Proceedings
of
National Conference
on

Data Mining

(NCDM–2011)

In the fields of
Bioinformatics, Medical-Informatics,
Agro-Informatics and Business Management

www.excelpublish.com

Proceedings
of
National Conference
on

Data Mining

(NCDM–2011)

In the fields of
Bioinformatics, Medical-Informatics,
Agro-Informatics and Business Management

Editors
Dr. H.S. Acharya
Prof. M.M. Junaid F.

Organized by
MCE Society's
Allana Institute of Management Sciences,
Azam Campus, 2390-B K.B. Hidayatull Road
Camp, Pune

Conference Sponsored by
Indian Association for Medical Informatics
Computer Society of India
International Neural Network Society

EXCEL INDIA PUBLISHERS
New Delhi

First Impression: 2011

***Proceedings of National Conference on Data Mining (NCDM–2011*)**

ISBN: 978-93-81361-26-9

DISCLAIMER

The authors are solely responsible for the contents of the papers compiled in this volume. The publishers or editors do not take any responsibility for the same in any manner. Errors, if any, are purely unintentional and readers are requested to communicate such errors to the editors or publishers to avoid discrepancies in future.

Published by
EXCEL INDIA PUBLISHERS
61/28, Dalpat Singh Building, Pratik Market, Munirka, New Delhi-110067
Tel: +91-11-2671 1755/ 2755/ 3755/ 5755 ● Fax: +91-11-2671 6755
E-mail: publishing@excelpublish.com
Website: www.excelpublish.com

Typeset by
Excel Publishing Services, New Delhi-110067
E-mail: prepress@excelpublish.com

Printed by
Excel Printing Universe, New Delhi-110067
E-mail: printing@excelpublish.com

Preface

Data warehousing and Data Mining are amongst the emerging trends of the day. India being the world's second most populous country, is a very significant contributor of data related to agriculture, healthcare, education etc, which are of tremendous social welfare importance. The National Informatics Centre (NIC), which is a prime government established organization, has developed its own data warehouse called GISTNIC , to take care of information services in this regard. It offers online information retrieval services from its wide ranging static and quasi-static areas such as Indian Economy monitor, IMF databases, 1991 population Census Database, Village Amenities Database, Rural Technology Database, Tourist Guide of India, District Profiles, University and College directory of India and a Database on Traditional Sciences and Technologies of India. In this context the areas of agroinformatics, medical informatics and datamining in general have more relevance to our scientists and research workers. The abundance of data available on this facility can be put to effective use by information scientists for research and development. Soon the world would be needing experts in this area to cater to the needs of information seekers. This volume starts with an excellent articles which is basically the key note address delivered by none other than Mr. M. Moni, the Deputy Director General of, NIC.

Natural resource conservation, minimum risk agriculture, quality medical care have been most fundamental requirements that determine the status of the welfare of the society. Domain knowledge clubbed with IT , can infuse precision and quickness and work wonders. Traditionally 'Statistics' , and now 'data mining' in the modern days of ICT adoption, have been providing information support to critical decision makers. This certainly will have a significant influence on performance of both the medical sector and agricultural sector, in addition to contributing to growth of IT sector itself. Best foot in this regard has been already put forth by the Allana Institute of Management Sciences in bringing together experts from Medicine, Agriculture, Biology and IT .

We have received more than 25 contributions out of which six are on medical informatics, five are on agro informatics , four on Bioinformatics, two on business management and three on generic concepts. The domains range from application of data mining to Unani diagnostics to recovery of NPA in banks. This Proceedings can be a useful resource for the research workers, students, IT professionals and academia alike. This publication is one of the series of quality technical publications the MCA department of AIMs is planning to bring out in the coming years. Our target would be mainly to produce value added books which aim at supporting interdisciplinary research and learning of subjects which are primarily interdisciplinary with IT as one of the components.

Such a product would not have become a reality without support from many. We would like to thank Hon P.A. Inamdar, President MCE Society Mrs. A.P. Inamdar Vice President MCE society, Prof. R,. Ganesan Director Allana Institute of Management sciences for the valuable infrastructural support. We are also thankful and acknowledge efforts of Prof Jawed Khan the chief organizer of the conference and his team members of organizing committee for such a wonderful conference. We will also thank all the participants for their contribution. The Technical contribution were reviewed and evaluated by our reviewers and the members of editorial board, we thank them for their invaluable contribution. Special thanks to our young editorial assistants Prof. Afroz Sheikh and Prof. Ashwini Mohan. Finally we thank our Excel India Publishers, New Delhi for careful handling of the manuscript both at the editorial and Production stages.

Editors
Dr. H.S. Acharya
Prof. M.M. Junaid

Foreword

Medical practitioners, drug discoverers, biotechnologists, agricultural scientists and practitioners, and business men are always on the search for innovative ideas. The ideas which may help them improve current processes, and take better risk free decisions. When a doctor diagnoses, minimization of the consequences of a wrong decision are always at the back of his mind. When a biotechnologist searches for a new sequence possibility of a better drug design is his ultimate goal. When an agricultural scientist looks for patterns in the past data he is looking for better practices with minimum risk and assured returns. They are fully aware of the uncertainties which are needed to be tackled. They need to unearth these from the huge past data that is lying with them in most cases.

Data mining, which is also fondly called as patterns analysis on large sets of data, uses tools like association, clustering, segmentation and classification for helping better estimations and reliable inferences. Analysis of the data help the business firms to discover new ways to lower costs while improving the product and delivery methods with better chance of countering the competition. It is a race where speed with which you discover new processes, new patterns is what lets you successfully compete.

A deep understanding of the knowledge hidden in the organizational data is vital to a firm's competitive position and organizational decision-making. Traditional data analysis methods often involve manual work and interpretation of data that is slow, expensive and highly subjective. Data Mining enables firms to make calculated decisions by assembling, accumulating, analyzing and accessing corporate data. It uses variety of tools like query and reporting tools, analytical processing tools, and Decision Support System (DSS)

Data mining has been used extensively in the banking and financial markets. In the banking industry, it is heavily used to model and predict credit fraud, to evaluate risk to perform trend analysis, to analyze profitability, as well as to help with direct marketing campaigns.

A particular active area of research is the application and development of data mining techniques to solve biological problems. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on micro array data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc. Therefore, we see a great potential to increase the interaction between data mining and bioinformatics.

In agriculture , data related to production, consumption, agricultural marketing, fertilizer consumption. Seeds, prices (wholesale as well as retail) technology, agricultural census, marketing regions, live stock, crops, agricultural credit, plant protection, watershed, area under productions yields, land use statistics, finance and budget etc, can be mined to reach to important conclusions and then to take decisions based on these conclusions.

Progress in data mining applications and its implications are manifested in the areas of information management in healthcare, health informatics, epidemiology, patient care and monitoring systems, assistive technology, large-scale image analysis to information extraction and automatic identification of unknown classes. Various algorithms associated with data mining have significantly helped to understand medical data more clearly, by distinguishing pathological data from normal data, for supporting decision-making as well as visualization and identification of hidden complex relationships between diagnostic features of different patient group

An exclusive conference, named the ***National Conference on Data Mining (NCDM–2011)*** , aimed at bringing researchers in data mining working in these diversified domains , was organized by us 19th and 20th February, 2011. We received an overwhelming response from the enthusiastic participants. The Policy of Allana Institute of Management Sciences is to nurture and promote talent in research. This volume is a direct outcome of the effort. I congratulate my staff who have put in lot of efforts in bringing up this publication and sincerely hope the contents will be a helpful to all those who want to pursue research in the area of data mining.

Prof. R. Ganesan
Director
Allana Institute of Management Sciences

Editorial Board

Patrons

Mr. P.A. Inamdar
President MCE Society

Mrs. Abeda Inamdar
Vice President MCE Society

Prof. R. Ganesan
Director, Allana Institute of Management Sciences, Pune

NCDM–2011Committees

Chief Organizer

Prof. Jawed S. Khan, HOD MCA
Allana Institute of Management Sciences

Advisory Committee

Prof. R. Ganesan	Allana Institute of Management Sciences, Pune
Dr. A.B. Rao	Allana Institute of Management Sciences, Pune
Dr. H.S. Acharya	Allana Institute of Management Sciences, Pune
Dr. Manik S. Kadam	JSIMR, Pune
Dr. S.B. Padhey	Inter Disciplinary Research Group Azam Campus Pune
Dr. Amol Goje	VIIT, Baramati
Dr. Mrs. Swati Sirdesai	NIC, Pune
Dr. Swarnalata Rao	Division V., Education and Research, CSI
Dr. Atanu Rakshit	Isquare IT, Pune
Dr. E.M. Khan	Abeda Inamdar Sr. College, Pune
Dr. Jalis Ahmed	Z.V.M. Unani Medical College, Pune
Dr. Kiran Bhise	Allana College of Pharmacy, Pune
Dr. K.V. Kale	Dr. BAMU, Aurangabad
Dr. H.K. Mishra	Institute of Rural Management, Anand, Gujarat
Dr. D.S. Bormane	Rajeshri Shahu College of Engg., Pune
Prof. Nandakumar Kachane	Institute of Management & Computer Studies, Pune
Dr. Hiten Lakhey	Trinity Consultant, Pune
Mr. A.S. Pawar	CSI, Pune
Dr. Roshan Kazi	Allana Institute of Management Sciences, Pune
Dr. S.C. Shirwaikar	Nowrosjee Wadia College, Pune

Organizing Committee

Prof. Farhana Sarkhawas
Prof. Tajuddin Bennur
Prof. Mehdi Ali Jafri
Prof. Abhijit Kaiwade
Prof. Jyoti Mulchandani
Prof. Afroz Shaikh
Prof. Tasnim Kayamkhani
Prof. Rajesh More
Prof. Rahila Sayed

Editorial Assistance

Prof. Afroz Shaikh
Prof. Ashwini Mohan

Session Chairs

Dr. H.K. Misra
Dr. H.S. Acharya
Prof. Suash Deb
Dr. Manik Kadam
Dr. Roshan Kazi
Dr. S.M. Abbas
Dr. Pravin Metkewar

Contents

Knowledge Discovery and Mining-Discovering Hidden Value in Warehouses to Strengthen G2C Model of E-Governance Programme in India

Madaswamy Moni
Deputy Director General National Informatics Centre Government of India New Delhi–110 003
e-mail: moni@nic.in

It is, indeed, my pleasure, to attend and deliver the Keynote Address in the National Conference on Data Mining (NCDM-2011), being orgnaised by M.C.E Society's Allana Institute of Management Sciences (AIMS), in collaboration with Computer Society of India (CSI).

I am very happy to note that the prime goal of the Allana Institute of Management Sciences (AIMS) aims at producing professionals in the area of Information Technology and Business, and empowering them to solve the technical and business issues, emerging out of global environment. It is good to know that AIMS is aiming at building Supply-Chain Models, Value-Chain Models, and Results-Chain Models etc., so as to have the Return-on-Investment (ROI) at the appreciatable level. Since the Institute is in the discipline of Management Science, I hope so, the AIMS introduced the theme: Business Management to understand supply-Chain Models (retail management) and CRM Models (Relationship management) in Agriculture and Medicine; and also as how to embed "business analytics" to exhibit "business intelligence".

The objective of NCDM 2011 is to bring hardcore IT professionals, academia and research workers in the areas of Data mining, to the fields of Bioinformatics, Medical informatics, Agro Informatics & Business Management, and also to enhance the skill levels in Data mining Applications. This is a laudable venture by an Institution of Merit viz., AIMS, in collaboration with Computer Society of India (CSI). I am very confident with respect to Agricultural Informatics and Business Management but I will venture into both Bioinformatics and medical Informatics as well.

Technologies such as Information Technology, Database Technology, GIS Technology, GPS technology, Remote Sensing Technology, Image Processing Technology, Data Warehousing / Data Mining techniques, etc are essential to convert data into information, knowledge and wisdom... in that order. However, according to the Economist (November 2001 issue), Agriculture (genetic modification), Medicine (genome research and bioinformatics) and Information & Communication Technologies (ICTs) are the three fields where diffusion of technology holds particular promise for the poor.

In the area of Bioinformatics, the Conference is looking into the research and applications of Pattern Recognition, Sequence Classification, Microarray Analysis, Phyto genetics and Public Databases. Recent advances in bioinformatics and high-throughput technologies such as microarray analysis are bringing about a revolution in our understanding of the molecular mechanisms underlying normal and dysfunctional biological processes. Microarray studies and other genomic techniques are also stimulating the discovery of new targets for the treatment of disease which is aiding drug development, immunotherapeutics and gene therapy [1]. This requires handling of protocols, instruments, hardwares, softwares, databases and validations. Capacity and capability building is more required.

In the area of Medical informatics, the Conference has gone one step forward to look into the aspects of: Text Mining and ontologies, Pattern Prediction, Knowledge management, Biomedicine and Health Care. Biomedicine is a branch of medical science that applies biological and other natural-science principles to clinical practice. Biomedicine involves the study of (patho-) physiological processes with methods from biology, chemistry and physics [2].

Medical Informatics and biomedical computing have grown in quantum measure over the past decade. We should recognise that an abundance of advances have come to the foreground in this field with the vast amounts of biomedical and genomic data, the Internet, and the wide application of computer use in all aspects of medical, biological, and health care research and practice. It is the research and applications in the areas of Knowledge Management and data mining in Biomedicine that are raising the technical horizons and expanding the utility of informatics to an increasing number of biomedical professionals and researchers.

Biology is rich in data, and is getting richer all the time. Recent advances in DNA sequencing, microarray data generation, high-throughput, gene-function studies, medical imaging, and electronic medical records (EMR) have resulted in the automatic generation of new, vast, and exciting databases. Deriving "big pictures from this sea of biomedical data," as described in the July 2005 issue of Science, is a major scientific challenge that will require the close collaboration of computer scientists, biologists, and mathematicians.

While tremendous progress has been made over the years, many of the fundamental problems in bioinformatics, such as protein structure prediction, gene-environment interaction, and regulatory pathway mapping, are still open, according to Internet resources. Data mining will play essential roles in understanding these fundamental problems and development of novel therapeutic/diagnostic solutions in post-genome medicine.

Bioinformatics offers numerous challenges. How to facilitate “Knowledge Discovery” in a complex biological system”? Different from analyzing single molecules, complex biological systems consist of components that are in themselves complex and interacting with each other. Understanding how the various components work in concert, using modern high-throughput biology and data mining methods, is crucial to the ultimate goal of genome-based economy such as genome medicine and new agricultural and energy solutions [3]:-

- Phylogenetics and comparative Genomics
- DNA microarray data analysis
- RNAi and microRNA Analysis
- Protein/RNA structure prediction
- Sequence and structural motif finding
- Modeling of biological networks and pathways
- Statistical learning Methods in Bioinformatics
- Computational proteomics
- Computational biomarker discoveries
- Computational drug discoveries
- Biomedical text mining
- Biological data management techniques
- Semantic webs and ontology-driven biological data integration methods

In the Area of Agro Informatics, the Conference is interested to understand issues related to: Risk Management, Precision Farming, GIS Applications and Meteorology. This area is of my interest and I am instrumental in evolving Agricultural Informatics as a discipline in India. I have been working in this area during the last 30 years and many of ICT projects in the country, which I will deal in my address later.

In a nutshell, the NCDM-2011 will discuss research papers in the areas related to: Pattern Recognition, Pattern Prediction, Sequence Classification, Text Mining and Ontologies, Knowledge management, GIS Applications, Public Databases, Risk Management, Microarray Analysis, Phyto genetics, Biomedicine and Health Care, Precision Farming, and Meteorology.

I do have every reason to hope that the Research Scholars, Academia and IT professionals would go back to their Institutions, with a lot of professional enrichment in the themes, proposed by this NCDM-2011.

I. Data Mining–A Powerful Technology Tool

Data Mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help organisations and Governments focus on the most important information in their data warehouses for decision making. Data mining tools predict future trends and behaviors, allowing organisations and Governments to make proactive, knowledge-driven decisions. We know that the Government departments collect massive data, using statistical methods, to facilitate policy formulations at macro and micro levels.

Data mining Algorithms are the result of a long process of research and product development, and can be adopted to enhance the values of existing information resources (i.e. data warehouses) in digital form. Data Mining is possible as (a) the Governments and Organisations collect massive data; (b) the Governments and Organisations have data centres, and (c) Data Mining Algorithms area available both in Open Source and Proprietary Source.

In India, the Public Sector has Internet Data Centres at Central level, Regional level and Provincial levels. The on-going National Knowledge Network (NKN), with the minimum internet speed of, not less than one GIGA bit, will strengthen adoption of data mining, in a large scale.

To refresh our memory, in 1960s, retrospective, static data delivery was facilitated through computers, Tapes and Disks. This is a “Data Collection Era”. In 1980s, retrospective, dynamic data delivery at record level was facilitated through Relational databases (RDBMS), Structured Query Language (SQL) and ODBC. This is termed as “Data Access Era”. In 1990s, retrospective, dynamic data delivery at multiple levels was achieved through On-line analytic processing (OLAP), multidimensional databases and data warehouses. This is termed as a “Data warehouse and Decision Support Era”.

