the tagging process.

v. *User Interface:* It is the front look of our software through which the user can give the respective four parameters as the input to the software.

vi. *Knowledge Base:* It is the collection of all the respective vocabulary used to define the different parameters of different bikes.

- Positive Dataset holds the higher grade vocabularies which speaks good for the respective product.
- Negative Dataset holds the lower grade vocabularies which speaks bad for the respective product.

If any new word is found while reviewing the knowledge base then a message will be displayed to user ' New Word Found' then the user has to label the word to the dictionary in positive or negative sense.

vii. *Inference Engine:* It is the main machine of the software which contains the entire functioning of the software which carry out the searching processes of the respective parameters in the knowledge base and then give back the result in the number of public opinion in favor of the product or against the product.

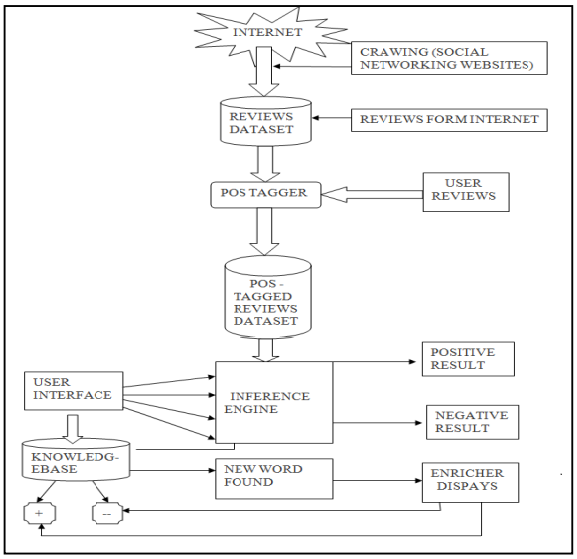


Fig. 2: Architectural of the RBMIS showing the whole Process included in it.

IV. ALGORITHM.

4.*1 General Algorithm:*

a. Crawling various social networking websites for the purpose of collecting reviews about different bikes.

b. Then converting this simple file of text to its POS format with the help of POS tagger.

c. Then the tagged file is given as the input to the inference engine to perform the hidden layer processing.

- For the particular bike four parameter which is given by the user is been selected.
- Searched the tagged reviews on that bike on the basis of the parameter mentioned.
- Then the respective bikes result is checked in the negative and the positive dataset.

If then

- The result is found in the respective positive and negative it well displays the number of public reviews in its positive and negative respects.

Else

- It will display the dialog box to the user and ask them to label the particular word in the positive or negative dataset.

d. Then the final output is displayed that is it shows the number of public reviews in respect to goodwill of product and number of reviews in against of the product.

e. Finally exit.

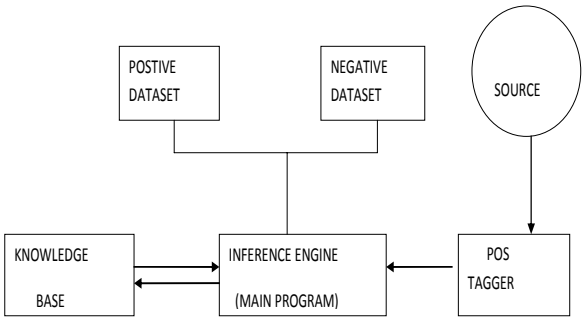


Fig. 3: Algorithm process

4.2 Learning Algorithm:

RBIMS uses the supervised learning algorithm in which there is the respective guide to carry out the learning of the software which is user in our software.

a. The four parameters given by the user are then fetched from knowledgebase.

b. Then the search of the respective parameter's adjectives is searched from the positive dataset then the negative dataset.

c. If found then successful.

d. Else the dialog box appears with instruct the user to label the new word on its bases to negative dataset or the positive dataset.

This is how our software follows the supervised learning.

V. RESULT

The output result of this screen will be shown like this :

Fig. 4: Represents the learning screen and first screen

VI. LIMITATION

a) Not to deal with the free text.

Now a day we are using slangs in conversation. and our software do not works on the text which we use in short messaging service(SMS).

For e.g.: Fine = f9.

Your = ur.

Great = gr8.

b) Reviews ever changing.

The social networking web sites are dynamic in nature e.g. twitters is one of the most dynamic sites of the world with data changing every second. On which reviews are continuously change at each instant of the time.

c) Single rule:

There are a number of rules applicable in NLP and we have designed a system works only for a single rule of NLP. Due to static nature we are not providing the facility of adding new rules.

VII. CONCLUSION AND FUTURE WORK

The software has developed in the way to overcome the problem regarding "online product information". As in this modern era were time is very precious to all, there meeting place is through online reviews even e-commencing is taking a deep root among the youngsters for shopping purposes. But they are not getting any such platform were they can get the assembled information about any of the product even there is problem regarding the comparison of different brand products

So by keeping today's scenario in mind we have facilitated the public with a platform where customer can use the public opinion for analyzing on any of the user defined parameters. Through this software customer can analyze any of the brands of with 78% accuracy.

As here is 'Jiffy – Analyzer—(a rule based intelligent mining system)' the solution to all the problems regarding the 'online product information'. In the future we will add some more rules so that it will be more dynamic and more accurate.

REFERENCES

[1] Robert M.Losee ,"Natural Language Processing In Support of Decision-Making: Phrases and Part-of-Speech Tagging", *Information Processing and Management*, 37 (6), pp. 769-787, November 2001

[2] Kristina Toutanova and Christopher D. Manning .." Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger". In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70..

[3] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer.2003."Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network". In Proceedings of HLT-NAACL 2003, pp. 252-259

[4] Nadeau, D., and Turney, P.D. (2005), A supervised learning approach to acronym identification, Proceedings of the Eighteenth Canadian Conference on Artificial Intelligence (AI-05), Victoria, Canada, pp. 319-329. (NRC #48121).

[5] Turney, P.D. (2003), Coherent keyphrase extraction via Web mining, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico, pp. 434-439. (NRC #46496).

[6] Susan Dumais,John Platt,David Heckman,Mehran Sahami Inductive learning algorithms and representations for text categorization CIKM '98 Proceedings of the seventh international conference on Information and knowledge management ACM New York, NY, USA ©1998 ISBN:1-58113-061-9 doi>10.1145/288627.28865

❑❑❑

Threat Evaluation Modelling for Dynamic Targets Using Fuzzy Logic Approach

Sushil Kumar & Arati M. Dixit

Department of Applied Mathematics, DIAT Deemed University, Pune, India
E-mail : sushil0402k5@gmail.com, aratidixit@diat.ac.in

Abstract **-** Threat evaluation is a critical component of the system protecting the defended assets against the hostile targets like aircrafts, missiles, helicopters etc. The degree of threat is evaluated for all possible hostile targets on basis of heterogeneous parameter values extracted from various sensors, to improve the situational awareness and decision making. Taking into consideration the amount of uncertainty involved in the process of threat evaluation for dynamic targets, the fuzzy logic turns out to be a good candidate to model this problem.