There is a growing gap between more powerful storage and retrieval systems and the users’ ability to effectively analyze and act on the information they contain. Both Relational and OLAP (On-Line Analytical Processing) technologies have tremendous capabilities for navigating massive data warehouses, but

brute force navigation of data is not enough. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools are making this leap [4]. Prospective, proactive information delivery is progressively facilitated using advanced algorithms, cloud computing and massive databases. This is termed as "Data Mining Era", as emerged today. The most commonly used techniques in data mining are: (a) Artificial neural networks (ANN), (b) Decision trees, (c) Genetic algorithms, (d) Nearest neighbor method, and (e) Rule induction.

II. Text Mining- to Glean Meaningful Information from Natural Language text

Labor-intensive manual text mining approaches first surfaced in the mid-1980s, but technological advances have enabled the field to advance during the past decade. Text mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics, and computational linguistics [5]. As most information (common estimates say over 80%) is currently stored as text, text mining is believed to have a high commercial potential value. Increasing interest is being paid to multilingual data mining: the ability to gain information across languages and cluster similar items from different linguistic sources according to their meaning. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities). Text mining results can be incorporated in Data Mining Projects viz., graphics (visual data mining methods), Clustering and factoring, and Predictive data mining.

In Text Mining, patterns are extracted from natural language text rather than databases (data mining), and the input is free unstructured text, whilst web sources are structured (web mining). Computation Linguistics (CPL) computes statistics over large text collections in order to discover useful patterns which are used to inform algorithms for various sub-problems within Natural Language Processing (NLP). Billions of documents must be handled in an efficient manner, as these are intended for different consumers, i.e. different languages (human consumers) and different formats (automated consumers). The National Centre for Text Mining (NaCTeM) is the first publicly-funded text mining centre in the world [6].

III. Web Mining - Information and Pattern Discovery on the World Wide Web

With the explosive growth of information sources available on the World Wide Web (WWW), it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources. Web mining (Web content mining and Web usage mining) can be broadly defined as the discovery and analysis of useful information from the WWW. Web mining has adapted techniques from the field of data mining, databases, and information retrieval, as well as developing some techniques of its own, e.g. path analysis. A lot of work still remains to be done in adapting known mining techniques as well as developing new ones. Specifically, Robert Cooley et al [7] suggested the issues viz., new types of Knowledge, improved Mining Algorithms, incremental Web mining and Distributed Web mining, to be addressed.

IV. Open Source Data Mining Tools

There has been a lot of debate with respect to Open Source Softwares and Proprietary Softwares. Open Source is a development methodology which offers practical accessibility to a product's source (goods and knowledge), according to [2]. The term Open Source has been gaining popularity with the rise of the Internet, which provides access to diverse production models, communication paths, and interactive communities. How open is Open Source Really? There are "hidden costs" of Open Source (i.e. Red Hat Linux subscriptions and training can cost thousands of dollars).

When a vendor installs open source, it also has terms of service to stand behind and Free downloads do not. Users love the low cost of open source but are still nervous about O&M support services cost for the Open Source Software Products and Services. Support is still a problem for Open Source. Most of the closed-source vendors have passed the stage of rejection and denial of open source and, instead, have turned to open source as a key part of their software development strategies, drawing on its technical quality, low cost and favorable licensing terms.

Let me look into some of the major Open Source Data Mining Tools. The NCDM-2011 aims at utilisation of Open Source Data Mining Tools by Researchers, Academia and IT professionals :-

a) Weka [GPL] is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization [8]. "Data Mining Solution" is available as part of "Pentogo Business Intelligence Tool" and SpagoBI "Business Intelligence Tool". Weka is developed at the University of Waikato, New Zealand.

b) Orange (GPL) is a tool for data visualization and analysis for novice and experts; data mining through Visual programming or Python scripting; components for machine learning; extensions for bioinformatics and text mining; packed with features for data analytics [9];Maintained and developed at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia.

c) Rapid Miner (AGPL) is a fully integrated for data mining, predictive analytics, and business intelligence; rapid prototyping and beyond; ETL, OLAP, Predictive modeling, and reporting combined in a single product; numerous connectors for all common databases and data formats as well as unstructured data like text documents [10].

d) Rattle (the R Analytical Tool to Learn Easily) –GPL - is a data mining toolkit used to analyse very large collections of data. Rattle presents statistical and visual summaries of data, transforms data into forms that can be readily modeled, builds both unsupervised and supervised models from the data, presents the performance of models graphically, and scores new datasets. It is a new data mining application based on the open source and free statistical language R using the Gnome graphical interface [11].

e) TANAGRA - is free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area [12].

f) SIPINA - is a Data Mining Software which implements various supervised learning paradigms; Classification Tree Software (specialized on Classification Trees algorithms such as ID3, CHAID, C4.5, ASSISTANT-86, etc.); other supervised methods are also available (e.g. k-NN, Multilayer perceptron, Naive Bayes, etc.) [13].

g) ALPHA MINER is an open source data mining platform that provides the best cost-and-performance ratio for data mining applications viz., clustering, product association analysis, classification and prediction [14].

V. AGRICULTURAL INFORMATICS

Computers have been used in various institutions to simulate models and to solve statistical problems in the agricultural sector, since early 1970s. In 1980s, when NICNET was expanded to district level, DISNIC Programme was launched, which included DISNIC-AGRIS as one project component for developing agricultural information system in the district. I was then the Programme Director of DISNIC. In 1995, I was given the responsibility to build up of Informatics for Sustainable Agricultural Development. At that time, in the country, nobody was interested to talk on "mainstreaming ICT in Agriculture". I took up the challenge, as supported morally by the then Director General of NIC and with the help of Public Sector Banks such as SBI, Indian Bank, IOB, Oriental Bank of Commerce, NABARD, the Ministry of Agriculture and Department of Fertilisers, I organised the First National Conference on "Informatics for Sustainable Agricultural Development" (ISDA-95), in May 1995, at Vigyan Bhawan, New Delhi.

The ISDA-95 Conference had 16 sessions on various themes and prepared a roadmap for mainstreaming ICT in the Agricultural Sector, i.e. development of Digital Network for the Farmers (DNF), in the form of:

- AGRISNET-an Infrastructure network up to block level agricultural offices facilitating agricultural extension services and agribusiness activities to usher in rural prosperity
- AGMARKNET with a road map to network Agricultural produce wholesale markets and rural markets;
- ARISNET - Agricultural Research Information System Network;
- SeedNet- Seed Informatics Network;
- CoopNet-to network Agricultural Primary Credit Societies (PACS) and Agricultural Cooperative Marketing Societies to usher in ICT enabled services and rural transformation;
- HORTNET-Horticultural Informatics Network;
- FERTNET-Fertilisers (Chemical, Bio and Organic Manure) Informatics Network facilitating "Integrating Nutrient Management" at farm level;
- VISTARNET-Agricultural Extension Information System Network;
- PPIN-Plant Protection informatics Network
- APHNET-Animal production and Health Informatics Network networking about Animal Primary Health Centres;
- FISHNET-Fisheries Informatics Network;
- LISNET-Land Information System network linking all institutions involved in land and water management for agricultural productivity and production systems, which has now evolved as "Agricultural Resources Information system" project during the Tenth Plan being implemented through NIC.
- AFPINET-Agricultural & Food Processing Industries Informatics Network

- ARINET-Agricultural and Rural Industries Information System Network to strengthen Small , Micro & Medium Enterprises (SMEs)
- NDMNET-Natural Disaster Management Knowledge Network in India
- Weather NET-Weather Resource System Information Network of India

The ISDA-95 has also recommended for an allocation of 3-5 per cent of the Agricultural budget for ICT in Agricultural sector. Later, the Government of India earmarked 2-3 per cent of the Departmental Budget for ICT, which was the stepping stone for many central government departments for launching progressive steps to mainstreaming ICT in their Departments. There was a fillip in mainstreaming ICT in departmental functions. The Ministry of Agriculture, especially the Department of Agriculture & Cooperation, decided to mainstream ICT in the agricultural sector as a whole, based on the recommendations of ISDA-95. During the last 15 years, agricultural sector have been witnessing ICT in a very systematic manner. ICT projects such as AGRISNET, AGMARKNET, SeedNet, AGRISNET, PPIN, HortNET, APHNET, FISHNET etc., are operational in the country.

Under the National e-Governance Programme, Agriculture is one the Mission Mode Projects. I will discuss this aspect, later, in my address.

VI. Distress at Grassroots Level–How to Mitigate?

Let us look into the existing grassroots level problems. One such issue is as to how to prevent distress such as Farmer's suicide / affected villages in future? In 2008, the Department of Science & Technology (DST), Government of India, took keen interest that a systemic response should be operationalised to prevent distress such as farmer's suicides in suicide-prone / affected villages in future. Professor Anil Gupta of IIM (A) was asked to prepare a project proposal facilitating an interdisciplinary action research and implementation framework so that farmers' distress could be anticipated in advance, analyzed and responded in time to prevent any serious social disorder. Professor Anil Gupta has conducted two consultative meetings on 10th July 2008 at IIM (A) and on 29th October 2008 at IARI. I was invited to participate in both the deliberations. These consultations were to prepare a Detailed Project proposal (DPR) for Department of Science and Technology, on setting up of Village Knowledge Management Systems (VKMS), to prevent distress among farmers. These deliberations highlighted the need for a comprehensive Village Information System for all stakeholders of village development and prosperity. I have suggested including the national initiatives such as the DISNIC-PLAN project, AgRIS Project, GRID project and SMART VILLAGE project as the components of the proposed VKMS. They are relevant, and institutions have to come forward and institutionalize them in the districts coming under their jurisdiction. I will detail these components later.

We are aware that extreme distress among many disadvantaged farmers in different parts of the country had led to farmers' suicide / distress in about 43 districts in the country. There are many factors as to why farmers might take such unfortunate steps, which include lack of access to formal credit, inability to pay back moneylenders debt, crop failure, health emergencies or inability of a farmer to get out of downward spiral due to natural disasters or state or market failure. There are many policy alternatives through which, the distress among the farmers could be anticipated and alleviated to prevent such desperate outcomes. The consultation has detailed that one of the ideas is the possibility of linking Village Information Systems (VIS) and Land Information Systems (LIS) so as to generate indicators for preventing distress and identifying science and technology based solutions.

The insights from agriculture, eco-system management, health, education, livelihoods, land-use, risk management, etc at the grassroots level, should be understood. Studies which show connections between soil/land use, between crop/livestock and human use options, etc., are valuable. Lot of data exists at village level, which can be interpreted to generate early warning signals. National aspiration for better quality of life for the majority will not be fulfilled if systematic mitigation of farmers' problems does not take place. Sustainable alternatives for agriculture, livestock, non-farm and related rural knowledge based industries have to be urgently explored. Managing local knowledge and blending it with modern science and technology offers a pathway. This necessitates the need for a comprehensive Village level Information System on what is available on the ground and below the ground. The DISNIC-PLAN project is meant for this. But nobody is interested to implement in the country.

We have been reasoned to trust that "the Road for Nation's Development goes through Villages" i.e. bottom-up development. Since independence, it has been mainly top-down approach for development. This trickle-down approach did not impact at the grassroots level and there have been patches of development. The NREGP 2005 has targeted beneficiaries directly but assets creation, as aimed, at the grassroots level has not matched the expectation of the Government. This programme had a strong social audit component. The Poor represent the BOP (Bottom-of-the-Pyramid) markets and this calls for repackaging products to suit the pockets of the poor – or micro-selling in a mega economy [15].

India has undertaken measures such as globalisation, liberalization at macro level, and decentralization at grassroots level, to achieve vibrant economy, growth and development. Contemporary globalisation has encouraged the movement of people, capital, knowledge and ideas. New links, networks and partnerships have been formed between developed and developing regions, remote and favoured regions, urban and rural areas. These challenges and opportunities pose the following questions:-

- How can we understand these changes at grassroots level?
- What challenges do they present to theory, policy and practice?
- What are the opportunities for new thinking and action?
- What impacts do they have on poverty and inequality, and on related issues such as the environment, rural and urban livelihoods, corporate social responsibility, conflict and security, and HIV/AIDS?
- Are grassroots level people competent enough to sustain these changes (Farmer's Suicide)? Can they see opportunities (global markets for their products) in these challenges?
- What are the measures to be taken to make them aware and competent?
- Will traditional, tacit as well explicit knowledge, well proven practices and technologies get lost in this socio-economical churning?
- How to record and achieve the amalgamation of these knowledge and technologies with new technologies?

A SWOT Analysis is thus required for every village, to understand problems, challenges and opportunities Various Study Reports corroborate that the current state of various government departments, in terms of usage of ICT, is not in a "holistic manner" so as to achieve profound impact on ROI [in terms of people, process and knowledge]. e-Governance Roadmaps of many Government Departments, as of now, do not reflect the "pyramid upside down". G2G, G2B, G2C components of e-Governance Framework require "institutional approach", i.e. training, extension, development, education and research approach. It requires moving beyond "technology" component. Mainly ICT Infrastructure is being used for email, word processing, and in some cases process based applications (File tracking, scheme monitoring, public grievances monitoring, etc). Content Generation, Workflow applications, Decision Support Systems, Data Analysis, Framework based Web Services etc., have taken a back seat. What we require at grassroots level is C2G, C2B, C2C components of e-Governance Model. As of now, there is only G2C component. The C2G, C2B, C2C and G2C are the components of Village Knowledge Management System (VKMS).

Agriculture challenges in India are fourfold: need for enhanced production and productivity; need to address the issues of equity and uneven development; need to understand and address issues of sustainability and the last challenge is to enhance profitability in agriculture. A farmer needs to be linked with the agri-business systems, research institutions, public administration, other farmers, open market and other unlimited partners.

The keywords are precision, potential yield, desired quality and commensurate appropriate technology. Information & Communication Technology (ICT) has opened a new mode of technology dissemination including the areas with disadvantaged locations. The whole paradigm shift is summarized as production revolution to quality revolution, commodity to integrated commodity, mono-disciplinary to multi-disciplinary and inter-disciplinary, general technology to precision driven technology, production security to quality security and overall shift from "on-farm" employment to "off-farm" and "non-farm" employment. This requires a shift from "transfer of technology" to "development of demand-driven knowledge system" for espousing the cause of the farmers.

VII. Agricultural Resource Information System (AGRIS)

A. A Much Needed Strategy for Sustainable Rural Livelihood

Agriculture is highly dynamic in nature, because of the changing phenomenon of agricultural crops, which is further complicated by the interaction of crops with environment. Despite potential of economic and ecological benefits, adoption of precision technologies is very slow throughout the World. The reasons for limited implementation of site-specification management (or precision farming) in Asian Countries are due to: small land holdings, cost-benefit aspect, heterogeneity of cropping system, lack of local technical expertise, and knowledge and technological gaps. 1 In India, about 57.8 per cent of operational holdings have size less than hectare. Farmers, Land and Natural Resources (supported by the Land) have intrinsic and dynamic relationship. The site-specific management (or precision farming) is farm-size and production-system neutral, and will make agriculture "information intensive". This will impact rural economies. A stocktaking and diagnostic survey is needed early in the planning process to provide information about the wide range of factors, among the others, influencing agricultural performance:

TABLE 1

Agro-climatic data	Agro-economic data	Agro-forestry Resources
Animal Resources	Capital Resources	Crops and Cropping Systems
Environment data	Fisheries Resources	Forestry Resources
Infrastructure data	Institutional Resources	Land owners data
Plant Resources	Soil Resources	Socio-economic Data

Agricultural Resources Information System (AgRIS) is the Central Sector Scheme for strengthening / promoting Agricultural Information System in the Department of Agriculture & Cooperation (DAC) Ministry of Agriculture. This Project is based on the recommendation of the Report of the Core Group- V of the Standing Committee on Agriculture and Soils, National Natural Resources Management System (NNRMS) of the Planning Commission (March, 2000). During the Tenth Plan, the Department of Agriculture & Cooperation, in association with National Informatics Centre (NIC), has undertaken "Proof–of-Concepts-Projects" in districts across the country, facilitating the followings:

a) Initiate pilot projects on "Agricultural Resources information System (AgRIS) in districts in order to work out the cost and efficiency of this project and then expand to the entire country;
b) Develop a comprehensive database on various parameters related to land use, inputs (seeds, fertiliser, agricultural technology, agricultural credit) use, and water use;
c) Development of decision support systems (DSSs) packages for strengthening advisory services to farmers; and
d) Capacity building through Human Resources Development.

The various types of District typologies considered for the pilot project are:

TABLE 2

Coastal	**Dairy-farming**	**Dominated by cash crops**	**Dominated by forest economy**
Dominated by one or two urban centers	Mining/ industrial belt	Arid-zone	flood-prone but having vast wasteland
Dry -farming	Green revolution	Hilly	Socially backward
Tribal	--	---	----

The AgRIS Project is expected to emerge as the richest "agricultural resources information system" covering all aspects of agricultural, natural resource, and food systems, to: enable farmers to locate needed information to improve yields, plan for weather contingencies, access research, calculate treatments and runoff, simulate the growing season, visualize precision data, manage finances, buy inputs and sell outputs, and monitor prices in local as well as world markets. The Guiding Principles of designing AgRIS will be as follows:-

- Focus on the Disadvantaged Communities, who otherwise will be excluded;
- Provide that information or service which otherwise will not be provided;
- Focus on utilizing and where possible building upon what is existing rather than thrusting a new intervention;
- Create an outcome which in absence of ICT, will not be produced efficiently or timely; and
- Understand the difference between direct benefits and trickle-down benefits for the disadvantaged community.