Keywords **-** *Threat Evaluation, Fuzzy Knowledge Based System, Membership Functions, Intent Assessment, Capability Assessment, and decision support system.*

I. INTRODUCTION

In order to support the security of any nation the places of significance are to be protected as defended assets. The various defended assets can be air bases, tourist places, bridges, camps, nuclear power plants, command post, harbors, radars, monuments, parliament's buildings, etc. In the war as well as peace keeping scenario it becomes critical to understand the possible enemy dynamic targets such as aircrafts (bomber, fighter, and transporter), missiles, helicopters, etc which can be manned or unmanned targets. The decision making is very critical with respect to available resources and time. The dynamic targets are those targets which are mobile and exhibit change in their characteristic behavior. Various factors are considered for a decision making augmented with human cognitive intelligence. An expert system built with help of fuzzy logic can play an important role in enhancing situation awareness and automated decision making.

The protection of defended assets is the prime objective of threat evaluation modeling of dynamic targets. An assumption is made that defending targets act as potential threats, but targets may be friend or enemy which is decided by IFF (Identification, friend or foe). The IFF is designed by command and control system. In this situation, prioritization of potential threats is very important according to threat level (Degree of threat) of detected enemy targets via multi-resources. Battle space and intelligent sensors help in target classification. Threat value quantifies the possibility of threat or danger imposed by a potential target. In this situation of possible multiple targets, it becomes critical to prioritize the degree of threat involved with them to decide which target is more dangerous via predicting the threat value. Threat value is directly proportional to the amount of danger a target produces towards the protected asset. The higher threat value implies more dangerous target. This analysis in turn will play a significant role in weapon allocation against suspicious targets.

A grid of sensors produces large amount of heterogeneous data which can be used to evaluate the degree of threat of a target. Thus threat evaluation is a high level information fusion process. At times the threat evaluation becomes challenging in the presence of multiple parameters and processes. There is some amount of uncertainty involved in these parameters depending on the nature of targets and assets involved. It is difficult to formulate mathematical model by using selected parameters as inputs to generate the threat value as an output. The fuzzy inference system turns out to be one of the most efficient methods for the threat evaluation of dynamic targets under uncertain condition.

II. THREAT EVALUATION

The threat is an expression of intension to inflict evil, injury, or damage [1]. The threat evaluation is significant component in target classification process. Small errors or mistakes in threat evaluation and target classification can result in huge damage of life and property. A misclassification of a non-threat as a threat

resulted in tragic consequences by the US Navy cruiser USS_Vincennes:

The US Navy cruisers USS Vincennes shoot down an Iranian commercial airliner in the late 1980s, assuming it to be an Iranian air force F-14 aircraft. This action resulted in destruction of the passenger aircraft, killing to 290 civilians traveling in the plane.

Similar incidents have taken place in the past at different instances resulting in embarrassing and catastrophic consequences. Threat evaluation helps in case of weapon assignment, and intelligence sensor support system. It is very important factor to analyze the behavior of enemy tactics as well as our surveillance. Disastrous situation in terms of loss of life and the valuable assets occur due to wrong evaluation of threat value. In this case we will suffer more as damages so it is important to evaluate more accurately.

Threat evaluation is a process based on defending targets to defended asset; here an assumption is to protect one asset against several defending targets but consideration of more number of assets will give realistic feel towards threat evaluation. It is a high level information fusion technique that belongs to third level data fusion model in Joint Directors of Laboratories (JDL) as seen in Figure 1.

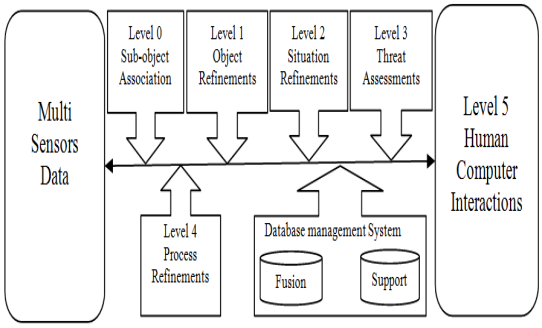


Fig. 1 : The JDL Model

The JDL is a conceptual information fusion Model, which describes the processes, functions and specific techniques used for information fusion. Data fusion is the process of combining data or information to estimate or predict entity states [2]. It describes how data from different sources is transformed to information. This information used by decision makers which improves the situational awareness. In this model, data utilized is obtained from different sources like radars, sensors and databases. After estimation of information, aggregation and improvement can be done to extract right information for the decision makers. The JDL model comprises different levels [2]:

A. Level 0: Sub-Object Data Assessment:

This level focuses on heterogeneous data collection. Assessment and prediction of data observable states on the basis of data association and characterization is done in this level. At this level, data is accessed from different sources, which may be localized or distributed. The main task of this level is to pre-process data by correcting biases and standardizing the input before the data from variety of sources is fused.

B. Level 1: Object assessment/ refinement:

The data collected in level zero is processed in this level to extract useful information. Assessment and prediction of entity states on the basis of observation-to-track association for continuous state estimation (e.g. kinematics) and discrete state estimation (e.g. Target type and ID) is done in this level.

C. Level 2: Situation Assessment:

The information extracted in level one is utilized to study the impact on current situation. Assessment and prediction of relations between the entities and relationship with the surrounding is focused in this level. This includes force structure, cross force relations, communications, perceptual influences, physical context, etc.

D. Level 3: Impact Assessment:

The situation information generated in level two is studied with respect to the role of possible contributors on the situation. Assessment and prediction of effects on situation of planned or estimated/predicted actions by the participants; to include interactions between action plans of multiple players (e.g. assessing susceptibilities and vulnerabilities to estimated/predicted threat actions given one's own planned actions) is the main focus of this level.

E. Level 4: Process Refinement (an element of Resource Management):

This level focuses on the optimization of over all information fusion process.

The decision making in limited time is very important in most of the peace-keeping and war scenarios. The Observe, Orient, Decide, and Act (OODA) loop is a concept originally applied to the combat operations process, often at the strategic level in military operations. The OODA loop is considered to be one of the most effective decisions making model in defence and security, often applied to understand commercial operations and learning processes today. The OODA loop can be seen in Figure 2. The whole idea of this loop is make faster decision for a quicker action with respect to the enemy. The battle space dominance can be achieved by:

1. *Observing* the entities and actions of the enemy,
2. *Orienting* resources to tackle current situation,
3. *Deciding* actions on basis of current situation,
4. *Acting* seamlessly on the decision made.