Deliverables of AgRIS will include:

- Decision Support Systems (DSSs) on Production Practices and Systems
- Creation of Metadata to become the Country's initiative of "National Spatial Data Infrastructure (NSDI)" on Agriculture
- Guidelines on standardized methodology/best practices to be used for building Agricultural Resource Information System in similar districts of the Country
- DSS proposed under AgRIS will facilitate farmers in adopting agricultural production practices.

The AgRIS Project aims at, among the others, mainly the development and deployment of the following DSSs and customized to the chosen district typology for pilot project: -

Crop Suitability based on factor endowment;
Land Suitability Assessment;
Land Productivity Assessment;
Population Supporting Capacity;
Land Evaluation and Land Use Planning;
Land Degradation Risk Assessment;
Quantification of Land Resources Constraints;
Land Management;
Agro-ecological Characterization for Research and Planning;
Agricultural Technology Transfer;
Agricultural Inputs Recommendations;
Farming Systems Analysis and Development;
Environmental Impact Assessment;
Monitoring of Land Resources Development;
Livestock (cattle, buffalo, goat, & sheep) Farming Systems;
Water allocation in an irrigation system;
Fodder Resources Development;
Water Bodies (Basin) planning systems using Watershed and Agro-Eco Region Planning Concepts;

There is a "need to bridge theory and reality at grassroots". Few are prepared to move beyond secondary data. Farmers, in particular SMFs, must be able to choose agricultural practices and technologies

from the full range of approaches available for tackling their problems: agro-ecological methods, conventional research methods, and molecular biology research methods. Converting millions of poor farmers to the use of new technologies has been a colossal task. The implementation of 'Agricultural Resources Information System (AgRIS)' will facilitate development of typology specific agriculture development plan in the country. As "resources application and agronomic practices" are to match with soil attributes and crop requirements, the Agricultural Resources Information System (AgRIS) is a "way-forward" to improve agricultural productivity in rural areas, and a much "needed domestic strategy" for sustainable rural livelihoods (Figure-1 & 2). The AgRIS is "a step towards establishing a location-specific e-Government model for the Poor".

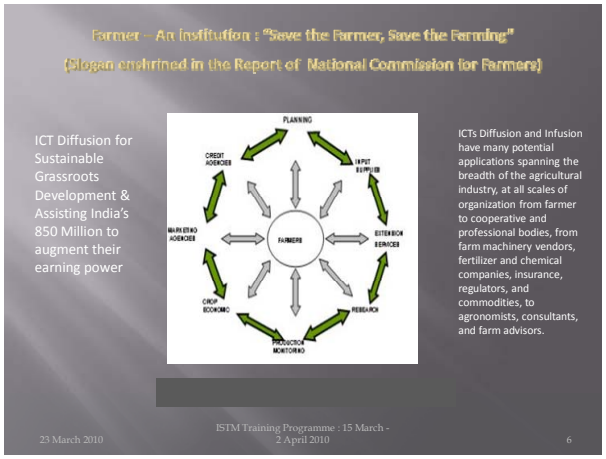


Fig. 1: Rural India's Stakeholding

Towards this objective, NIC has invited proposals from experienced Agriculture Management Organizations of repute i.e. Government Organizations, NGOs/ Voluntary Organizations, Individual Companies, having experience in the integration of ICTs into structure and function of rural economy, for submitting Expression of Interest (EOI) to provide comprehensive services for setting up Agricultural Resources Information System (AgRIS) for the district of Rohtak, Haryana. The deliverables of this Pilot project includes, among others, the followings:-

a) Development of AgRIS Portal, Unicode compliance, to provide one stop interface for all farming community services with institutional linkages etc.
b) Decision Support System on Production Practices and Systems for strengthening Advisory Services to facilitate farmers in adopting good agricultural production (GAP) practices.
c) Building Agricultural Resources Information System using Standardized methodology/ best practices.
d) Providing guidelines for establishing producers companies to maximize price realization by producer farmers
e) To deliberate on Potential Geographical Indicators (GI) and IPR including Traditional Knowledge for benefit to farming community for produces ,
f) Preparation of a technological architecture converging GIS, GPS, Web based solutions (SOA, Web services and software Engineering), Database, Expert Systems, Knowledge Base and other related technologies
g) Developing of a comprehensive spatial database on various parameters related to land use, input use (seed, fertilizer, agricultural technology, agricultural credit), water use, etc.; including creation of Metadata of Spatial Data Infrastructure.
h) Framework for collection/ validation/ analysis and input of spatial and non-spatial village level data from agro-economic sectors of importance; Localization of contents
i) Identification of intervention areas through generation of thematic maps;
j) Data sharing, dissemination and updation framework;
k) Reporting model and knowledge processing framework with every level of Decision Support System (DSS);
l) Application Software Development & Implementation;
m) Farmers Training System and Extension activities; Capacity building of stakeholders
n) Change Management Support and System Maintenance, Post implementation support
o) Critical Failure Analysis and Knowledge Warehouse

The AgRIS solution proposed should cover all aspects of agricultural and allied sectors including natural resources management to:

Enable farmers to access desired information for improving productivity, quality and profitability;

Plan for weather contingencies; Access research and technology database;

Suggest efficient and effective water and soil management strategies;

Simulate for alternative farm production and management strategies in view of weather forecasts (the growing season);

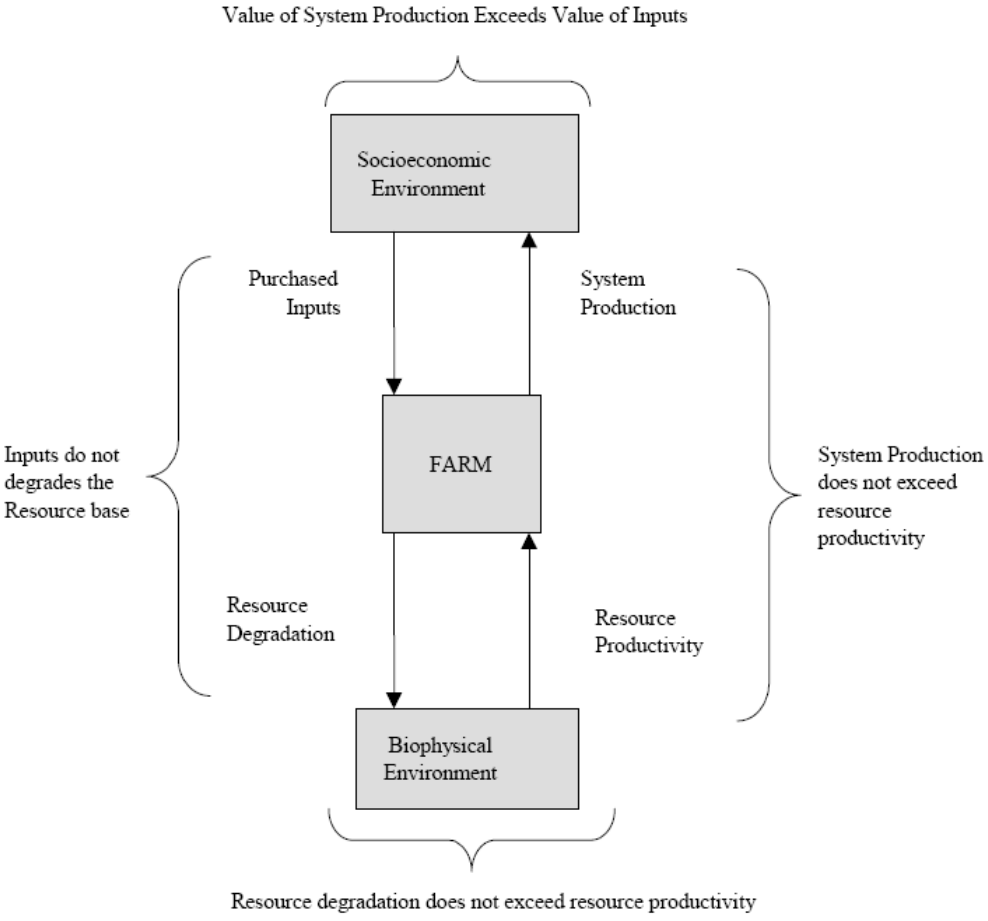
Visualize precision management strategies;

Advise on availability of financial resources and their judicious utilization;

Guide on input and output management and use ;

Monitor prices policy and fluctuations for timely advice to farmers;

The Nature of this Task may raise the requirement for the following among others:



Source: Proceedings of an International workshop on Sustainable Land Use Systems Research held at New Delhi on Feb. 12-16, 1990

Fig. 2: Sustainable Land Use System

Review of progress as per farm plan and natural resource base especially water, soil health and environmental degradation;

Project Details can be seen at the Website [16].

a) Preparation of a technological architecture converging GIS, GPS, Web based solutions (SOA, Web services and software Engineering), Database, Expert Systems, Knowledge Base and other related technologies;
b) Developing a comprehensive spatial database on various parameters related to land use, input use (seed, fertilizer, agricultural technology, agricultural credit), water use, etc.;
c) Localization of content
d) Framework for collection/ validation/ analysis and input of spatial and/or non-spatial village level data from all socio-economic sectors of importance;
e) Identification of intervention areas through generation of thematic maps;
f) Data sharing, dissemination and updation framework;
g) Reporting model and knowledge processing framework with every level of Decision Support System (DSS);
h) Application Software Development & Implementation;
i) Capacity building, Farmers Training System and Extension activities;
j) Change Management Support and System Maintenance
k) Post implementation support
l) Critical Failure Analysis
m) Knowledge Warehouse

Agricultural development is knowledge intensive and information intensive (both non-spatial and spatial). Decision Support System (DSS) on Agricultural Production requires information on the following:-

Information on physical feature [topography, geology, soils, natural vegetation, and hydrology (surface and sub-surface)] to determine the land's capability for agricultural development;

Maps depicting differences in physical land characteristics, meteorological, climatological, hydrological, geological, and geo-morphological conditions; population densities, types of land tenure systems used, proximity to markets and urban centres, transportation and other infrastructures;

Areas of immediate growth potential (where climate, soil and water conditions are favourable for agriculture and where technology needed to substantially increase output of major crops, already being grown, is available);

Areas of future growth potential (where favorable climatic and soil conditions exist but lack one or more elements of (i) adequate & controlled supply of water, (ii) technology required for substantially increasing production of a major crop or crops, currently grown, or capable of being growing, and (iii) transportation needed to bring the areas into national economy);

Areas of low growth potential (where climatological, soil, topological or other deficiencies without economic means for correcting them, exists) which require technological breakthroughs before substantial increases in output are possible.

I am very happy to realize that I have been instrumental in a big manner for realizing "agricultural informatics" way back in 1985-86 while establishing "DISNIC-AGRIS" Project under the DISNIC programme, for which I was the Founder Programme Director. Sustainable Development, prima facie, demands "natural resources management system", at grass roots level, to facilitate sustainable agricultural production in the country. The proposed National Mission on Sustainable Agriculture (NMSA) has recommended institutionalization of "Agricultural Resources Information System (AgRIS) and DISNIC-PLAN Programme in the country. When the DISNIC programme was launched in 1986-87 by the then Prime Minister, Shri Rajiv Gandhi, the programme was conceived to develop databases in 28 sectors to facilitate decentralised planning and administration in the country. "Database development" was given importance. This is still relevant and essential even now, for e-Governance programme. NIC has lost its interests in such programmes.

VIII. National Animal Disease Reporting System (NADRS)

The National Animal Disease Reporting System, in short NADRS, is a new Centrally Sponsored Scheme of the central Department of Animal Husbandry, Dairying and Fisheries, proposed for implementation during last three years of the 11th Five Year Plan.

India has a large animal population comprising, as per Livestock Census (2003), 485 million of livestock and a one-time count of 489 million poultry. Majority of the livestock, including poultry, are reared in rural areas where two-third of the people owns one or the other animal. These living assets contribute to the poor in a wide variety of ways, providing supplementary income and much needed nutrition for the family. Livestock also plays an important role in India's economy, contributing (along with fisheries) 5.21% to the country's GDP and 31.6% to the agriculture GDP in 2007-08. Their share in the GDP of the arid regions is as high as 70% and that of the semi-arid regions 40%. Progress of this sector results in balanced development of the rural economy, particularly in reducing poverty among weaker sections of the rural population. This is one sector where poor people contribute to growth directly instead of just benefiting from the growth generated elsewhere.

The livestock sector has immense potential. It has emerged as the key driver of agricultural growth in the country. The biggest impediment to growth of this sector, however, is the large-scale prevalence of diseases such as Foot and Mouth Disease (FMD), Haemorrhagic Septicaemia (HS), Black Quarter (BQ) in cattle, Enterotoxaemia, Peste des Petits Ruminants (PPR) & Sheep-Goat Pox in sheep and goats and Swine Fever in pigs, which drastically affect the productivity of animals. The presence of animal diseases also deters domestic and foreign investment in the livestock sector. These diseases not only wreck havoc on the existing stock but also constrain market access to our livestock sector, in spite of the fact that we have ample scope to participate in the global trade. It is projected that by the year 2020, over 60% of meat and 50% of milk will be produced in the developing countries. Within the developing countries, Asia will be the key production hub and India and China the primary producers of milk and meat. The country needs to gear itself for the opportunity.

The economic impact of the diseases in livestock results from both morbidity and mortality and the consequent production losses. This includes the direct losses due to mortality, reduced production in terms of milk, meat, wool, hide and skins, as well as indirect loss due to abortions, subsequent infertility, sterility, and deterioration of semen quality.

Controlling animal diseases is the best way to take rural poor out of poverty. By improving the productivity of animals on which people depend for their livelihood, it offers them a definite source of income. The pathway out of poverty involves improving the volume of the product marketed, and / or the quality of product, thereby increasing the revenue obtained. Access to this pathway is dependent on the control of diseases that either limit the movement of livestock or their products, or constrain the potential

purchasers investing in them due to their poor quality with respect to food safety.

At present, an animal disease is primarily recorded by the veterinary doctor working in a Government hospital / dispensary on the basis of clinical diagnosis. This information is passed on to the Taluka / Block level and then to the District and the State veterinary authorities. Disease information is also generated from the disease diagnostic laboratories at the District, State or regional level on the basis of laboratory diagnosis. Finally, information from State level is transmitted to the Central Government, i.e., the Department of Animal Husbandry, Dairying & Fisheries (DADF) in New Delhi. The DADF notifies the World Animal Health Organisation (OIE) and other international organizations, as appropriate.

The present system of animal disease reporting is not satisfactory for the following reasons:

- The disease reporting is neither timely nor complete. As a result of reliance on postal means of communication, the reports and returns take considerable time and some are also lost in transit. Hence, the compiled information does not represent true picture of the disease situation at any given point of time.
- The veterinary services available in the country are grossly inadequate. As a result, a large portion of the livestock owners do not have access to the Government veterinary services. These people rely on either the traditional systems of veterinary medicine or the private veterinary services. These incidences of animal diseases remain out of the reporting system. Their number is believed to be significant.
- In the prevailing situation, many times animal diseases assume serious proportion before control and containment steps can be initiated, thereby causing avoidable social and economic costs on the livestock owners and the country's economy.
- In order to bring about desired change to the existing situation, it is proposed to introduce a computerized system of animal disease reporting, linking each Taluka / Block, District and State Headquarters to a Central Disease Reporting and Monitoring Unit at the DADF in New Delhi.

The reporting system envisaged will enable the Block, District and State animal health officials to report the disease information and render reports and returns prescribed in this regard via internet. The system will be so designed as to assure secure data transfer and confidentiality of information. At the apex level, NADRS will compile and generate animal disease information for the country as a whole. The users will have access to the information as per permissions in consonance with their role and responsibilities envisaged under the system. This computerized system, proposed to be called 'National Animal Disease Reporting System' (in short NADRS), will enable fuller and timely reporting of the animal disease situation in the country, enabling its effective management.

The livestock diseases cause huge economic losses not only to their owners but to the economy at large. These losses are both of direct and indirect nature on account of the morbidity and the mortality in affected animals. While direct losses occur due to mortality in animals, indirect losses happen due to their reduced production of milk, meat, wool, hide and skins, abortions, subsequent infertility, sterility and deterioration in semen quality. Effective monitoring of diseases will enable their early control, prevention of their spread and reduction of economic losses caused by them. This will also help meet trade commitments related to a national surveillance system.

As a result of the information that would emerge from the NADRS, it would be possible to develop disease forecasting models, leading to development of disease prevention strategies. As the proposed scheme aims at effective monitoring the occurrence of livestock diseases with a view to enabling their early control, it will result in improving the livestock health in the country. By the very nature of the benefits that would accrue, these cannot be quantified in concrete terms. There is, however, no doubt that implementation of the scheme will yield immediate benefits to the livestock owners and to the economy by way of better health status of animals, prevention of losses due to their morbidity and mortality and improvement in the quality of their products. The benefits likely to accrue to livestock owners and to the economy may be summarized below:-

IX. Benefits to Livestock Owners

Better management of diseases of their livestock.