When the enemy aircraft comes into radar contact, more direct information about the speed, size, and maneuverability of the enemy target becomes available.

To determine which of several threats that represent the highest danger is of great importance, since errors such as prioritizing a lesser threat as a greater threat can result in engaging the wrong target, which often will have severe consequences [1].

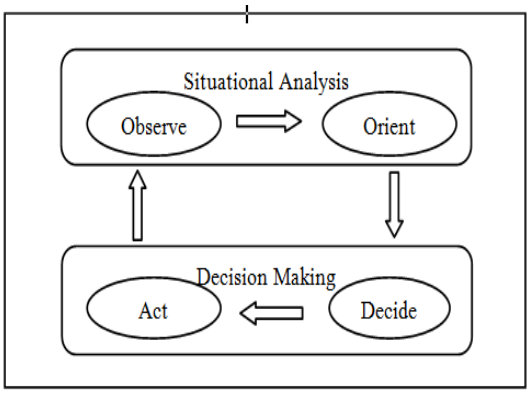


Fig.2 : Situational Awareness OODA LOOP

III. THREAT EVALUATION MODEL

The threat evaluation model with 'm' number of defended assets and 'n' number of defending targets is proposed in Figure 3. The model consists of:

1. $DA(i)$ = ith available defended asset to protect.
2. $DT(j)$ = jth attacking defending target detected.
3. $DA \times DT$= pair of available defended assets and defending targets ($DA(i),DT(j)$).
4. $Th(i,j)$=threat value from of ith available defended asset from jth attacking target.
5. $Tv(j)$=Overall Target value for defended target $DT(j)$.

$DA = \{DA_1, DA_2, \ldots, DA_m\}$=Set of 'm' defending assets to be protected. $DT = \{DT_1, DT_2, \ldots, DT_n\}$=Set of 'n' defending targets detected from radars or sensors. Threat value $Th(i,j) : (DA \times DT) \rightarrow [0\ 1]$ which lies between 0 to 1. Overall target value for defended asset is

$$Tv(i) = \frac{Pv(i) * \sum_{j=1}^{n} Th(i,j)}{n}$$

Where $Pv(i)$'s represent the protection value of assets lies between 0 and 1 ($Pv(i) \in [0\ 1]$) which is assigned by the decision maker. The protection value of an asset is the weight-value associated with the significance associated with the asset.

Threat value lies in between 0 and 1. Higher threat value indicates more severe danger threat. Threat Value helps in the decision making involved in the process of engaging the target. The overall threat values generate the knowledge which improves the situational awareness for the region of interest, which could be a city, a state, a nation or a continent. The situational awareness factor (SAF) of the region of interest is cumulative average of the target values associated with the set of defended assets.

$$SAF = \frac{\sum_{i=1}^{m} (1 - Tv(i))}{m}$$

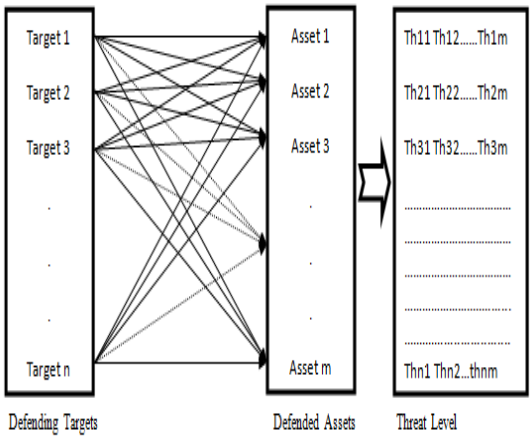


Fig. 3 : Asset- target pairs

1) *Parameters for Threat Evaluation*

The variety of parameters are proposed and used by researchers for threat evaluation [1]-[11]. These parameters have varying degree of effect on the threat value. Some parameters for calculating threat value are dependent on other parameters. A number of parameters [6] are discussed with their discriptions in Table I.

TABLE I : PARAMETER TABLE

Attribute	**Description**
Air lane	A published or otherwise known commercial air route.
Altitude	Approximate feet above ground or an indication of change (e.g., climbing).
Coordinated activity	Track is communicating with, or nearby, another track.

Course Heading	Exact compass heading or indication of heading relative to own ship (i.e., opening or closing).
CPA(Closest Point of Approach)	Closest Point of Approach Estimated distance that track will pass by own ship if the track and own ship remain on their current courses. (Figure 4)
ESM/Radar	Electronic Support - Electronic emissions from the track (typically indicates the type of radar system the track is using).
Feet Wet/Dry	A Feet Dry track is flying over land. A Feet Wet track is flying over water.
IFF Mode	Identify Friend or Foe. Signals from a track that indicate if it is a friendly, or perhaps neutral, aircraft.
Maneuvers	Indicates the number of recent maneuvers, or if the track is following the ship.
Number/ Composition	Number of aircraft in the formation.
Origin/Location	Indicates the country from which the track most likely originated.
Own Support	Availability of nearby friendly ships or patrol aircraft.
Range/ Distance	The track's distance from own ship.
Speed	Approximate airspeed or an indication of change (e.g., increasing).
Visibility	Approximate number of miles, or an indication of atmospheric conditions (e.g., haze).
Weapon envelope	The track's position with respect to its estimated weapons envelope.
Wings Clean/Dirty	A track without weapons is designated *Wings Clean*. A rack with weapons is designated *Wings Dirty*.

Closest Point of Approach (CPA) is point where the distance between asset and the direction of velocity of target will be the shortest one as seen in Figure 4, in which two targets (Target1 and Target2) are moving in different direction at particular instant with different CPA values CPA1 and CPA2.

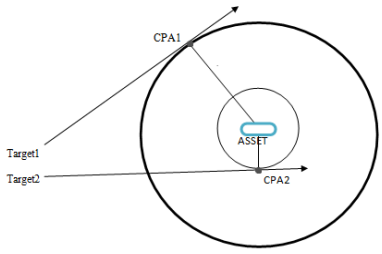


Fig. 4 : CPA information

2) Fuzzy Inference System

Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made, or patterns determined. The process of fuzzy inference involves all of the sections: Membership Function, Logical Operation, and If-Then Rules. Fuzzy inference systems have been successfully applied in fields such as automatic control, data classification, decision analysis, expert systems, modeling & simulation, and computer vision. Because of its multidisciplinary nature, fuzzy inference systems are associated with a number of names, such as fuzzy-rule-based systems, fuzzy expert systems, fuzzy modeling, fuzzy associative memory, fuzzy logic controllers, and simply (and ambiguously) fuzzy systems. There are two types of fuzzy inference system implementations, which vary somewhat in the way outputs are determined:

1. Mamdani-type and
2. Sugeno-type.

Mamdani's fuzzy inference method [5] is the most commonly used fuzzy methodology. Mamdani's method was among the first control systems built using fuzzy set theory. It was proposed in 1975 by Ebrahim Mamdani as an attempt to control a steam engine and boiler combination by synthesizing a set of linguistic control rules obtained from experienced human operators. Mamdani's effort was based on Lotfi Zadeh's 1973 paper on *fuzzy algorithms for complex systems and decision processes*. Mamdani-type inference expects the output membership functions to be fuzzy sets. After the aggregation process, there is a fuzzy set for each output variable that needs defuzzification. It is possible, and in many cases much more efficient, to use a single spike as the output membership functions rather than a distributed fuzzy set. This type of output is sometimes known as a *singleton* output membership function, and it can be thought of as a pre-defuzzified fuzzy set. It enhances the efficiency of the defuzzification process because it greatly simplifies the computation required by the more general Mamdani method, which finds the centroid of a two-dimensional function. Rather than integrating across the two-dimensional function to find the centroid, we use the weighted average of a few data points. Sugeno-type systems support this type of model. In general, Sugeno-type systems can be used to model any inference system in which the output membership functions are either linear or constant.

IV. IMPLEMENTATION

By using fuzzy logic, it will not give exact value because it is based upon ambiguous, imprecise, missing

information. The *Threat Evaluation Fuzzy Model* is a Fuzzy inference System as seen in Figure 5 with:

Input parameters= {speed, altitude, range and CPA}

Output = {threat value}

The steps involved for threat value:

1. Select target's information as inputs and threat rating as output. The target's information is collected from radar, and processed to set some information as fuzzy input evidences like: *Speed, Altitude, Range, and CPA*. The *threat value* rating is set as a fuzzy output.

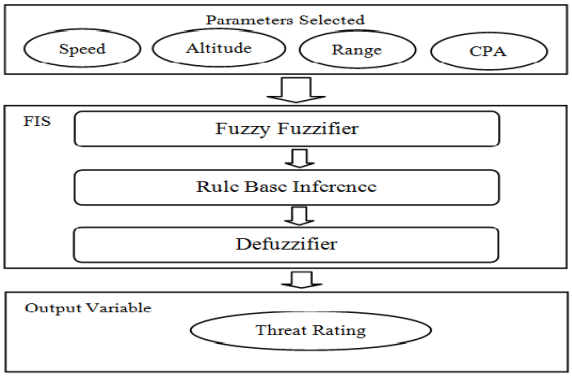


Fig. 5 : TEFM(Threat Evaluation Fuzzy Model)

2. Decide membership functions for each input and output parameters. Decide the membership functions (which lie between 0 and 1) for each input e.g. Speed, Altitude, Range, CPA and output e.g. Threat rating.

Membership functions for Speed: Generally Targets have maximum 1400 knot speed. E.g. Speed will lie in between 0 to 1400 as seen Figure 6, [0 1400]. Membership function of speed (μ_{Speed}) is triangular. Here μ_{Speed} is represented in mathematical expression:

$$\mu_{Speed=} \begin{cases} 0 & \text{if } Speed <= Sp \\ (Speed-Sp)/(Cp-Sp) & \text{if } Sp <= Speed <= Cp \\ (Ep-Speed)/(Ep-Cp) & \text{if } Cp <= Speed <= Ep \\ 0 & \text{if } Speed >= Ep \end{cases}$$

Where *Sp, Cp,* and *Ep* stand for Starting point, Center point, and End point respectively.

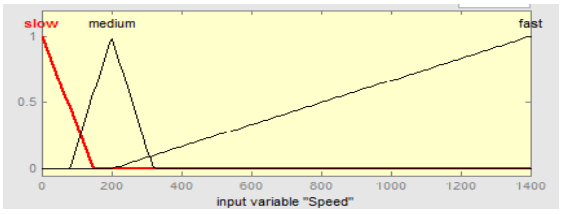


Fig. 6 : Membership functions for Speed

Membership Functions for Altitude: Generally targets can achieve maximum 50000 ft Altitude but it depends on the type of target. E.g. Altitude will lie in between 0 to 50000 as seen in Figure 7. [0 50000]

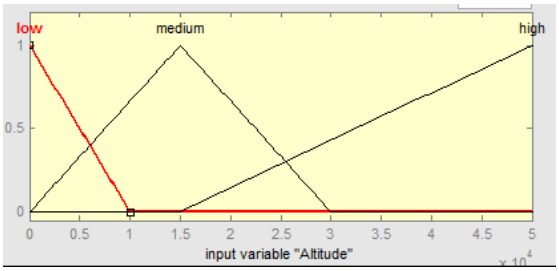


Fig. 7 : Membership functions for Altitude

Membership Functions for Range: Maximum range detected by the radar system will be 200 nautical miles but this range depends on the power of radar system. So range will lie in between 0 to 200 as seen in Figure 8. [0 200]

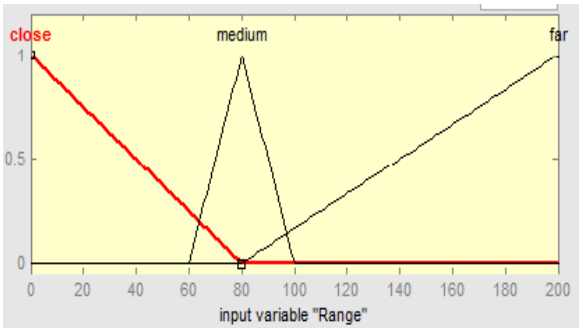


Fig. 8 : Membership functions for Range

Membership Functions for CPA: CPA can be calculated from velocity vector and position of asset. Maximum CPA is considered 200 feet CPA [0 200]. E.g. CPA will lie in between 0 to 200 as seen in Fig 9.

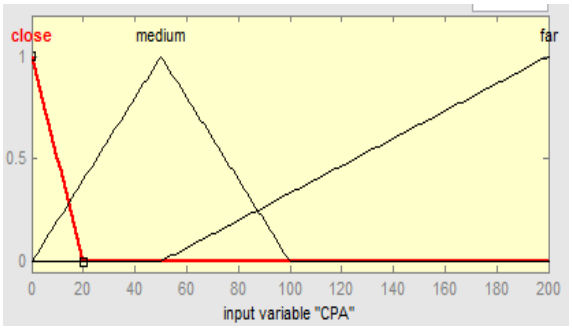


Fig. 9 : Membership functions for CPA

Membership Functions for Threat rating[0 1]: Selecting threat rating between 0 and 1 as seen in Fig.10.

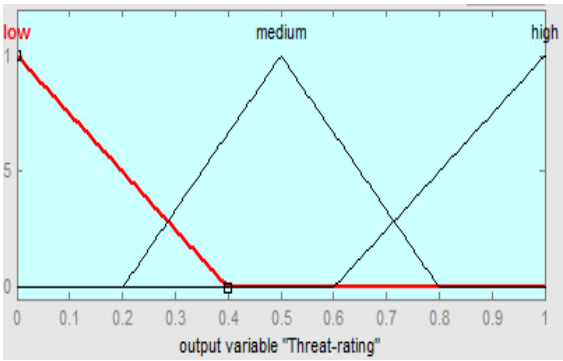


Fig. 10 : Membership functions for threat rating

3. Determine fuzzy rules by using inputs and output: Determine fuzzy inference rules using some standard data available and the expert's comments on the relation between the inputs Altitude, Speed, Range, CPA and output Threat rating. Some tentative rules are framed and the results are evaluated for the validity of the results with respect to the real time scenario. These inputs change the threat rating via rules as seen Figure 11. There are several rules given below for example:

Rule1 : If (Altitude is low) and (Speed is fast) and (Range is close) and (CPA is close) then (threat-rating is high). (Weight: 1)

Rule2 : If (Altitude is high) and (Speed is slow) and (Range is far) and (CPA is far) then (threat-rating is low). (Weight: 1)

Rule3 : If (Altitude is medium) and (Speed is medium) and (Range is medium) and (CPA is medium) then (threat-rating is medium). (Weight: 1)

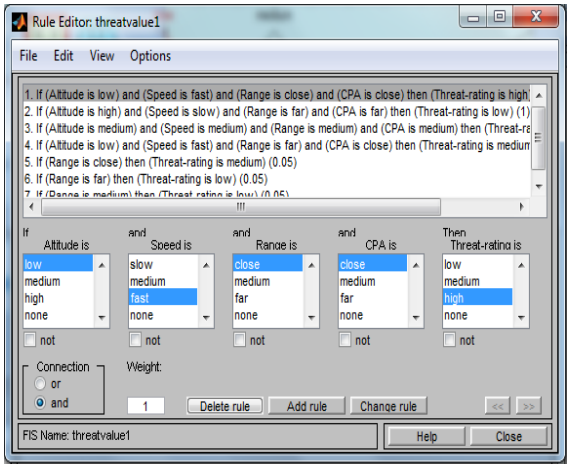


Fig. 11 : Fuzzy rule formations

4. Display and check the output using fuzzy inference systems rule viewer window, as seen in Figure 12.

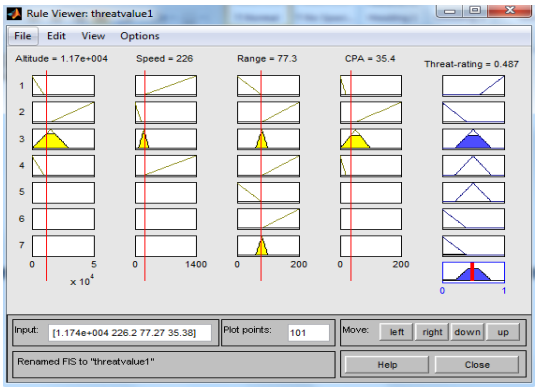


Fig. 12: Fuzzy rule viewer: to check all fuzzy rules

V. SIMULATION AND RESULTS

Simulation of proposed fuzzy model is completed for threat evaluation of targets by using GUI developed using the MATLAB software as seen in Figure 13. The figure shows a window which is reads the input parameters from user: altitude, speed, range and CPA, which are obtained from the radar system connected in the command and control unit. The underlying Fuzzy Inference System evaluates the value of threat value for the defended asset.

Simulation of this fuzzy model is done for the multiple set of inputs for the various example targets. Ex. For the input information like Altitude 1000 ft, Speed 500 knot, range 50 in nautical miles and CPA 15 in ft. the output generated is the threat rating 0.793506 (which lies between 0 and 1). It will change when values of parameter change time to time. Higher the threat rating identifies more dangerous target. The value of the threat rating will guide the decision making to engage the weapons in the process of protecting the assets from the targets.

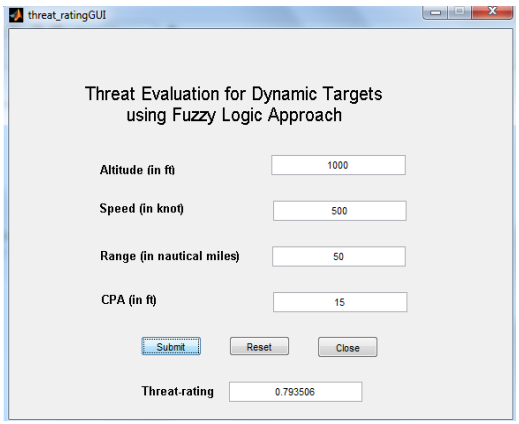


Fig. 13 : GUI for evaluating threat rating at current instant

VI. CONCLUSIONS AND FUTURE WORK

The fuzzy logic based multi objective decision making system is an excellent tool available to deploy a decision support system. It simplifies the task of human decision maker to a great deal. Each target has different threat value at different time. In this paper, threat rating of targets is effectively estimated between 0 and 1 by using fuzzy inference system which is giving accurate result.

Four parameters are introduced for threat evaluation such as altitude, speed, range and CPA as input in fuzzy inference system. Some more parameters can be introduced to improve the accuracy in the threat value evaluation. The future work would involve introduction of more parameters and the verifying the results with the actual data obtained from various sensors and the radars.

The threat evaluation system can contribute significantly in the process of the improvement of situational awareness in peace and the battlefield scenarios in a network centric operation setup. This will add value to the battle space entity in a network centric platform operations with respect to the automated decision making support.

REFERENCES

[1] Roy, J., Paradis, S., Allouche, M., "Threat evaluation for impact assessment in situation analysis systems," In: Kadar, I. (ed.) Proceedings of SPIE: Signal Processing, Sensor Fusion, and Target Recognition XI, vol. 4729, pp. 329–341, 2002

[2] Johansson, F., "Evaluating the performance of TEWA System," Orebro University, 2010

[3] Johansson, F., Falkman, G., "A Bayesian network approach to threat evaluation with application to an air defense scenario," In: Proceedings of the 11th International Conference on Information Fusion, 2008

[4] Lampinen, T., Ropponen, J., Tommi Laitinen, T., "Joint Threat Assessment with Asset Profiling and Entity Bayes Net," In: Proceeding of the 12th International Conference on Information Fusion, Seattle, WA, USA, July 6-9

[5] Liang, Y., " A fuzzy knowledge based system in situation and threat assessment', Journal of Systems Science & Information, 4, 791–802, 2006