Availability of veterinary service.

Increased economic gain from higher productivity of animals.

Improved market acceptability of their livestock products.

Benefits to animal husbandry administration

Availability of a common channel for dissemination of animal disease information to all stakeholders.

Availability of SMS-based instant alert system for outbreak of diseases, spread of diseases, remedial measures and expert advice, enabling prompt control of diseases.

Availability of enhanced decision support system with GIS integration for effective and timely decision making.

X. BENEFITS TO ECONOMY

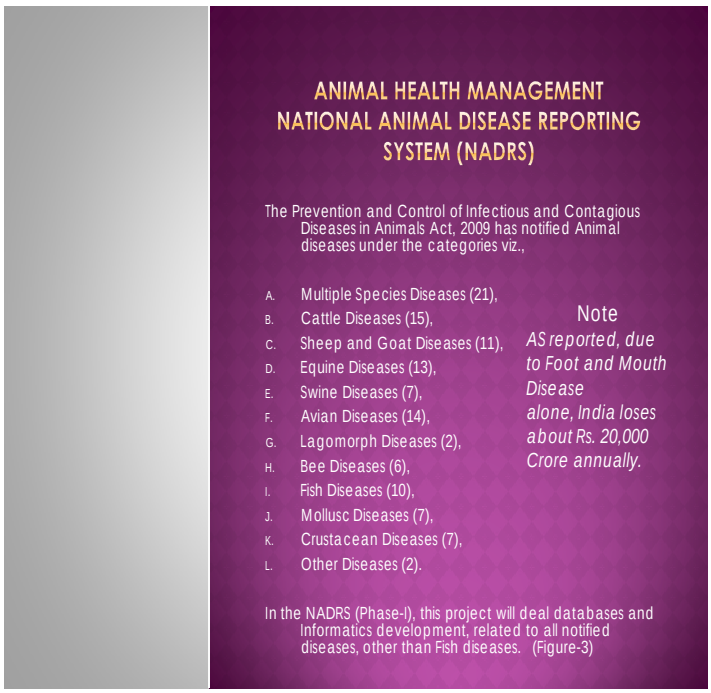


Fig. 2

An illustrative technical architecture diagram is shown below:

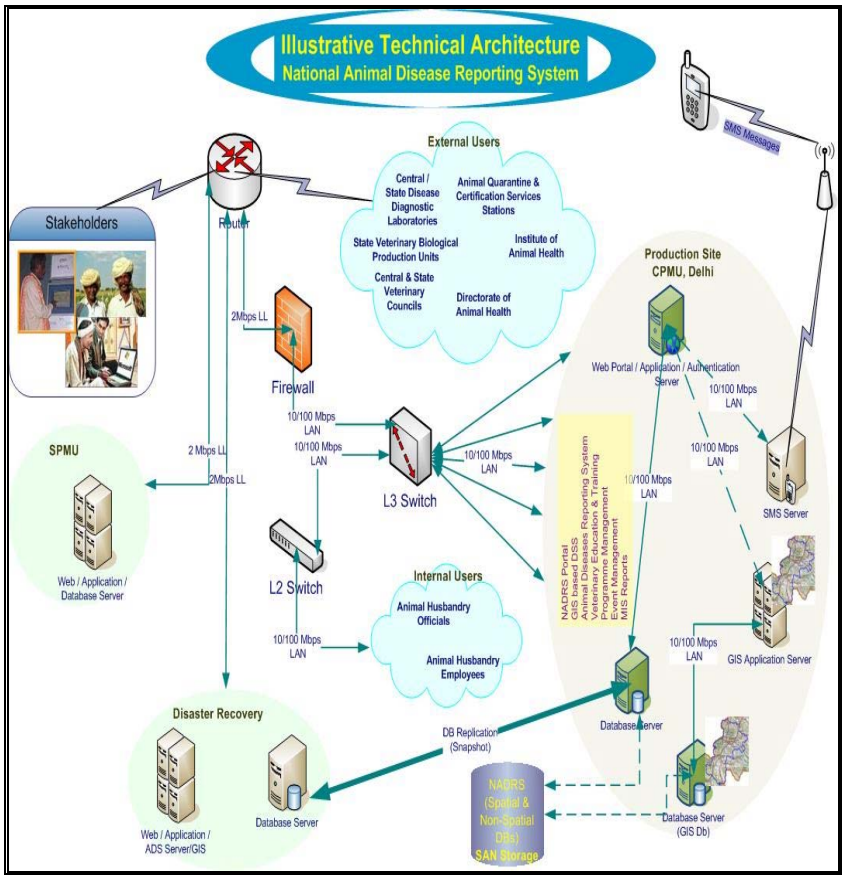


Fig. 3

Increased livestock production and productivity.

Improved market acceptability of domestic livestock products in international trade.

Saving of costs otherwise incurred for treatment of animals.

Fillip to the growth of the livestock sector, leading to increased employment generation and higher availability of animal protein to the population.

The NADRS will involve a computerized network, integrating both MIS and GIS, which would link each block, district and the State/UT headquarters in the country to the Central Disease Reporting & Monitoring Unit (CDRMU) in the DADF at New Delhi. All the notifiable diseases scheduled in the `The Prevention and Control of Infectious & Contagious Diseases in Animals Act 2009' (27 of 2009) will be included in the reporting system(Figure-3)).

The Disease Diagnostic Laboratories at the District, State and the National level will also be part of the computerized network. The veterinary colleges / universities will also form part of the NADRS.

XI. National e-Governance Programme Agricultural Mission Mode project

The Union Cabinet has approved the National e-Governance Programme (NeGP) with the cost of estimate of Rs. 23,000 Crores on 18th May 2006 and all measures are underway to accelerate the pace of implementation of its various components Under the National e-Governance programme, 27 Mission Mode Projects (Central, State and Central-cum-State), State Wide Area Networks (SWANs), One Lakh Common Services Centres (CSCs) which has now been re-christened as Bharat Nirman Common Services Centres (and scaled up to about 2.65 Lakhs in number) were undertaken to make the G2G, G2B, G2C components of e-Governance/e-Government Scheme operational in the country. The Status of these sub-components of the NeGP Schemes can be seen from the website [17].

Under the Agricultural Sector, Agricultural Mission Mode Project (MMP) has been announced. The following services are included /suggested:-

XII. Included

Service 1: Providing information on Quality Pesticides

Service 2: Providing information on Quality Fertilizers

Service 3: Providing information on Quality Seeds

Service 4: Providing information on Soil Health

Service 5: Providing information on Crop diseases

Service 6: Providing information on forecasted weather

Service 7: Providing market information on prices and arrivals of agricultural commodities

Service 8: Providing related market information to facilitate farmers get better prices

Service 9: Providing interaction platform for producers, buyers & transport service providers

Service 10: Providing information on minimum support price and government procurement points

Service 11: Providing electronic certification of imports and exports

Service 12: Providing information on Marketing Infrastructure and Post Harvest facilities

Service 13: Providing information on storage infrastructure

Service 14: Monitor the implementation of schemes / programs

Service 15: Providing training support to farm schools for adoption of good agricultural practices

Service 16: Sharing Good Agricultural Practices with farmers & trainers and providing extension support through online video

Service 17: Providing information on fishery inputs

Service 18: Providing information on irrigation infrastructure

XIII. Suggested New Services

Service 19: Service 19: Providing Information on Crops Development Programme and Production Technologies to increase Production and Productivity.

Service 20: Providing Information on farm Machineries & Implements

Service 21: Providing Information on Drought related aspects

Service 22: Providing Information on Livestock Development

Service-23: Providing Information on Financial Services available from PACS, RRBs and Public Sector Banks

Service-24: To provide information on financial security to persons engaged in Agriculture and Allied Activities through Insurance Products and other Support Services (Agricultural Insurance Services).

Service-25: To provide Information on Use of Plastics in Agriculture, Horticulture and Floriculture

Service -26: To provide information on Medicinal Plants.

Service-27: To provide information on Patent on traditional practices

Service 28: To provide information on Allied sectors like Sericulture, Floriculture, Horticulture, Bee-keeping

Service 29: To provide information to Farmers on Food Processing Technologies

Service 30: To provide information on agricultural wages;

To provide the above services, one can imagine the type of databases, applications software, Service Oriented Architecture (SOA), end-user empowerment tools, decision support systems, expert systems, knowledge bases, workflow, portal services etc required to put in place. Knowledge discovery and Data Mining will play a key role in making the services meaningful.

India has experienced during the last two months, the volatile nature of ONION prices in the country. The consumer price has gone up to Rs. 85 per kilo gram to Rs. 9 per kilo gram. Initially the Consumers cried and now the Producers are crying. The farmer might not have got more than 25 per cent of the consumer price as the farm gate price. Agricultural farmers and the Consumers have paid the price for this volatile situation, whereas the "middlemen" reaped the benefit. What went wrong? Is it the Supply-side problem?, the Demand-side problem?, or "the economy is getting over-heated", which we normally hear nowadays? How long do we have to take decision based empirical analysis? Forewarning and Forecasting should proceed further.

If India would have got one Agricultural TV Channel out of about 450 Channels, the stakeholders would have got benefited from the price analysis from all angles. There are about 300 agricultural commodities and each commodity has its own marketing channel. Price discovery of agricultural commodities is the need of the hour. Instead of political mining, India needs "data Mining" for Rural India to shine, smile and roar.

At the grassroots level, there is a question of production increase, productivity increase, income rise, employment opportunity looming large, even though National Rural Employment Guarantee Scheme (NREGS) has become the life line of the rural poor, since 2005.

XIV. Government Transformation and G2C Model of E-Governance Programme

I wish to highlight the issues impacting "Government Transformation", even though both the Central Government and various State Governments, NCT & UT Administration have allocated more than adequate budgetary provisions, year after year, since the pronouncement of e-Governance/e-Government programmes during this decade. We may look at the technologies which can usher in tremendous amount of ROI in e-Governance:

- Internet, Open Standards and Open Protocols: Needed for "Transformational" Government
- Identity and Access Management: to be Delivered as a Set of Standard Web Services
- eForms- Enabling Governments to Reorganize Data Collection Activities in the Government through ISO Standards
- Extensible Business Reporting Language (XBRL): A Specification That Exchange Financial Data across the Internet
- Cloud computing as enabler of Government Transformation: Challenges ahead
- PC (Personal Computers) Computing Trends
- Universal Serial Bus (USB)
- BI+MDA+BPM+SOA+EA leads to Better Service Delivery
- Enterprise Service Bus (ESB)
- Browser War, Rich Internet Application (RIA) & WebTop Application
- Application Performance Monitoring (APM)-Power of Network Forensics for e-Governance Application in India
- Internet Data Centers- Mainframe Architecture of Yester Years
- Information System Security-Internet as well as external threats
- Unified Threat Management (UTM) Solutions: Integrated Security Appliances & Locating Rogue wireless Access Points (AP)
- Inforamation Security Research & Training (ISRT): Need of the Hour
- A Rs 1500 Internet-Ready WIMAX Phone for Internet Access
- "Speech Recogntion"-the next most disruptive technology
- Beyoud Broadband Access: Data-Based Information Policy for a New Administration
- Web Use & Remote Workers: Managing the Risk

This proposed Transformation is possible only if the Academia adopts Departments and put these technologies in better perspective. The technologies like Cloud Computing and Virtulisation are technology components to utilize the ICT infrastructures, already created by the Central Government, State Governments and UT Administration. These technology components continue to get developed to strengthen the back-end ICT infrastructure and will facilitate ROI more overwhelming.

When we look at the implementation of G2C component of e-Government/e-Governance Scheme, it requires e-form technology adoption, workflow application, database development, essential decision

support systems (DSSs), and web services in 22 constitutionally recognised languages, if not in 1600 languages, being spoken in the Country. The Government Transformation and hence the e-Governance demands both "bottom-up" and 'top-down" process. The G2C component is for the "bottom of the Pyramid", to achieve social inclusion in the country. The theme of 11th Five year Plan has been outlined as "inclusive growth" and the 12th Five Year Plan is envisaged to embark upon "sustainable development for inclusive growth".

As most of the Government websites are in English language and more than 95 per cent of the Indian populations speak other languages, there is a mismatch for the "government transformation". This gap is widening, exponentially, in view of fast exploding social networking and services. Cloud computing is a cost-effective way to deliver innovative government services over the network. Virtualization is a computing technology that enables a single user to access multiple physical devices. The computing infrastructures include hard disk, development platform, database, computing power or complete software applications.

Virtualisation is an essential component of Cloud Computing. The widely used applications viz., Twitter, MySpace, Wikipedia, YouTube, faceBook, LinkedIn, Google docs and blogger are examples of Cloud computing. B2G (the inverse of G2B) and C2G (the inverse of G2C) services are the example to be based on Cloud Computing Model, whereas the G2G services can be based on Cloud Computing and Virtualisation model. Through Cloud computing, a world-class data center service and co-location provider is achievable in India.

The Web 3.0 Technology is now revolutionizing marketing and advertising, content distribution and customer engagement [18]. The "Web of tomorrow" will be about data and not document. Creating semantic web (Web X.0 where X>2) requires a different set of building blocks: protocols and standards. Universities and peer groups are creating "web ontologies" (i.e. association with the data will be the primary building block).

Many governments worldwide ate establishing one-stop portals to provide access to various public services based on the needs of citizens or businesses and not the internal structure of the government. A critical support for such one-stop portals is a workflow infrastructure, supporting the matching of the needs against provided services and coordination of the implementing processes, often spanning several government agencies. Olumide Oteniya et al (2007)[19] describe a generic workflow infrastructure for one-stop government – GovWF. This GovWF supports the operations of a Virtual Government Organization - a hierarchy of agencies providing collectively a set of public services, while offering a uniform one-agency view to its customers. One-Stop Portal should not be a "bug-stops-here-Portal". A Virtual Office which can be described as an integration of e-Form, Workflow, Decision Support Systems, Portals and Database, is the one that G2C Beneficiary (i.e. the Poor) will bank upon.

Let me conclude now.

As I reiterated earlier, the Agriculture (genetic modification), Medicine (genome research and bioinformatics) and Information & Communication Technologies (ICTs) are the three fields where diffusion of technology holds particular promise for the poor. The outcome of the NCDM-2011are expected to be translated into "actionable points" so as to undertake "Data Mining", "Text Mining" and "Web Mining" in the areas of agricultural informatics, bioinformatics and medical informatics in a dominant way in all relevant disciplines of research and education. It is required to work on algorithms, develop algorithms, improve algorithms, and products, and populate in Open Source as well as Closed Source. I have not seen any such open source products from IITs, IISc, IIITs and NITs, on the Internet. Let us emulate at least one such Weka in this country.

I thank the Organisers, Delegates, the AIMS and CSI for giving this opportunity to deliver the Keynote Address on the topic which is relevant now.

—Madaswamy Moni

REFERENCES

[1] http://www.microarrayworld.com
[2] http://www.en.wikipedia.org
[3] 8th International Workshop on Data Mining in Bioinformatics (BIOKDD '08), August 24-27 2008, Las Vegas, NV, USA;
[4] http://www.thearling.com/text/dmwhite/dmwhite.htm
[5] http://en.wikipedia.org/wiki/Text_mining
[6] http://www.nactem.ac.uk
[7] Robert Cooley, Bamshad Mobasher and Jaideep Srivastava : "Web Mining: Information and Pattern Discovery on the World Wide Web", Department of Computer Science University of Minnesota, Minneapolis, MN 55455, USA; {cooley, mobasher, srivasta} @cs.umn.edu; http://maya.cs.depaul.edu/~mobasher/webminer/survey/ survey.html.
[8] http://www.cs.waikato.ac.nz/~ml/weka
[9] http://www.ailab.si/orange
[10] http://rapid-i.com/content/view/181/190
[11] http://rattle.togaware.com
[12] http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
[13] http://eric.univ-lyon2.fr/~ricco/sipina.html
[14] http://www.eti.hku.hk/alphaminer
[15] Prahalad, C.K. (2006): The Fortune at the Bottom of the Pyramid: Eradicating Poverty Through Profits (Wharton School Publishing Paperbacks), http://www.amazon.com.
[16] http://www.agris.nic.in
[17] http://mit.gov.in.
[18] http://www.digitalmediaenclave.com
[19] Olumide Oteniya, Tomasz Janowski and Adegboyega Ojo (2007): "Government-Wide Workflow Infrastructure – Enabling Virtual Government Organizations", UNU-IIST Report No. 365, United Nations University, International Institute of Software Technology, April 2007.