[6] Liebhaber, M., Feher,B., " Air threat assessment: Research, model, and display guidelines," in Proceedings of the 2002 Command and Control Research and Technology Symposium

[7] Liang Y., "An approximate reasoning model for situation and threat assessment," in Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery

[8] Nguyen X., "Threat assessment in tactical airborne environments," in Proceedings of the Fifth International Conference on Information Fusion, 2002

[9] Paradis, S., Benaskeur, A., Oxenham, M., Cutler, P., "Threat evaluation and weapons allocation in network-centric warfare," In: Proceedings of the 8th International Conference on Information Fusion , 2005

[10] Ross, T.J., "Fuzzy Logic with Engineering Applications," Second Edition, John Wiley and Sons, 628

[11] Roux, J.N., Van Vuuren J.H., "Threat evaluation and weapon assignment decision support: A review of the state of the art, ORiON," vol. 23, pp. 151–186, 2007

◈◈◈

€XML Tree Pattern Matching Algorithms

Lakshmi Tulasi.Ambati, Y.SSR.Murthy & L.Balaji

Computer Science Engineering, Shri Vishnu Engineering College For Women,Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India.
E-mail : lakshmitulasi.ambati@gmail.com, yssrmoorhty@gmail.com, balucse@svecw.edu.in

Abstract **-** In the present day digital world, it is imperative that all organizations and enterprises facilitate efficient processing of queries on XML data. XML queries typically specify patterns of selection predicates on multiple elements that have specified tree structured relationships. The primitive tree-structured relationships are parent-child and ancestor-descendant. Finding all occurrences of these relationships in an XML database is a core operation for XML query processing. In this paper the pattern matching algorithms TwigStack and TwigStackList are discussed. The behavior of TwigStack is analyzed, and a comparison of these two algorithms is attempted. The TwigStack algorithm the initial holistic algorithm, has features of performing simultaneous scan over streams of XML nodes to match their structural relationships holistically, reducing a number of unnecessary intermediate results, and skipping XML nodes that will not contribute to final answers. The family of holistic pattern matching algorithms has appeared as the major important algorithms for processing XML query patterns due to its efficiency and performance advantage. The experimental results show that the query performance is significantly improved especially for queries having relatively more complex structures and/or higher selectivities.

Keywords **-** *Xml,TwigStack,TwigStackList;*

I. INTRODUCTION

XML employs a tree-structured model for representing data. In Xml, XPath and XQuery [1] are used for addressing the parts of an xml document and for specifying patterns of selection predicates on multiple elements that have specified tree-structured relationships. For example, the XQuery path expression

Book [author=suciu] // [title=XML]

An XML tree pattern query, represented as a labeled tree, is essentially a complex selection predicate on both structure and content of an XML. Tree pattern matching has been identified as a core operation in querying XML data. The data in Xml is arranged by using the grammar DTD (Document Type Definition) fig 2. In web mining, the data is retrieved from web through XML tree. The XML tree gives all relevant information to the users of the web. Xml allows for structuring of data on the web. The structure of XML data is represented in fig 1. An XML document is made of elements limited by tags and is hierarchically structured.

II. BACKGROUND

The extensible markup language XML has recently emerged as a new standard for information representation and exchange on the Internet. XML allows users to make up any new tags for descriptive markup of their own applications. Since XML data is self-describing, XML is considered one of the most promising means to define semi-structured data, which is expected to be ubiquitous in large volumes from diverse data sources and applications on the web. In Xml Tree there is a Parent-Child (P-C) and Ancestor and Descendant (A-D) relationships which are represented as / and // in fig 3. A tree which is maintained by both Parent-Child (P-C) and Ancestor and Descendant (A-D) relationships is presented in fig.3. There are some pattern matching algorithms [4][5], which are not much efficient than TwigStack [2]discussed in III. Twigstack implementation is discussed in section IV. To overcome some limitations of TwigStack a TwigStack List is discussed in section V. Section VI concludes the paper. TwigStack [2] is one of the pattern matching algorithms, which can efficiently retrieve information much faster than many other algorithms [4][5]. TwigStack [2] is optimal for tree pattern queries with only A-D edges. In other words, TwigStack [2] processes the tree pattern holistically without decomposing into several small binary relationships. TwigStack [2] guarantees that there is no useless intermediate result for queries with only Ancestor-Descendant (A-D) relationships.

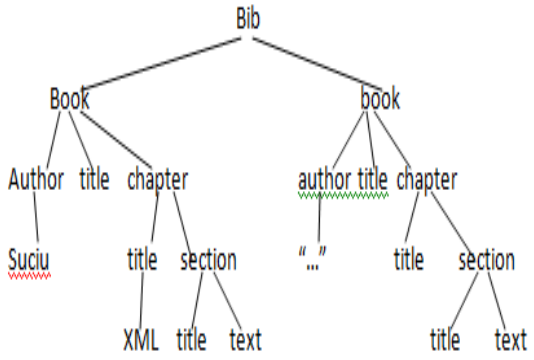


Fig. 1: An XML Tree Representation

Algorithm TwigStack operates in two phases. In the first phase (lines 1-11),some (but not all) solutions to individual query root-to-leaf paths are computed. In the second phase (line-12), these solutions are merge-joined to compute answers to the query twig pattern as delineated in fig 4.

<!ELEMENT bib (book*)>

<!ELEMENT book (author+, title, chapter*)>

<!ELEMENT author (#PCDATA)>

<!ELEMENT title (#PCDATA)>

<!ELEMENT chapter (title, section*)>

<!ELEMENT section (title, (text | section)*)>

<!ELEMENT text (#PCDATA|bold |keyword| emph) *>

<!ELEMENT bold(#PCDATA|bold|keyword | emph)*>

<!ELEMENT keyword (#PCDATA | bold | keyword | emph)*>

<!ELEMENT emph (#PCDATA | bold | keyword | emph)*>

Fig. 2 : An DTD For XML Data

//Phase 1

1 while ~end(q)
2 q_{act} = getNext(q)
3 If (~isRoot(q_{act}))
4 cleanStack(parent(q_{act}), nextL(q_{act}))
5 If(isRoot(q_{act}) V ~empty(S parent(q_{act})))
6 cleanStack (q_{act}, next(q_{act}))
7 moveStreamToStack (T_{qact},S_{qact}, pointer to top($S_{parent(q\ act)}$))
8 if (isLeaf(q_{act}))
9 showSolution WithBloacking(S_{qact},1)
10 Pop(Sq_{act})
11 else advance(T_{qact})

//Phase 2

12 mergeAll PathSolutions()

Function getNext(q)

1 if (isLeaf(q)) return q
2 for q_i in children(q)
3 n_i=getnext(q_i)
4 If(n_i isnotEqualto q_i) return n_i
5 n_{min} = minarg n_i , nextL(T_{ni})
6 n_{max} = maxarg n_i , nextL(T_{ni})
7 while (nextR(Tq) < nextL(T_{nmax}))
8 advance(Tq)
9 If (nextL(Tq) < nextL(T_{nmin})) return q
10 Else return n_{min}

Procedure cleanStack(S, actL)

1 while (~empty(S) and (topR(S) <actL))
2 pop S.

Fig. 4 : TwigStack Algorithm.

III. TWIGSTACK IMPLEMENTATION

TwigStack [2] 1) avoids generating large intermediate results which do not contribute to the final answer, 2) avoids unnecessary scanning of source documents, 3) avoids unnecessary scanning of irrelevant portions of XML documents.

For example, the query is

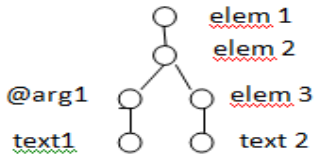
/library/category[@name=*France*]/book/title[@language=*English*]

At each node a stack is maintained by the TwigStack [2] algorithm. A diagrammatic representation of the processing of a query is made in fig.5. And how the data is arranged in the stack in each and every node is presented in fig 6.

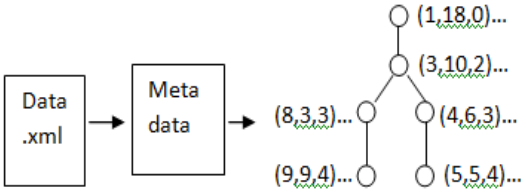
XPath Query String

Elem1/elem2{@arg1=[text][/elem3=text2]

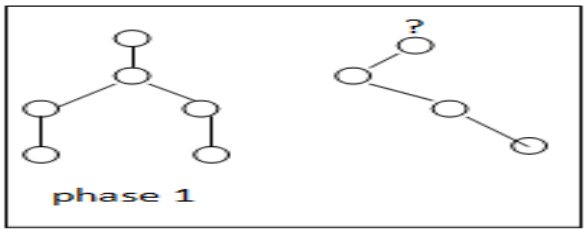
Query Tree



Query Tree With Metadata

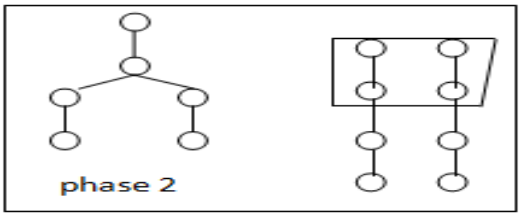


Phase1:



Path List=Intermediate result

Phase:2



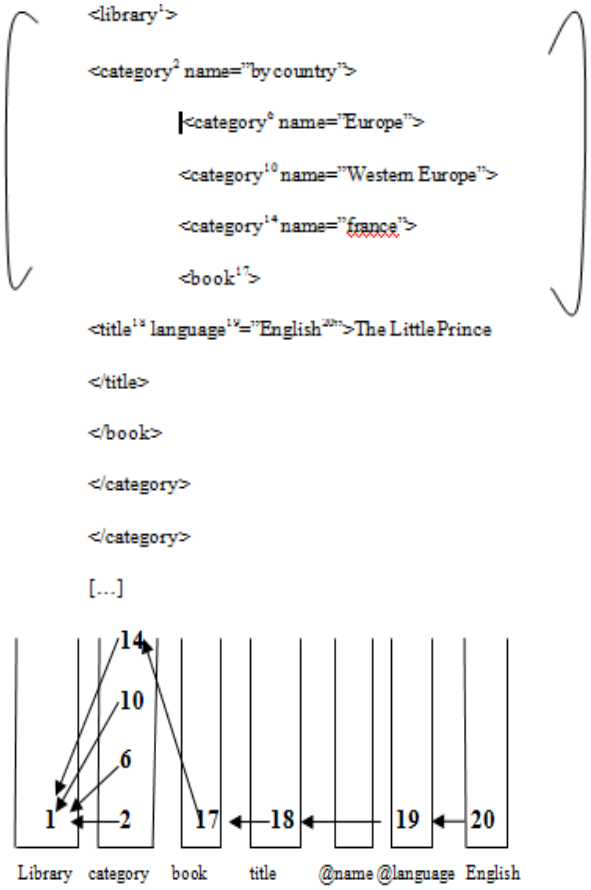
TreeList=result.

Fig. 5: TwigStack implementation.

In fig 5, the TwigStack algorithm comprises two tasks. The first task is to perform query pattern matching against XML data and to generate partial solutions. Meanwhile, the second task is to merge the partial solutions generated by the first task for final solutions.

/library//category[@name=france]//book/title[@language=English]

Same XML tag Can be nested



Query Node Stack

Fig. 6 : Arranging of data in stack at each node.

The values of a stack for a query are shown in fig.6 Query Node Stack. The limitations of TwigStack [2] Algorithm are redundancy is maintained, retrieving of data through XML is not much faster than TwigStackList [3], the efficiency of retrieving large queries in XML data is not effective and the intermediate results are not reduced.

IV. TWIGSTACKLIST

TwigStackList [3] is combination of TwigStack [2] and Lists. It improves efficiency of large queries on XML data and overcomes the limitations of redundancy in TwigStack [2]. The tree structure of XML data using

TwigStackList [3] is shown in fig 7. At each node the stack and lists are maintained.

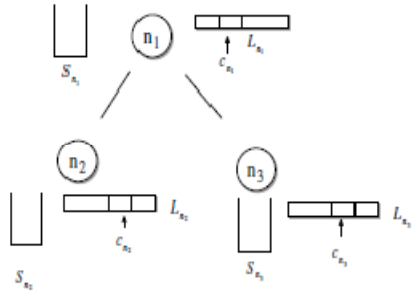


Fig. 7 : TwigStackList [3] where 'Sn' stands for Stacks and 'Ln' for Lists.

TwigStackList [3] operates in two phases. In the first phase (line 1-11), it repeatedly calls the *getNext* algorithm with the query root as the parameter to get the next node for processing. We output solutions to individual query root-to-leaf paths in this phase. In the second phase (line 12), these solutions are merge-joined to compute the answer to the whole query. The *getNext* algorithm is presented in fig 8 and TwigstackList [3] Algorithm in fig 9.

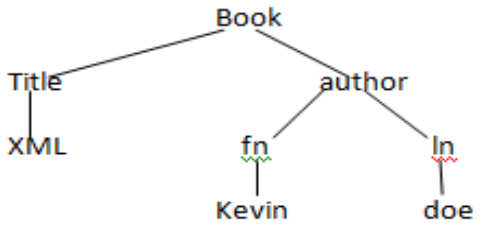


Fig. 3 : XML tree with A-D and P-C relationships.