An Analytical Model for Evaluating Public Moods Based on the Internet Comments

Chan Io Weng[1], Simon Fong[1] and Suash Deb[2]
[1]*Faculty of Science and Technology University of Macau Macau SAR, China*
[2]*Dept. of Computer Sc. & Engineering C.V. Raman College of Engineering Bidyanagar, Mahura, Janla Bhubaneswar–752054, Orissa, India*
E-mail: macau.oliver@gmail.com, ccfong@umac.mo suashdeb@gmail.com

***Abstract*—It has become a prevalent lifestyle nowadays that netizens voice their opinions on social networks (Web 2.0), for matters of all sizes, and on a regular basis. The opinions which initially should be intended for their groups of friends propagate to all public users. This pond of opinions in the forms of forum posts, messages written on micro-blogs, Twitter and Facebook, are largely contributed by communities of online users (or sometimes bloggers). The messages though might seem to be trivial when each of them is viewed singularly, the converged sum of them serves as a potentially useful source of information to be analysed. A government of a city, for instance, may be interested to know the response of the citizens after a new policy is announced, from their voices collected from the Internet. However, such online messages are unstructured in nature, their contexts vary greatly, and that poses a tremendous difficulty in correctly interpreting them. In this paper we propose an innovative analytical model that evaluates such messages by representing them in different moods. The model comprises of several data analytics such as cultural moods analyzer implemented by neural networks, text mining and hierarchical visualization that reflects public moods over a large population of Internet comments.**

Keywords: cultural moods engines; Event driven artificial neural network models; hierarchical visualization

I. INTRODUCTION

Netizens nowadays develop a habit of whining out their opinions in the Internet world, through blogs, social networks as well as community forums. Their purpose may just to share their views, both casually and deliberately, in response to all kinds of world events or topics of interest. From the postings and counter-replies, it has evolved into a trend of social acquaintance in the virtual world [10]. Twitter has more than 180 million unique visitors per month, and a total amount of messages close to a trillion. Facebook also has a population of 166 millions active users whose posts amount to a similar astronomical figure. And these figures are still undergoing some phenomenal growth.

Recently government agencies established their community groups on Facebook. The motive could be in two fold: to disseminate information to the online users, and to probably listen to their opinions. However, to the second motive, assuming that the government agency bothers to pay attention to the opinions, there is an inherent challenge in the format of the data. They are unstructured both in grammar and context. As users are free to post anything under the sun, the format is not in formal writing (unlike official letters); slangs may be used and they differ from culture to culture. On the brighter side, netizens are responsive to new posts and new events. For example, any world news, such as earthquake, terrorist attacks or economic crisis would attract them to post and encounter post on each other's messages. They share their views in different emotions, pertaining to the subject that they are commenting about. The messages come in very different types of wish-making, suggestion, political opinion or dissatisfaction to share with friends and the rest of the world in a cyber-world of social network.

In addition to the obstacles of data formats and contexts, a government or organization may face another challenge due to the dynamic nature of the distributed Internet comments, which arise both in tremendous quantity and at a very high speed. The contents of the comments may change over time too; for example, an invention of a vaccine for a global epidemic disease may first be cheered as a "happy" news. Should it be later found as a hoax, the general comments may gradually switch to mood of "disappointed" or even "debate".

Organizations do need some autonomous method to classify the messages into different moods and kinds of opinions, in contrast of the previous works of deciphering their actually meanings. Currently manual work is required by a human user to comprehend the messages by his knowledge background and relate them as opinions being talked about of a particular event. Because there are large diversities of words and vocabularies representing different emotions, it is important to tap on the cultural background knowledge.

Emotion is a complex psycho physiological experience of an individual's state of mind as interacting with biochemical and environmental influences. In humans, emotion fundamentally involves "physiological arousal, expressive behaviors, and conscious experience" [1]. Emotion is associated with

mood, temperament, personality and disposition, and motivation. Emotion doesn't exist in computers that are based on logics. Emotion also may not be easily calculated by a formula. Emotion is a fuzzy state, hence artificial neural network (ANN) is well-known for handling this type of problems [9]. Before entering the comments in to an ANN, we need to translate sentences to relative metadata which are represented in abstract levels. To translate sentences to metadata we make use of a linguistic dictionary to categorize and group the words into meta-data.

"Point of view" is another important factors that contribute to understanding emotions. We utilize information from online newspapers to establish a neutral evaluation standard. Since opinions in newspapers are in journalistic and supposedly objective style, we adopt so as a standard for describing neutral opinions. The other usage of newspaper is it may contain the background story of an event. Comparing the evaluation standard and Internet comments we can have some benchmark for positioning a neutral point in our visualization which shows information in different levels, thus we call it "Hierarchical Visualization". The Hierarchical Visualization can provide the trend of public mood, detail of the range about mood and it can be directly used for government or organization to understand their citizens' or customer's feedbacks. The Hierarchical Visualization reveals the moods of the public in general, instead of displaying a long list of individual comments. Our proposed Hierarchical Visualization is designed for high-level users who often prefer to glimpse at an overall view of the public opinions, without going into details. This method proposed in this paper is subtle which doesn't require massive scale of survey questionnaires to probe answers directly from citizens.

II. Our Proposed Model

Before collecting information from the Internet, we should define the Research Topic (or topic of interest) to be analyzed. Research topics could be any current affair or the latest government policy which netizens are keen to comment about. The next step is to download the data from relevant sources. The easiest date of the event that happened can be referred from the news. News published on the Internet usually would have highlighted by some keywords that can be extracted from part of the URL link – the keywords are useful for us to define the metadata of a research topic. Overall, for each research topic, we use the time, the metadata, as settings of parameters for the web downloading software to congregate Internet comments with a reasonable time range (e.g. 80% of netizens talked about Michael Jackson's death within only 5 months). The information downloaded will be used to build up two kinds of databases. One is a repository of online News that are tagged with easiest date of occurrence, plus the related metadata for ontology [8]; the second one is the postings extracted from some social networks and micro-blogging sites. Twitter and Facebook are used as experiments in this paper. HTML tags are cleansed in the preprocessing step. The information about the poster's information, such as IP (which may not always be available), time of posting, user's background or other will also be collated.

The constitution of Moods engine was based on a standard dictionary for embracing the keywords by using an ANN. We select all the words that are related to emotions from the dictionary and use those words as training data to build up a number of ANNs. Optionally, one may incorporate MSN-style of acronyms or emoticons to represent emotions [6], e.g. a smiley is a symbol of happiness written as :-) Short-names commonly used as cyber etiquettes like W.T.H. (anger plus astonishment) and I.M.H.O (neutral narration) could also be added on. One ANN is to be trained and employed to describe one type of mood. This collection of ANNs make up of the mood engine. The mood engine is to be fine-tuned with users' subjective experiences for improving the accuracy. So the major function of the mood engine is to distinguish a resultant mood by reading through a pile of text messages.

The evaluated news will also train the ANN of different moods. Since this method is Event driven based, the new training dataset will be no longer be used for another event. The ANN of a particular mood is represented by the weights trained by the training data. If a piece of news was marked or flagged as "happiness", the news would be used to train the "happiness" ANN model. Data from the Internet comments databases would be used as testing data to be tested in the ANNs for deciding which mood(s) they belong to.

The judgment of internet comments will accumulate to a dataset. The comments should be cleaned or restructured before inputting to moods engine, the detail will be discussed later. The output dataset will be used in presentation engine and displayed in the Hierarchical Visualizations. Multiple levels of visualizations are used because the details of the results could be shown in different depth, depending on the choice of the user for the desired resolution. Too much information in visualization is confusing to the users. Users can opt to choose a viewing level interactively in the control panel of the visualization software.

The results are shown in graphical form so that it is easy to captivate the attention of the user, and possibly to spot any special patterns visually. The user can zoom in and out at will, or to display the full details for further analysis when necessary.

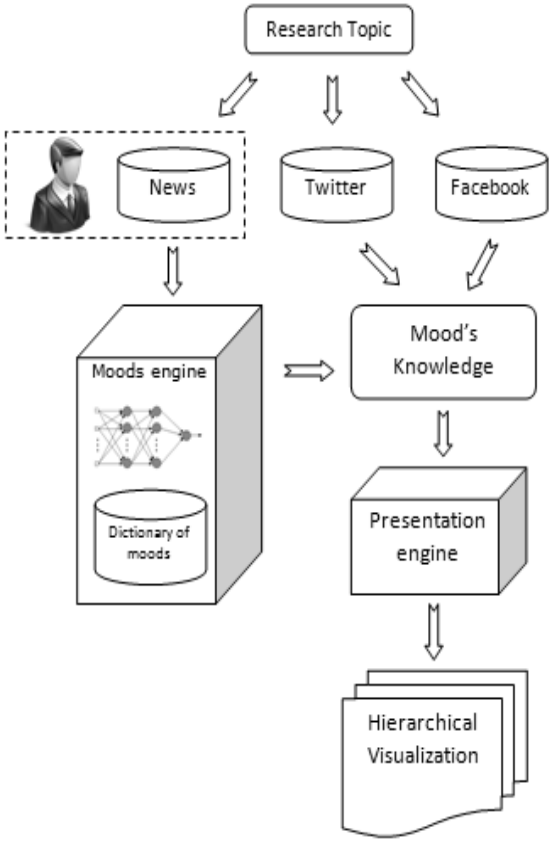


Fig. 1: A General Model for Processing Public Emotions

III. Processing Mechanism

The basic mechanism of moods engine is by using dictionaries to build Mood's artificial neural network models. Each model is only used for one mood detection. We should find out all the words to describe moods manually or from the list in Wiki [5]. We can get some reference of moods from Linguistics and psychology books. There exist a lot of examples to describe mood of what people will say or do. There are synonyms, usage examples and web definitions. The synonyms can link the moods together and build up a mood's net. The extraction of the usage examples and web definitions in dictionary inputted to artificial neural network and the word of moods to be the artificial neural network ID, the output will be a flag determining the mood. The flag is a real number between 0 and 1. We define a clear function which uses rule of English grammar to simplify the sentences. The notation of clear function is $cf(S, L)$ where L is the level of simplification. Suppose we have an article A, A's sentences is in a set $\{S_1, S_2 \ldots S_n\}$. $S_1' = \text{cf}(S_1, L)$, $S_2' = cf(S_2, L) \ldots S_n' = cf(S_n, L)$. All the sentences in A' should be simplified sentences. A simplified sentence contains a single subject, a verb and a predicate. It describes a single thing, an idea or a question, and has only one verb. An example of a dataset to represent happiness is shown in the following table:

Table 1: Sample dataset That Represents Happiness

Subject	Verb	Predicate	Happiness
I	like	holiday	Yes
Mary	lost	100 dollars	No
100 dollars	found		Yes
The weather	is	fine	Yes

The dataset should be processed further. The subject and predicate can be categorized. It is very important to make ANN precisely to detect the mood of a sentence with the concise information (and to suppress noise). The categorizations make the sentences has more compact since we do not care about the details. E.g, 100 dollars is categorized into Money and Mary is She, by using The bag-of-words model that is a simplifying assumption used in natural language processing and information retrieval. The dataset is condensed to items that include a subject, a verb and a predicate to be used as input to the ANN, and a binary verdict of output - happiness or otherwise.

Table 2: Sample of catagorized dataset

Subject	Verb	Predicate	Happiness
I	like	lesisure	Yes
She	lost	money	No
Money	found		Yes
Weather	is	fine	Yes

The accuracy of mood's artificial neural networks model depends on the number of records in the training dataset and the categorization of subject and predicate. Detection of moods can be defined as either positive or negative. There can exist some opposite moods, such as happiness and sadness, or even a mix of moods.

By using words in dictionary and their synonyms, we can build Mood's Artificial Neural Network Models (MANNM) to a net, and according to Synonyms we calculate the weight of connections between MANNMs. The weight is meansured by the similarity between synonyms as follow:

$$\text{weight}(A, B) = \frac{\text{Total number of B's Synonyms i.e. same as A}}{\text{Total number of A's Synonyms}}$$

According to the weights of each artificial neural network models that are in the form of a net of relationship, every sentence can be tested in the MANNM's net. The mood can be simply classified in positive mood ornegative mood. Testing the sentences alternatively in positive and negative can speed up the mood detection. E.g, testing A' mood, each sentence must be inputted to a Mood's artificial neural network model, if output is 0 then go to reverse site of mood, from one node goes to the other node and find out which is its Mood belong to. Finally we sum every sentence with the entire mood and calculate the percentage of mood of the article A. E.g, A has 100 sentences, 78 sentences are classified as happiness, 15 as general mood and 17 as surprised. We then deem that A contains 78% happiness, 15% general mood and 17% surprised.

The second important element of mood engine is the mood standard. The system may choose one article from the earliest collection of news. That article will let researcher to read and feedback the mood of the article as expert advice. The researcher decides that it belongs to which mood that is research standard mood to understand the event. Based on the mood's artificial neural networks model net which is generated by the dictionary, here we use the article that we chose and defined the sentences to which mood. Those dataset will add to the training dataset and rebuild the mood artificial neural network model. Finally we have a set of standard mood's artificial neural network models net to evaluate internet comments about a specific event. The theory of this method is the new dataset of standard mood will let the artificial neural network to learn more and the News words about the events will be considered by the ANN models. The net can understand what the events are talking about.

Now we have the engines to analyze the moods but the result is in a lengthy list that is not so readable to human users, because typically it may produce approximately 10000 records for a research topic. Then we need to develop a new kind of interactive visualization to present the result for researcher after the moods engine finished its processing. The Hierarchical visualization is used to show the moods in colors and distributions graphically. According to the weight of similarity it should classify the moods in serial groups. Each group has similar feelings. In this paper we experimented on Facebook and Twitter comments and built a dataset to store all moods and internet comments with their information, such as IP address, user information that include gender, age, education background, etc.

The follow example in our experiment is on the topic of "Hengqin Campus Project of University of Macau". We set up a list of colors [2] to represent different moods. The circle represents moods of an event and we use the angles represent the percentage of the moods. In Figure 2, it is easy to get to know the public mood about an event based on the comments collected from the Internet. In the software, we can scale to level 2 of the hierarchical visualization as shown in Figure 3 when we select the two colors in the circle. Since the colors are not fixed, the visualization is interactive with researcher. Researcher can focus on which moods that the general public feels about the specific topic.

In this zoomed in level of visualization we may want to analysis about genders of the users who posted their opinions (and subsequently reflected by their moods), just for example. The wave line in Figure 3 is representing the number of people is different moods with different gender, one on each side of the belt. In the next level, we can select a location to be analyzed. The locations, gender and moods relationships are presented in this level. In Figure 4, the chart shows visually that how users in different locations carry certain moods. The level of details can be increased optionally; for example, the locations can further break down to streets. Other dimensions can be added or switched too.

Fig. 2: Level 1 of the Visualization

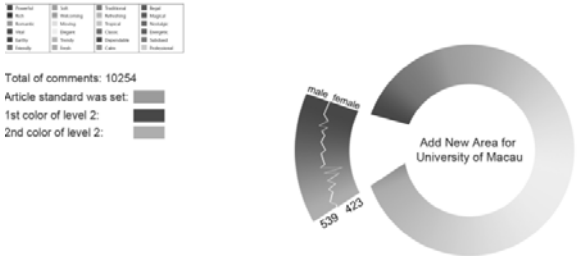


Fig. 3: Level 2 of the Visualization

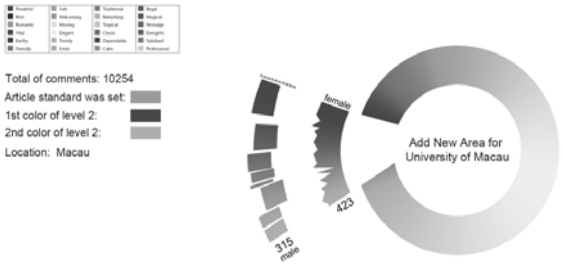


Fig. 4: Level 3 of the Visualization

IV. Discussion on the Outcomes

The proposed system can help organizations or government to understand the opinions which are in response to a news event or a policy announcement, without doing meticulous survey to collect public opinions. The System is easy enough to use, for revealing the public moods based on a given event. The system features about consideration of a culture of a country. Cultural practices influence on the way the citizens say and post. For example, the word "cool" may mean a cheerful mood in the Western culture, but otherwise in oriental or other conservative cultures who take the word "cool" literally as "indifferent". However, different versions of the system may be needed to be built for different cultures [7], but the training sets and hence the ANNs would be unique in each culture.

As such, an analyst who uses the system by different cultures can understand where, who, which age group, how many people and what they feel in the visualization, in response to an event, based on the analysis from the Internet comments.

V. Related Works

There is a similar project named "Twitter mood maps reveal emotional states of America" in America. It has an idea to present human mood in a timeline superimposed on an America map with different colors. This method takes individual words out of context. If someone tweets "I am not happy", the team's method counts the tweet as positive because of the word "happy". It is based on current state of Twitter user and by possibly keyword matching methods. Unlike our proposed model, it does not shows results from multi-dimensions, and the fuzzy natures of mood matching and different cultural aspects were not considered.

Another academic research work that is very similar to ours is [3], by the authors Tao et al. Tao regarded that the Web has become an excellent source for gathering and realizing public voice. The paper discusses a method for exploring the public mood levels at the time of posting. Hence the results are presented as two-dimensional graphs with the y-axis being the mood level, and the x-axis as a time-line. Again, the two dimensions of variables for presenting mood levels can be extended to multiple dimensions as proposed in this year. Also the paper [3] used corpus aggregating method for measuring mood level, and the case study was on emergency scenarios, where ANNs are used in our model for handling fuzzy situations.

In addition, our model incorporates with a hierarchical visualization program, and our experiments showed that the prototype can be used in many general situation.

VI. Conclusion

In this paper we proposed and defined an analytical model for evaluating Internet users' comments in response to a given event. Our model features a Mood engine made up of a number of artificial neutral networks, one for each type of mood, and different sets of ANNs for different cultures. The ANNs are trained by using standard words that describe the corresponding moods taken from dictionaries as well as online news websites that supposedly report events in objectively correct moods (unbiased). The trained networks are used to detect types of moods and their intensities from a pool of messages and postings collected from micro-blogs and social networks that constitute largely online communities. The Internet comments are inputted to the mood engine and the comments are categorized in types of moods. Aggregating categorized comments and their moods are fed into a hierarchical visualization program that shows interactively different dimensions of the information with respect to the public Mood. A prototype is built and results show that the model is feasible.

.

References

[1] Myers, David G. (2004) "Theories of Emotion." Psychology: Seventh Edition, New York, NY: Worth Publishers, p. 500 (Wiki)

[2] Chuan-Kai Yang, Li-Kai Peng, "Automatic Mood-Transferring between Color Images", IEEE ComputerGraphics, Issue No.2, March/April 2008, pp.52–61.

[3] Tao Xu, Qinke Peng, Chengwei Li, "A Method of Capturing the Public Mood Levels in Emergency Based on Internet Comments", 7th World Congress on Intelligent Control and Automation, WCICA 2008, 25–27 June 2008, pp.3496–3499.

[4] Zhihang Chen, Chengwen Ni, Murphey, Y.L., "Neural Network Approaches for Text Document Categorization", International Joint Conference on Neural Networks, IJCNN 2006, pp.1054–1060.

[5] Wiki Emotion list: http://en.wikipedia.org/wiki/Emotion

[6] Yamashita, Ryo; Yamaguchi, Sanae; Takami, Kazumasa, "A Method of Inferring the Preferences and Mood of Mobile Phone Users by Analyzing Pictograms and Emoticons Used in their Emails", Third International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services (CENTRIC), 2010, pp.67–72

[7] Zhijun Zhao; Lingyun Xie; Jing Liu; Wen Wu, "The analysis of mood taxonomy comparision between chinese and western music", 2nd International Conference on Signal Processing Systems (ICSPS), 2010, pp.606–610.

[8] Seheon Song; Minkoo Kim; Seungmin Rho; Eenjun Hwang, "Music Ontology for Mood and Situation Reasoning to Support Music Retrieval and Recommendation", Third International Conference on Digital Society, ICDS 2009, pp.304–309

[9] Elaleem, O.A.; Elragal, H.M.; Shehata, H.M, "Voice Message Priorities Using Fuzzy Mood Identifier", Proceedings of the Twenty Third National Radio Science Conference, NRSC 2006, pp.1–6.

[10] Tanase, S., "When web 2.0 sneezes ...everyone gets sick", Engineering & Technology, Volume 5, Issue 5, 27 March–23 April 2010, pp.28–29

Advanced Techniques for Regression and Classification in Mining of Biomedical Data

M. Abbas, Mukesh Srivastava and Mohammad Imran Siddiqi
Biometry and Statistics Division
Computationa lBiology & Bioinformatics Lab, Division of Molecular and Structural Biology
Central drug Research Institute, Lucknow
E-mail: abbas@cdri.res.in, mukeshlko@yahoo.com

***Abstract*—In biomedical research there are situations where the independent variables [X] may be highly correlated, the number of variables are greater than the number of observations and number of dependent variables may be more than one. In such situations, classical methods such as Multiple Linear Regression and Linear Discriminant Analysis are not feasible. For such situations, Partial Least Squares [PLS] and Artificial Neural Network [ANN] are very efficient methods. Concepts behind the PLS, the algorithmic variants of PLS, network topology and learning rule associated with Feed Forward Back Propagation [FFBP] are presented. Methods are illustrated by two examples from the literature.**

Keywords: Regression, Classification, Partial least squares, Feed-Forward-Back-Propagation, QSAR, MS/MS Spectra

I. INTRUDUCTION

Data, like raw material, is a valuable resource. Managing data effectively and efficiently is very challenging. Database technology provides ease in organizing and using the data. Relational database has usually been used to create operational data and Online Transaction processing [OLTP] system .Next important development was the concept of data warehouse and its applications in business management. Thus the field of Business intelligence was developed. Important functions of business intelligence technologies are, online analytical processing, analytics, data mining, , text mining, and predictive analytics. A Data warehouse is supposed to be a repository of all types of data from different sources concerning any organization .Warehouses are characterized by large bulks of data coming from disparate systems. Thus warehouses are centralized in nature. At times, it may be very difficult task to get the relevant data needed for statistical analysis as well as trend and pattern spotting. Data mining provides tools to overcome this difficulty.

For a clear understanding of data mining concept, let us take some cues from coal mining industry. In coal mining industry terms like prospecting, exploration, mine development and extraction are used. Prospecting is the actual search for coal where as exploration is to assess the viability of mining operation. Mine development is to prepare the site for mining, and actual mining is to remove the coal from the developed mine. Like-wise, we start with a research question in data-mining. Prospecting is to discover relevant data sources to get the answer. Data Exploration is to see whether a data base is minable, then we go for convergence of databases and data integration. After this we get required data from the database and finally we go for analytics. In practice, one does not differentiate these steps clearly but in principle they are being followed. In modern time, data mining also includes visualization and analysis of medium and large datasets independent of any database

Bio-medical research has become a data-rich field and needs applications of data-mining techniques. Section II presents scope of biomedical data. Sections III and IV deal with statistical and machine learning methods, respectively. PLS methodology is presented in Section V, whereas Section VI is for FFBP. Section VII presents two illustrations.

II. BIO-MEDICAL RESEARCH: A DATA-RICH FIELD

Biomedical field, be it in the basic research in academia or governmental laboratories or its applications in diagnosis and pharmaceutical industry, has become an information driven science. An enormous amount of data is either present in published literature and patents or being generated using high-throughput technologies. Also patient-related data is being generated at a rapid rate. The resulting computational biology challenges currently involve the development of databases that facilitate the hosting and integration of these diverse data-sets and the development of intelligent data mining methodologies for these high-dimensional data

A. Laboratory Data

Recent advances in experimental technologies are generating vast amount of data from complex biological systems for biological discoveries in science .These new and exciting data include genomes of a variety of model organisms[1],

genetic variations of individuals[2], gene expressions[3], high-throughput proteomic[4] & metabolomics data[5], gene ontology[6] and pathway databases[7].

B. *Patient-related Data*

Recent advances in sensor technologies have enabled long term recordings of numerous physiologic parameters in patients, generating very large data sets. This phenomena extends to implantable sensors as well as non invasively applied external sensors: electroencephalographic [EEG] recordings [8], hyper spectral imaging for tongue fissures [9], for blood vessel determination and artery-vein differentiation [10] and diabetic foot ulceration[11].

In contemporary world, all practical data is multivariate in nature. Essentially, more than one variable is recorded for a number of objects or experimental units. Variables may be independent [predictor] or dependent [response] types .Numbers of observations, independent variables and dependent variables are denoted by n ,p and q respectively. As for analytics, data-mining approaches may be divided into two types: Classical where statistics plays a central role and modern wherein machine learning approach is used.

III. Classical Approaches

Techniques from multivariate statistics play an important role in data mining .These techniques may be divided in two major classes

A. *Regression & Classification*

Regression analysis includes any techniques for modeling when the focus is on the relationship between independent and dependent variables. Ordinary least squares regression is parametric method of establishing mathematical relationship. Classification is a special case of regression when responses are classified according to some criteria. Fisher's linear Discriminant Function analysis, Logistic Regression and Naive Bayes Classifier are important techniques for linear classification.

B. *Clustering*

Clustering is disintegration of a set of observations into subsets so that observations in the same subset are similar in some sense. It is used in order to get some insight of the data, usually applied in the beginning of data mining operation.

IV. Machine Learning Approaches

Machine learning is a sub-field of Artificial Intelligence [AI] concerning with the design and development of algorithms that allow computers to spot pattern and trend in large data. A major focus of machine learning research is to automatically learn and recognize complex patterns. One of the earliest methods was decision tree learning. This learning is a decision tree [12] as a predictive model which maps observations to response. However, the most general approach is Artificial Neural Network [ANN] learning. An ANN [13] uses a connectionist approach to computation. We can have different types of network topology and accordingly learning rules. ANN extends regression and clustering methods to non-linear multivariate models. Other approaches include Association–rule learning [14] and Support Vector Machine [15]. In machine learning, the learning may be of two types:

A. *Unsupervised learning*

One form of unsupervised learning is clustering. Among neural network methods, the Self-organizing Map [16] and Adaptive Resonance Theory [17] are commonly used. Fuzzy ART [18] has recently been introduced.

B. *Supervised Learning*

A supervised learning algorithm analyzes the training data and produces a trained model. The trained model should predict the correct output value for any valid input object. This requires learning algorithms to generalize from the training data to test data.

Having noted important methods in statistics and machine learning, we shall, now, focus on PLS regression and ANN Classification

V. Partial Least Squares

A. *When PLS should be used?*

When the independent variables [X] are highly correlated or the number of variables in X [p] > the number of observations [n] or number of dependent variables is more than one**,** PLS may be used. With advancing of modern technologies, high-dimensional data are prevailing in computational biology where such conditions are met.

B. *Basis*

When Y is a vector and X is a full-rank matrix, then OLS is appropriate. When p>n, X is likely to be singular and MLR is no longer feasible. One approach could be to eliminate some predictors. Another is to perform PCA of the X matrix and then use the PCs of X as regressors on Y. Nothing guarantees that the PCs which explain X, are relevant for Y. When structures of X and Y are complex, even PCR does not give reliable results. PLS generalizes and combines features from PCA and MLR. PLS is still a highly active research area.

In PLS, the row vector of variable means α , the score vector t and the loading vector p are used to model X

$X = \alpha + tp + E$

and Y is computed as;

$Y = btc + F$

Where c is weight matrix for Y and 'b' is the regression coefficient between t and u, to connect X and Y, E is the matrix of residuals in X, F is the matrix of residuals in Y. It can be seen that the latent variable t of X is the common part of the two matrices, X and Y. Computationally, the goal is to find two sets of weights W and C in order to create a linear combination of the columns of X and Y. t= X.W and u= Y.C. The latent vectors t and u could be chosen in a lot of different ways.

PLS methodology is to relate two co-ordinate systems **X** and **Y** via score vectors **t** and **u** respectively. By fitting a line in each co-ordinate system to the points and then increasing the correlation between **t**-scores and the **u**-scores by tilting both lines, the PLS solution is obtained. The **Y** values of a new subject can be predicted using the **t** and other scores obtained during the calculations

Principal components regression and partial least squares regression [19] differ in the methods used in extracting component scores. In short, principal components regression produces the weight matrix **W** reflecting the covariance structure between the predictor variables, while partial least squares regression produces the weight matrix **W** reflecting the covariance structure between the predictor and response variables.

C. PLS Algorithms

There are a large number of algorithmic variants of PLS. Some PLS algorithms are only appropriate for the case where *Y* is a column vector, while others deal with the general case of a matrix *Y*. **PLS1** is a widely used algorithm appropriate for the vector *Y* case. There are two main variants for multivariate PLS regression. The first variant is denoted as PLS2 and the second variant as SIM-PLS. PLS2 is multivariate PLS. NIPALS and Kernal-PLS are two important algorithms.

D. PLS Applications

Application of PLS has been made in regression [20]-[23] as-well-as in classification [24] –[30].

VI. Feed Forward Back Propagation [FFBP] Neural Network

FFBP is the most important type of network used for supervised learning. We shall discuss this type of network in terms of its topology and learning rule.

A. FFBP Network Topology

The data flow from input to output units is strictly feed-forward. The data processing can extend over multiple [layers of] units, but no feedback connections are present, that is, connections extending from outputs of units to inputs of units in the same layer or previous layers. Such Networks are commonly called *feed-forward.* In terms of Graph theory it is a *directed acyclic graph.*

B. Learning

In learning, for a measure of how far away a particular solution is from an optimal solution to the problem, the concept of cost function is introduced. Learning algorithms search through the solution space to find a function that has the smallest possible cost.

The learning is framed as an optimization problem with mean-squared error as its cost function, which is minimized between the network's output, f[x], and the target value y. When one tries to minimize this cost using gradient descent method one obtains the well-known back propagation algorithm [31] for training neural networks. Other learning algorithms have been developed.

Once learning is complete, the weights are stored and can be used to predict future cases in separate test data sets.

E. Training set, Validation set and Test set

All available observations [cases] are divided into two groups: training set and test set. The training set is used to develop the ANN model. During the training, a small part of the training data [may be just a single observation] is not used in the training, but used to validate the model. This part is known as validation set.

F. Assessing the Predictive Power of Trained Model

The external test set is used to assess the real predictive power of the trained ANN. RMSE, R^2 and MEP are some important measures for goodness.

G. Recent Applications of ANN

DNA micro-arrays are one of the methods commonly used for sample profiling at the transcript level. Such data is referred as Gene Expression data. References [32]-[36] present the use of ANN in gene expression data. Mass-spectrometry approaches are being used to generate profiles of Biological samples as well as any analytes which have been ionized. References [37-40] present use of ANN for the analysis of MS data in proteomics.

VII. Illustrations

A. PLS Regression for 3D-QSAR

Quantitative structure-activity relationship [QSAR] is a general process by which chemical structure is quantitatively correlated with a well defined process, such as biological activity.

Traditional QSAR models are unable to explain complex structure–activity data because the extreme specificity of biological activity is described by 3-D intermolecular forces and predicated on 3-D molecular structures. A three-dimensional quantitative structure activity relationship [3D-QSAR] study of diaryloxy-methano-phenanthrene derivatives by comparative molecular field analysis [CoMFA], which uses PLS regression, was performed. A total of 44 diaryloxy-methano-phenanthrene based inhibitors with CoMFA descriptors and minimum inhibitory concentration [MIC] values as bio- activities were used in the study [41]

1] Training set

Thirty seven compounds were used as training set

2] Development of regression model

To derive 3D-QSAR models, the Co MFA descriptors were used as independent variables. The MIC values were converted to the corresponding [−log MIC] and used as dependent variables. Partial least squares [PLS] regression analysis was conducted with the standard implementation in the Sybyl 7.0 package. The predicted values of the models were evaluated by leave-one-out cross-validation. The cross-validated coefficient [$\mathbf{q^2}$] was calculated using Eqs.1 and 2.

$$q^2 = 1 - \frac{\text{PRESS}}{\sum_{i=1}^{N}(y_i - y_m)^2} \quad [1]$$

$$\text{PRESS} = \sum_{i=1}^{N}(y_{pred_i} - y_i)^2 \quad [2]$$

Where y_i is the activity for training set compounds, y_m is the mean observed value corresponding to the mean of the values for each cross-validation group, and $y_{pred,i}$ is the predicted activity for y_i.

The optimal number of components to be used in the analysis significantly influences the prediction ability of the model. The number of components describes the degree of complexity of the model; at some point adding more detail corresponds to fitting the data to noise, and the predictive ability begins to diminish. Usually, the optimal number of components is determined by selecting the highest q^2 value, which most often corresponds to the smallest S value. Whenever the last added component improved q^2 by less than about 5%, the less complex model was chosen. To find out the optimum number of components to be used in CoMFA studies, CoMFA models with different numbers of components were generated. The best CoMFA model [q2=0.625] and minimum standard error value [0.021] are obtained with five components, and further increase in the number of components has no effect on the q^2 value. Hence, five is selected as the optimum number of components for further analysis.

3]Assessment of the developed model on the test set

To validate the derived Co MFA models, biological activities of an external test set of seven compounds were predicted using models derived from the training set. The predictive ability of the models is expressed by the predictive r^2 value, which is analogous to cross-validated r^2 [q^2] and is calculated using the formula

$$r^2_{pred} = \frac{SD - PRESS}{SD} \quad [3]$$

SD is the sum of the squared deviations between the biological activities of the test set molecules and PRESS is the sum of the squared deviations between the observed and the predicted activities of the test molecules. The predicted r^2_{pred} was 0.99 for the model. Thus the predicted values fall very close to the actual MIC.

H. ANN Classification for Mass Spectrometry Data

Carbohydrates which have become the third bio-informative macro-molecule after nucleic acid and protein are considered promising for diagnosis and cure of several diseases. Mass spectrometry is faster and has greater sensitivity in characterizing various carbohydrates. Structural characterization of carbohydrates involves several features [42]. The samples of carbohydrate molecules are introduced into the ionization source of the instrument, and are ionized. The separated ions are detected and this signal called relative abundance of ions is recorded at different positions of m/z value,where m is mass and z is charge.

Independent identification of aldohexoses and ketohexoses is not easy because of high dimensionality in the data and also the number of hexoses may often be fairly large. There is a need for full characterization of various kinds of hexoses. Mass-spectrometry experiment data was analyzed by FFBP network to classify hexoses in eleven types.