At line 2-5, in Algorithm *getNext*, we recursively invoke *getNext* for each n_i *2 children* (*n*). If any returned node g_i is not equal to n_i , we immediately return g_i (line 4). Line 6 and 7 get the *max* and *min* elements for the current head elements in lists or streams, respectively. Line 8 skips elements that do not contribute to results. If no common ancestor for all *C* n_i is found, line 9 returns the child node with the smallest start value, i.e. *gmin* . Line 10 is an important step. Here we look-ahead read some elements in the stream *Tn* and cache elements that are ancestors of *Cnmax* into the list *Ln*. Whenever any element n_i cannot its parent in list *Ln* for n_i *2 children*(*n*), algorithm *getNext* returns node n_i (in line 17). In *TwigStack[2]*, *getNext*(*n*) return *n0* if the head element *en0* in stream *Tn0* has a descendant *e* n_i in each stream Tn_i , for *ni 2 children*(*n0*) and *getNext*(*root*) in *TwigStackList[3]* returns *b*1.

By using TwigStackList Algorithm, we can reduce the intermediate results of a query on xml data, and thereby reduce the redundancy level in TwigStack[2].

Algorithm 1 getNext(n)

1 If isLeaf(n) retun n
2 For all node n, n children(n) do
3 g_i = getNext(n_i)
4 If (g_i isNotEqualTo n_i) return g_i
5 End for
6 n_{max} = maxarg n_i ← children(n) getStart(ni)
7 n_{min} = minarg n_i ← children(n) getStart(ni)
8 while (getEnd(n) < getStart(nmax)) proceed(n)
9 if (getStart(n) > getStart(nmin)) return n_{min}
10 MoveStreamToList(n, n_{max})
11 For all node n_i in PCRchildren(n) do
12 If (there is an element e_i in listLn such that e_i is the parent of getElement(ni)) then
13 If(n_i is the only child of n) then
14 Move the cursor p_n of list Ln to point to e_i
15 end if
16 End for
17 Return n

Procedure getElement(n)

1. If ~empty(Ln) then
2. return Ln.elementAt(p_n)
3. Else return cn

Procedure getStart(n)

1. return the start attribute of getElement(n)

Procedure getEnd(n)

1 return the end attribute of getElement(n)

Procedure MoveStreamToList(n,g)

1 while Cn.start < getStart(g) do
2 if Cn.end > getEnd(g) then

2 Ln.append(Cn)

3 end if

4 advance(Tn)

5 end while

procedure proceed(n)

1 if empty(Ln) then

2 advance(Tn)

3 else

4 Ln.delete(Pn)

5 Pn =0 {Move pn to the point to the beginningof Ln}

6 End if

Fig. 8 : *getNext* algorithm

Algorithm 2 TwigStackList

1 While ~end() do

2 n_{act} = getNext(root)

3 If (~isRoot(n_{act})) then

4 cleanparentStack(n_{act}, getStart(n_{act}))

5 end if

6 if (isRoot(n_{act})V~empty(Sparent (n_{act})) then

7 clearSelfStack(n_{act}, getEnd(n_{act}))

8 moveToSatck(n_{act},S_{nact},pointertotop($S_{parent(nact)}$)

9 if (isLeaf(n_{act}) then

10 showSolutionsWithBloacking(S_{nact},1)

11 pop(S_{nact})

12 endIf

13 else

14 proceed(n_{act})

15 endif

16 end while

17 mergeAllPathSolutions

Function end()

1 return n_i subtreeNodes(n): isLeaf(n_i) and endC(n_i)

Function moveToStack(n, S_n,p)

1 push (getElement(n),p) toStack S_n

2 proceed(n)

Procedure clearparentSatck(n, actStart)

1 while(~empty$S_{parent(n)}$)

^ topEnd($S_{parent(n)}$)<actStart)) do

2 pop($S_{parent(n)}$)

3 end while

procedure clearSelfStack(n, actEnd)

1 while (~empty(S_n) and topEnd(S_n)<actEnd) do

2 pop(S_n)

3 end while.

Fig. 9: TwigStackList Algorithm

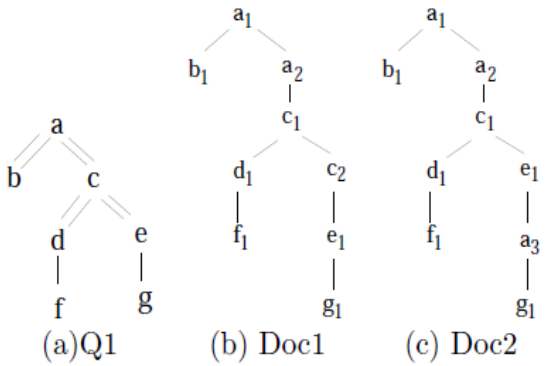


Fig.10:Example TwigQuery and Documents

TwigStack [2] pushes *c*1 to *stack Sc* and outputs two *useless*" intermediate path solution <*a*1*; b*1> and <*a*1*; c*1*; d*1*; f*1>. The behavior of TwigStack[2] is also reasonable because based on *region coding* of *g*1, one cannot decide whether *g*1 has the parent tagged with *e*. TwigStackList[3] does not hastily push *c*1 to stack , but first checks the parent-child relationship between *e*1 and *g*1. If *e*1 is not the parent of *g*1, then TwigStackList[3] caches *e*1 in a list and reads more elements in *Te*. In this simple case, *e*1 is the only element in stream *Te*

V. CONCLUSION

The XMl tree construction and importance of pattern matching algorithms for searching the data is discussed. The TwigStack Algorithm has a Time complexity but, the limitation is space complexity. How the TwigStackList overcomes the limitations of TwigStack in reducing the intermediate results in a query on XML data has been elaborated upon.Our experiments have demonstrated that these pattern matching algorithms have an edge over other pattern matching algorithms

REFERENCES

[1] A. Berglund, S. Boag, and D. Chamberlin, XML Path Language (XPath) 2.0, W3C recommendation, http://www.w3.org/TR/xpath20/, Jan. 2007.

[2] N. Bruno, D. Srivastava, and N. Koudas, "Holistic Twig Joins: Optimal XML Pattern Matching," Proc. ACM SIGMOD, pp. 310-321, 2002.

[3] J. Lu, T. Chen, and T.W. Ling, "Efficient Processing of XML TwigPatterns with Parent Child Edges: A Look-Ahead Approach," Proc. 13th ACM Int'l Conf. Information and Knowledge Management (CIKM), pp. 533-542, 2004.

[4] Q. Li and B. Moon, "Indexing and Querying XML Data

[5] For Regular Path Expressions," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 361-370, 2001.

[6] H. Jiang et al., "Holistic Twig Joins on Indexed XML Documents,"Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 273-284, 2003.

◈◈◈