1]Training set

Following eleven hexoses were selected for classification study:

- D-allose
- D-altrose
- D-glucose
- D-manose
- D-gulose
- D-idose
- D-galactose
- D-talose
- D-fructose
- L-sorbose
- D-tagatose

Observations of 5 spectra from each hexose were used in the training set

2] Development of ANN model

We applied the ANN module of STATISTICA 7.0 using logistic transfer function on training data set and obtained Best Classified NN Model with

- No. of nodes in Input layer =15
- No. of nodes in a single Hidden layer = 17
- No. of nodes in Output Nodes = 11

3] Assessment of the developed model on the test set

The trained model was then used for five spectra each of Glucose and Manose. The model correctly identified all the ten spectra.

VIII. Concluding Remarks

Data mining is a process involving several steps to help in making intelligent decision based on data. Normally, it starts with a research question. It encompasses search for the relevant data bases and extraction of suitable data from the data base. After this, Statistical and Machine learning methods are applied. When the problem of multi-collinearity occurs, number of predictor variables is more than number of observations and number of response variables more than one, both PLS and ANN are suitable for Regression and Classification. ANN is a generalized computational model, range of problems may be formulated within a framework of ANN. Both the methods have been used for solving many important problems in Bio-Medical area.

IX. Acknowledgement

Authors are grateful to the Director CDRI for granting oermission to present the paper. Authors are thankful to Ms Richa Srivastava for assistance in the preparation of the manuscript. The ideas and views expressed in the paper are those of the authors and not necessarily of the institute they serve.

References

[1] Genome Home.Available: http//wwwNcbi.nlm.nih.gov/.
[2] A. Johnson; C.O 'Donnell, "An open access database of genome-wide association results", BMC medical genetics, 10: 6, 2009.
[3] FGED Society. "Big news: MGED has a new name: FGED - Functional Genomics Data Society" Twitter.http://twitter.com/FGED
[4] JA. Vizcaíno, R. Côté, F. Reisinger, H. Barsnes, JM. Foster, J. Rameseder, H. Hermjakob, L. Martens, "The Proteomics Identifications database: 2010 update", Nucleic Acids Res. 2010 Jan; 38[Database issue]: D736–42. Epub 2009 November 11
[5] DS. Wishart, D. Tzur, C. Knox, et al. "HMDB: the Human Metabolome Database", Nucleic Acids Research, 35 [Database issue]: D521–6, January 2007
[6] The Gene Ontology Consortium, "The Gene Ontology project in 2008", Nucleic acids research,36 [Database issue]: D440–4, Jan 2008.
[7] R. Caspi, H. Foerster, CA. Fulcher, et al., "The Meta Cyc Database of metabolic pathways and enzymes and the Bio Cyc collection of Pathway/Genome Databases", Nucleic Acids Res,. 36 [Database issue]: D623–31, January 2008.
[8] M. Lee, D. Kim, HS. Shin, HG. Sung, JH. Choi, "High-density EEG Recordings of the Freely Moving Mice using Polyimide-based Microelectrode", J Vis Exp,. 11; [47]. pii: 2562., 2011 Jan
[9] Q. Li, Y. Wang, H. Liu, Z. Sun, Z. Liu, "Tongue fissure extraction and classification using hyperspectral imaging technology", Appl Opt. 2010 Apr 10; 49[11]: 2006–13, 2010.
[10] Akbari H, Kosugi Y, Kojima K, Tanaka N., "Blood vessel detection and artery-vein differentiation using hyperspectral imaging", In Conf Proc IEEE Eng Med Biol Soc. 2009: 1461–4.
[11] D. Yudovsky, A. Nouvong, L. Pilon., "Hyperspectral imaging diabetic foot wound care". J Diabetes Sci Technol. 1; 4[5]: 1099–113. 2010 Sep.
[12] Decision Tree Analysis. Available:http://www.mindtools.com/
[13] K. Gurney, "An Introduction to Neural Networks", London: Routledge, 1997.
[14] R. Agrawal; T. Imielinski; A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", In SIGMOD Conference 1993: 207–216.
[15] Ingo Steinwart and Andreas Christmann. Support Vector Machines. Springer-Verlag, New York, 2008.
[16] Haykin, Simon, Self-organizing maps,In Neural networks - A comprehensive foundation ,2nd ed.. Prentice-Hall, 1999.
[17] G.A. Carpenter, & S. Grossberg,, Adaptive Resonance Theory, In Michael A. Arbib [Ed.], The Handbook of Brain Theory and Neural Networks, Second Edition [pp. 87–90]. Cambridge, MA: MIT Press, 2003.
[18] G.A. Carpenter, S. Grossberg, & D.B. Rosen, Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system, Neural Networks [Publication], 4, 759–771, 1991.
[19] S. Wold, M. Sjöström, L. Eriksson."PLS-regression: a basic tool of chemometrics". Chemometrics and Intelligent Laboratory Systems, 58, 109–130, 2001.
[20] S. Datta, "Exploring the relationships in gene expressions: a partial least squares approach". Gene Expression; 9: 257–64, 2001.
[21] LP. Bras, JC. Menezes, "Dealing with gene expression missing data", IEE Syst Biol, 153: 105–19, 2006
[22] DV. Nguyen, N. Wang, RJ. Caroll, "Evaluation of missing value estimation for microarray data". J Data Sci; 2: 347–70. 2004
[23] X. Huang, W. Pan, S. Park, et al., "Modeling the relationship between LVAD support time and gene xpression changes in the human heart by penalized partial least squares". Bioinformatics; 20: 888–94, 2004.
[24] MZ. Man, G. Dyson, K. Johnson, et al. "Evaluating methods for classifying expression data", J Biopharm Stat;14:1065–84, 2004
[25] X. Huang, W. Pan, S. Grindle, et al., "A comparative study of discriminating human heart failure etiology using gene expression profiles". BMC Bioinformatics; 6: 205, 2005
[26] X. Huang, W. Pan, "Linear regression and two-class classification with gene expression data". Bioinformatics; 19: 2072–8, 2003.
[27] TR. Golub, DK. Slonim, P Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". Science; 286: 531–7, 1999
[28] U Alon, DA Barkai, K. Notterman, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays". ProcNatl Acad Sci; 96: 6745–50, 1999.
[29] M Perez-Enciso, M Tenenhaus, "Prediction of clinical outcome with microarray data: a partial least squares approach", Hum Genet; 112: 581–92, 2003.
[30] G Musumarra, V. Barresi, DF Condorelli, et al. "Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by PLS discriminant analysis". J Chemom; 18: 125–32, 2004.

[31] Chapter 7 The backpropagation algorithm in Neural Networks - A Systematic Introduction by Raúl Rojas [ISBN 978–3540605058]

[32] J Khan, JS. Wei, M. Ringner, et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks". Nat Med; 7: 673–9, 2001.

[33] NR Pal, K. Aguan, A. Sharma, et al. "Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering". BMC Bioinformatics; 8: 5, 2007.

[34] C. Peterson, M. Ringner, Analyzing "tumor gene expression profiles". Artif Intell Med; 28: 59–74, 2003.

[35] LE. Peterson, MA. Coleman, "Machine learning-based receiver operating characteristic [ROC] curves for crisp and fuzzy classification of DNA microarrays in cancer research". Int J Approx Reason; 47: 17–36, 2008.

[36] J. Xuan, Y. Wang, Y. Dong, et al. "Gene selection for multiclass prediction by weighted fisher criterion". J Bioinform Syst Biol; Article No. 64628, 2007.

[37] MA. Rogers, P. Clarke, J. Noble, et al. "Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility". Cancer Res; 63: 6971–83, 2003.

[38] YD. Chen, S. Zheng, JK. Yu, et al., "Artificial neural networks analysis of surface-enhanced laser desorption/ionization mass spectra of serum protein pattern distinguishes colorectal cancer from healthy population". Clin Cancer Res; 10: 8380–5, 2004.

[39] DG. Ward, N. Suggett, Y. Cheng, et al., "Identification of serum biomarkers for colon cancer by proteomic analysis". Br J Cancer; 94: 1898–905, 2006.

[40] JM. Luk, BY. Lam, NP. Lee, et al. "Artificial neural networks and decision tree model analysis of 40 liver cancer proteomes". Biochem Biophys Res Commun;361: 68–73, 2007.

[41] Shagufta, ashutosh Kumar, Gautam Panda, Siddiqi MI , "CoMFA and CoMSIA 3D-QSAR analysis of diaryloxy-methano-phenanthrene derivatives as anti-tubercular agents". J Mol Model; 13: 99–109, 2007.

[42] KP. Madhusudanan, and Mukesh Srivastava "Identification of hexoses diasteromers by means of tanden mass spectrometry of oxocarbenium ions followed by neural network analysis" J. Mass Spectrom,. 43: 126–131, 2008.

Finite Automata Based Pattern Mining

Kavita S. Oza
Department of Computer Science Shivaji University, Kolhapur
e-mail: kavita_oza@rediffmail.com

***Abstract*—In this paper, we propose the use of simple Automata theory approach to mine for particular pattern in a given dataset. Previous approaches to solve this problem were focusing on its Data structure used for mining and most of the data structures used were tree based. On the contrary, here the focus is completely on the given input data and pattern to be mined without using any complex data structure or pattern matching algorithms. Here goal of data mining is compromised as we are not mining any unknown pattern but checking the presence of a user defined pattern. It counts the occurrences of patterns contained in the sequence. It also gives association between subsequences of a pattern to be mined. The pattern is frequent if it satisfies the given support threshold. The proposed solution has the interesting features like : it performs a unique pass over the input database, and it is minimum support threshold independent.**

Keywords: Association, Pattern mining, Deterministic Finite Automata, support threshold

I. INTRODUCTION

The increased storage of voluminous data in digital form has increased the interest in the automatic discovery of hidden information, and data mining techniques. Usually, the term data mining is used to mean "the nontrivial extraction of implicit, previously unknown and potential useful information from data" Frawley et al [5]. In general, data mining techniques have been successfully applied from commercial domains, like customer relationship management, market basket analysis or credit card fraud detection, to scientific and engineering applications.

However, data mining algorithms are usually unable to produce optimal results with respect to all the trade-offs that they account for: sample size versus error rate, or simply model expressiveness versus compute time, to name a few, Bayardo. [6]. In particular, algorithms for discovering frequent patterns discover large amounts of patterns, most of the times, uninteresting and useless to the final user. The inability to focus the discovery process on user expectations and background knowledge, has lead to a process that is, in many cases, prohibitively expensive and very difficult to deal. A particular case of pattern mining usually suffers from these drawbacks.

In order to minimize this problem, in pattern mining, recent approaches use constraints to restrict the number and scope of discovered patterns. Renata Ivancsy et al. [1] classified pattern discovering into two main classes, namely, in the class of the level wise methods and that of the database projection-based methods. For discovering frequent sequences and tree-like patterns efficiently they introduced the idea of using automaton theory. Roberto Trasarti et al. [2] studied the problem of mining frequent sequences satisfying a given regular expression. They introduced a sequence mining automata to mine sequences satisfying the given regular expression. Sandrade Amo et al. [3] proposed to use tree automata as a mechanism to specify user constraints over tree patterns. They presented the algorithm CoBMiner which allows user constraints specified by tree automata to be incorporated in the mining process. Algorithm (ε-accepts) that verifies if a sequence is approximately accepted by a given regular language is proposed by Cláudia Antunes [4]. ε-accepts also uses regular language as a constraint.

This paper presents a simple pattern mining approach, which keeps the focus on user expectations, and the process will find only user defined patterns. Here automaton theory is used for discovering the support of the candidate patterns efficiently and also association between the subsequences of a pattern. Proposed algorithm is based on pattern inclusion test discussed in section-III.

The organization of the paper is as follows. Section -II introduces the problem of pattern mining. Section - III presents the pattern inclusion test which also gives details of creating the automatons for pattern mining. . In Sections- IV the details of algorithm are explained Experimental results are shown in Section-V. Conclusion can be found in Section- VI.

II. PROBLEM DEFINITION

Let D be a database of transactions, where each T Є D is a finite pattern of symbols from an alphabet Σ: T = (t_1, …..t_n) where t_i Є Σ, for all i Є {1, .. ,T_n}.

The support of a pattern P is the number of transactions in D that contain the given pattern P. Given a database D and a minimum support threshold σ, the frequent pattern is

$F(D, \sigma) = \{P \in \Sigma^* \mid sup_D(P) \geq \sigma\}$.

III. PATTERN INCLUSION TEST

For testing Pattern inclusion deterministic finite state machines can be used.

A deterministic finite state machine is a 5- tuple, $(Q, \Sigma, \delta, q_0, F)$ consisting of:

- a finite set of states (Q),
- a finite set called the alphabet (Σ),
- a transition function ($\delta: Q \times \Sigma \rightarrow Q$),
- a start state ($q_0 \in Q$),
- and a set of accept states (F is a subset of Q).

The state machine accepts the input string if the string contains the candidate pattern. For this reason the patterns are represented with strings from the alphabet Σ.

Definition: Let $P=p_0,p_1,\ldots,p_s$ be the string representation of a candidate pattern of size *k*, where *s+1* equals to the length of the pattern *P*. The rules for generating a deterministic finite state machine for the pattern *P* are given in Table I, where $Q_i(Q_i \in Q, i=0\ldots s+1)$ denotes the states of the machine, and $\Sigma\text{-}c_i$ denotes all characters in the alphabet Σ except c_i and the following conditions hold: $Q_0 = q_0$ and $Q_{s+1} \in F$.

TABLE 1: TRANSITION FUNCTIONS OF THE FINITE STATE MACHINE OF THE CANDIDATE PATTERN $P=p_0,p_1,\ldots,p_S$

Input items	Transition function
$p_0 \in \Sigma$	$\delta(Q_0, p_0) = Q_1$ $\delta(Q_0, \Sigma - p_0) = Q_0$
$p_i \in \Sigma$	$\delta(Q_i, p_i) = Q_{i+1}$, i=1…s $\delta(Q_i, \Sigma - p_i) = Q_i$

The machine starts in the start state Q_0. For each new character a new state is created and the transition between the states contains the character. These are represented in the state diagram of the finite state machine as forward edges [1]. The backward edge is created between the states having no transition with new character read. From each state there exist transitions to all the items such that the state will be same i.e. self loop. The accept state of the machine is the state for the last symbol of the pattern. As shown in the following example.

Example: Given pattern P = CDGS, we show how the transaction T

T = {ACDACDGNU } over the alphabet set Σ= { A,B,…..,Z} is processed by DFA- S and the whole process is graphically described in Figure 1.

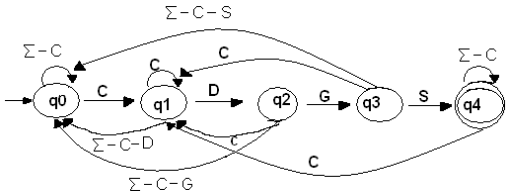


Fig. 1

It starts with the initial state q0. The input sequence is processed from left to right. The first symbol of the pattern A is read, there is no transition aimed at processing the input A, so no change in the state of the automaton S. When the second input symbol C is read, there is a transition on reading C, which changes the state from q_0 to q_1. When the third D symbol is read at the state q1, there is a transition aimed at processing the input D, so there is change in the state of the automaton S from q1 to q2. Now at state q2 the next input symbol read is A. It causes the transition (backward edge) from q_2 to q_0, which changes the state of automaton from q_2 to q_0. Now fifth input symbol C is read at state q_0, causing the transition from q_0 to q_1. Next input symbol read at q_1 now is D, there is a transition aimed at processing D, causing the transition from q_1 to q_2. Seventh input symbol G is read at state q_2 causing the transition from q_2 to q_3. When the eight input symbol N is read at state q3, there is a transition for this input causing change of state from q_3 to q_4, q_4 is the final state. Any symbol read at this state, there is no transition to process it so no change in the state of the automaton S. After reading the last input symbol U, automaton S remains in the same state i.e. q_4 (final state). If while reading the input a final state is reached, means the pattern is embedded in the transaction.

IV. ALGORITHM

In proposed algorithm each state of automaton is treated as a counter which increases whenever an input symbol read causes the change in the state. The state reached after reading the input symbol is the state whose counter will be increased by one. After the whole input database D is processed using a DFA, we get total number of transactions having pattern P embedded in them. It also gives the occurrences of sub patterns of the pattern P. Like if ABC is the pattern than we also get the count of sub patterns of ABC ie count of A, AB, and ABC, if the counter of all the sub patterns and pattern P is same then there is some association between the items in the pattern. The association rule will be if a customer buys item A, he also Buys item B and C.

Input to our algorithm is the database consisting of set of transactions T.

T= { $t_1,t_2, \ldots.t_n$ } , where t_i is the item of transaction T.

Algorithm 1

Input: D, P, Σ , σ

D- Database, P- pattern, Σ- input alphabet set, σ – support threshold

Output: (Frequency of Pattern P and sub patterns of P) >=σ

1. for all T $\in$ D do
2. compute δ (t_i, q)
3. for all $t_i \in$ P do
4. ++q. counter
5. end for
6. end for

V. EXPERIMENTAL RESULTS

Algorithm is implemented using core Java. All the experiments have been run on a Windows XP machine.

The Chess dataset used in experiment is downloaded from http://www.fimi.cs.helsinki.fi/data. Experiments are carried out on using different patterns to be mined in the dataset. This approach is support-independent and thus its run-time stays constant as the support changes. This consideration makes proposed algorithm an efficient one for mining at low support

TABLE 2

Dataset	#items	# Avg. Length	# Transactions
Chess	75	37	3,196

Following graph shows the frequency of the pattern < 11-13-15-17>, and its sub patterns (<11-13-15>, <11-13>, <11>) with the support factor σ = 0.2.

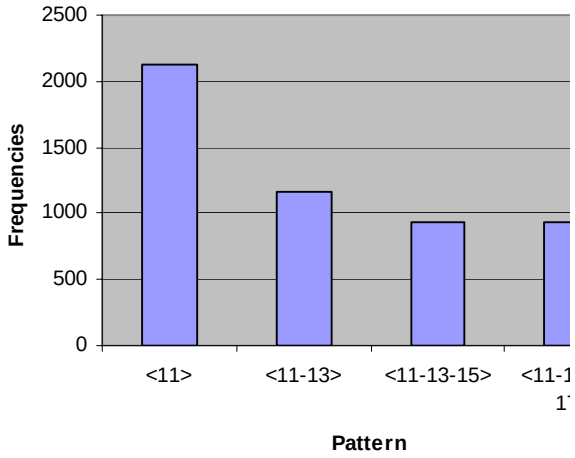


Fig. 2

In fig-2, it is observed that frequencies of patterns <11-13-15> and <11-13-15-17> are same. It indicates there is some association between the items <15> and <17>. Whenever item <15> is bought item <17> is also bought. This information can help in arrangement of items in a super market.

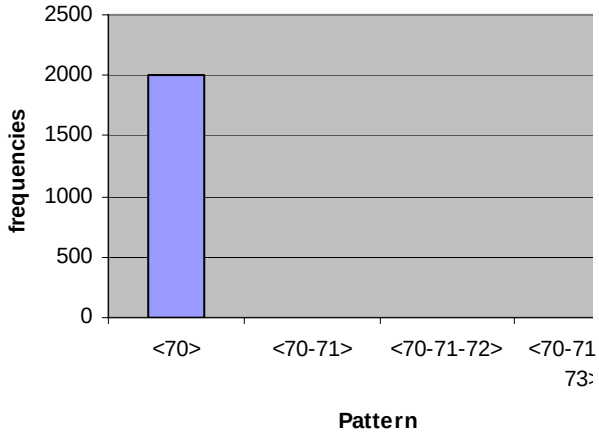


Fig. 3

Fig-3 gives frequencies of pattern <70-71-72-73> and its sub patterns. It is observed that this pattern does not occur in the database and its sub patterns have no association between them.

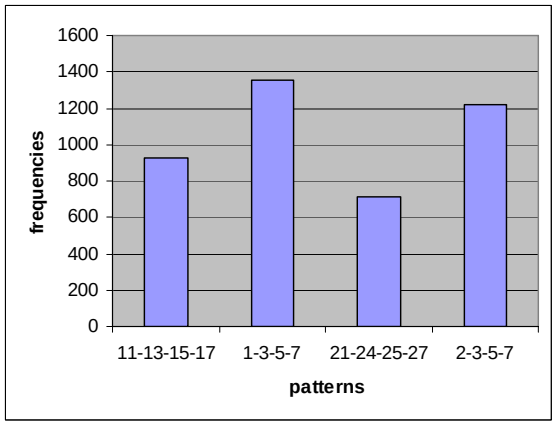


Fig. 4

Frequencies of patterns <11-13-15-17>, <1-3-5-7>, <21-24-25-27> and <2-3-5-7> are

927, 1351, 709 and 1217 respectively are reported in the fig-4.

VI. CONCLUSION

This algorithm is efficient if the user is not interested in any unknown pattern and just wants the status of any known pattern. In case of string search proposed algorithm gives better results. It can work for patterns of any length. Proposed solution has two interesting features: (1) it performs a unique pass over the input database (2) it is minimum support threshold independent. This algorithm does not require any complex algorithm or data structure for processing the data. It is memory efficient as nothing is stored except the values of the counters. Drawback of this method is, it does not perform any frequency-based pruning. It does not mine any unknown patterns but checks the presence of given pattern along with association between its sub patterns. Enhancement of this work will be to accept a regular expression for mining rather than a pattern.

REFERENCES

[1] Renáta Iváncsy, and István Vajk , *Automata Theory Approach for Solving Frequent Pattern Discovery Problems*, World academy of Science, Engineering and Technology 8, 2005.

[2] Roberto Trasarti, Francesco Bonchi and Bart Goethals, *Sequence Mining Automata: a New Technique for Mining Frequent Sequences Under Regular Expressions, 2008.*

[3] SandradeAmo, NyaraA.Silva, Ronaldo P.Silva and Fabiola S. Pereira, *Tree pattern mining with tree automata constraints* , Information Systems 35(2010) 570–591, Elsevier Journal.

[4] Cláudia Antunes and Arlindo L. Oliveira, *Sequential Pattern Mining With Approximated Constraints,* IADIS International Conference Applied Computing 2004, pg. no -131-138.

[5] Frawley, W. et al, *Knowledge discovery in databases: an overview. AI Magazine*, vol. 13, no. 3, pp. 57–70., 1992.

[6] Bayardo, R.J., *The Many Roles of Constraints in Data Mining*, SIGKDD Explorations, vol. 4, no. 1, pp. i-ii, 2002.

Determination of Soil Type from Farmer's Description of Soil: A Natural Language Processing Tool

Syed Khizer[1] and H.S. Acharya[2]
Department of Computer Science, Al-Namas Community College, King Khalid University, Saudi Arabia
Alana Institute of Management Sciences, Camp, Pune–411001, (MS) India
e-mail: syed_khizer@yahoo.com, haridas.undri@gmail.com

***Abstract*—In any advisory system, advises are given on the basis of the user's input. Accepting user's views (input) in his/her own natural language is very essential for generating useful and correct advices. In turn it determines the success and acceptance of the system, especially when the users of the advisory system are illiterate and/or poses poor knowledge. Very next difficulty lies in transforming user's views to the standard domain data, information and knowledge. In this article solution to the problem of accepting user's soil description and transforming it to standard soil type, using Natural Language Processing methods is discussed. The system was evaluated by comparing its results to those provided by the experts in the panel and farmers; a strong agreement was found between decisions by the system and the panel experts. The tool is a valuable and important component in any agriculture advisory/management system (example: Crop, Irrigation, Manure/Fertilizer …). The solution suggested here can be applied to the similar problems from any domain.**

Keywords: Natural language processing (NLP), Soil type determination, Expert systems, Artificial intelligence, Automated Text Processing, Computational Linguistics, Information extraction.

I. Introduction

Agricultural systems are quite soil-specific. Crop selection is an important decision [1]. Soil's descriptions are quite complex. The origin of the soil, colour, texture, chemical properties and depth, all need to be specified while describing the soil. The specific words used are likely to vary quite a bit as people tend to describe the same soil in different manner. Even their perceptions seem to be different [2, 3]. Determination of correct soil type from soil description provided by the user/farmer thus becomes a very complex process. A major challenge here is to map the farmer's description of the soil to the correct scientific one. This is very important in the process of generating the appropriate crop advice for the famers, using any automated system. The work reported in this article focuses on process of developing an intelligent Natural language processing (NLP) tool which can be used in larger advisory system [4].

II. Knowledge Base Design

Knowledge comes from research, experience of experts and various published works where most of the experience is documented. To begin with a panel of experts with adequate experience in the field was identified. They were asked to help in identifying reliable published sources where basic scientific descriptions of soils were available. A brief description of sources [5, 6] is given here. The experts also guided us at various stages of designing rule base.

A. Knowledge Sources

"Soil Resource Inventory of Marathwada" [2] and "Soils of Maharashtra for Optimizing Land Use" [3, 7], were important sources which gave scientific descriptions of soils needed in the development of this tool.

"Krishi Dainandini" [5] and "Krishi Darshani" [6] are yearly publications by departments of extension of regional Agricultural Universities, published basically for benefit of farmers, extension workers, students and researchers. Information regarding crops, varieties, soil, irrigation, metrology (weather), insecticides, pesticides, fertilizer, cropping patterns, economics, export and agriculture related business such as poultry farming, silk worm, goat farming, dairy are found here. Most of the information in this book is textual.

"Handbook of Agriculture" [8] contains the information related to all aspects of Indian agriculture on the national level.

III. Collection of Phrases

The first step in designing the NLP tool is to determine the morphology of the text to be processed [5, 6, 8] i.e. to determine structure and form of words. We gathered all the words and phrases that are used in the possible descriptions of the soils (Table I) in the English and local language (Marathi). These are collected from different sources as indicated in Section (2) earlier.

TABLE I: SOIL DESCRIPTIONS

Sr. No.	Soil Descriptions (English and Local Langue)
1	Very Shallow / *अंत्यत हलकी*
2	Hilly Shallow / *बारडाची डोंगराळ*
3	Crumby / *भूशभूशीत*
4	Hilly Sloped / *डोंगर उताराची*
5	Hilly / *डोंगराळ*
6	Shallow / *हल□ी*
7	Light Lateritic / *हल□ी मुरमाळ*
8	Poor Drained / *कमी नीचरियाची*
9	Red Shallow / *लाल हल□ी*
10	Reddish / *लालसर*
11	Lateritic / *मुरमाळ*
12	Brownish / *तांबळी भूरकट*
13	Shallow Light / *उथळ आणि हलकी*
14	Light to Medium Heavy / *हलकी ते मध्यम भारी*
15	Light to Medium / *हलकी ते मध्यम*
16	Medium to Light / *मध्यम ते हलकी*
17	Shallow to Medium Deep / *उथड ते मध्यम खोल*
18	Medium and Black / *मध्यम आणि काळी*
19	Medium and Deep / *मध्यम आणि खोल*
20	Medium and Moderately Well Drained / *मध्यम आणि निचरा होणारी*
21	Medium and Moderately Well Drained / *मध्यम आणि निचरियची*
22	Medium and Well Drained / *मध्यम आणि उतम निचरियाची*
23	Medium Shallow Moderately Well Drained / *मध्यम हलकी उतम निचरियाची*
24	Medium / *मध्यम*
25	Medium Black / *मध्यम □ाळी*
26	Medium Deep / *मध्यम □ोल*
27	Medium Fertile / *मध्यम प्रतिची*
28	Medium to Medium Deep / *मध्यम ते मध्यम खोल*
29	Medium Black to Heavy / *मध्यम काळी ते भारी*
30	Medium Black to Heavy Black / *मध्यम काळी ते भारी काळी*
31	Medium Black to Black / *मध्यम काळी ते काळी*
32	Medium to Heavy / *मध्यम ते भारी*
33	Medium to Heavy Deep Well Drained / *मध्यम ते भारी खोल उतम निचरियाची*
34	Medium to Heavy Moderately Well Drained / *मध्यम ते भारी निचरा होणारी*
35	Medium to Heavy Fertile / *मध्यम ते भारी सुपीक*
36	Medium to Heavy Fertile Well Drained / *मध्यम ते भारी सुपीक उतम निचरियाची*
37	Medium to Deep Well Drained / *मध्यम ते खोल उतम निचरियाची*
38	Deep and Black / *भारी आणि काळी*
39	Heavy / *भारी*
40	Heavy Moderately Well Drained / *भारी निचरियाची*
41	Well Drained / *चांगल्या निचरायची*
42	Black and Fertile / *काळी आणि कसदार*
43	Black / *□ाळी*
44	Fertile / *□ सदार*
45	Deep Black / *□ोल □ाळी*
46	Deep Black / *खोल मातीची*
47	Medium Moderately Well Drained Deep / *मध्यम निचरियाची खोल*
48	Highly Acidic / *अती अमलयूक्त*
49	Calcareousness / *चुनखडी असलेली*
50	Sodic / *चोपण*
51	Water Logged/ *दलदलीची*
52	Alluvial / *गाळाची*
53	Saline / *खारवट*
54	Water Logged / *पानथड*
55	Clayey / *पोयटयाची*
56	Sandy / *रेताळ*
57	Salt Affected / *क्षारयूक्त*
58	Alkaline / *विमलयूक्त*
59	Sand mixed / *वालूमिश्रित*
60	Sandy / *वालू□मय*

IV. VOCABULARY CONSTRUCTION

TABLE II: VOCABULARY OF WORDS USED IN DESCRIBING SOIL

Sr. No.	Occurring in First position	Occurring in Second position	Occurring in Third position
1	*अंत्यत*	*आणि*	*भारी*
2	*अती*	*अमलयूक्त*	*हल□ी*
3	*बारडाची*	*असलेली*	*□ाळी*
4	*भारी*	*भूरकट*	*□ सदार*
5	*भूशभूशीत*	*डोंगराळ*	*□ोल*
6	*चांगल्या*	*हल□ी*	*मध्यम*
7	*चुनखडी*	*□ाळी*	*निचरियाची*
8	*चोपन*	*□ोल*	*निचरा*
9	*दलदलीची*	*मातीची*	*ते*
10	*डोंगराळ*	*मिश्रित*	*उतम*
11	*डोंगराळ*	*मुरमाळ*	
12	*गाळाची*	*निचरियाची*	
13	*हल□ी*	*प्रतिची*	
14	*□ाळी*	*ते*	
15	*□मी*	*उताराची*	
16	*□ सदार*		
17	*खारवट*		
18	*□ोल*		
19	*लाल*		**Distinct words :**
20	*लालसर*		*अमलयूक्त*
21	*मध्यम*		*आणि*
22	*मुरमाळ*		*असलेली*
23	*पानथड*		*अती*
24	*पोयटयाची*		*अंत्यत*
25	*रेताळ*		*बारडाची*
26	*क्षारयूक्त*		*भारी*
27	*तांबळी*		*भूरकट*
28	*उथळ*		*भूशभूशीत*
29	*विमलयूक्त*		*चांगल्या*
30	*वालू*		*चुनखडी*
31	*वालू□मय*		*चोपन*

Occurring in Fourth position	Occurring in Fifth position	Occurring in Sixth position
भारी	*हो□ारी*	*निचरियाची*
हो□ारी	*□ाळी*	
□ाळी	*निचरियाची*	
□ोल	*उतम*	
निचरा		
निचरियाची		
सुपी□		
उतम		

दलदलीची	*लालसर*	*सुपी□*
डोंगर	*मध्यम*	*तांबळी*
डोंगराळ	*मातीची*	*ते*
गाळाची	*मिश्रीत*	*उताराची*
हल□ी	*मुरमाळ*	*उथड*
हो□ारी	*निचरा*	*उतम*
□ाळी	*निचरियाची*	*विमलयूक्त*
□मी	*पानथड*	*वालू*
□ सदार	*पोयटयाची*	*वालू□मय*
खारवट	*प्रतिची*	
□ोल	*रेताळ*	
लाल	*क्षारयूक्त*	**Total = 45**

The second step is the determination of syntax i.e. rules for putting words together to form phrases/sentences/descriptions [9]. In the second step unique words are separated, with possible order of their appearance. Table (II) lists all the words and the sequences used by farmers to describe soils. The words are identified from Table (I), column 2 that contains sixty different soil descriptions. The Table (II) is used to provide drop down list boxes for accepting user soil description.

V. Identifying Common Words

Table (III) contains all the common, qualifier and connecting words used in the user soil description; this is identified by using Table (I).

TABLE III: COMMON, QUALIFIER AND CONNECTING WORDS

Sr. No.	Common, qualifier and connecting words (Words to be removed from user soil description)
	Marathi
1	आणि
2	असलेली
3	अती
4	अत्यत
5	मिश्रित
6	ते
7	होऊशरी

There are lots of redundancies in the way farmers describe the soils. The next (i.e. third) step was to determine distinct descriptors in the description. It is accomplished by removing common, qualifier and connecting words (Table III) from sixty different soil descriptions listed in the Table (I).

VI. Building Semantic Rule Base

TABLE IV: SOIL TYPES

Sr. No.	Soil Type	Description
1.	Light	Shallow, Low in clay , good for root proliferation, but depth only up to 30 cm Clay % < 10 , slope < 5 %
2.	Light to Medium	Shallow , Low in clay , good for root proliferation, depth 31cm - 90 cm Clay % < 10 , slope < 5%
3.	Medium	Clay < 20%, depth 90 cm –1.5m, Slope < 5%
4.	Medium to Heavy	Clay <40%, depth 90cm -1.5m, Slope < 5%
5.	Heavy	Clay < 40%, depth > 1.5m, Slope < 5%
6.	Problematic	Acidic/ Alkaline/ Waterlogged/Undulating/Slope > 5%

We have categorized normal soils into five common soil types and in the sixth category all problematic soils are put together. The association with closest soil type which corresponds to the soil description is finalized in consultation with the panel of the experts. A reference table of various soil descriptions and the equivalent class identifiable is given in Table (V).

The fourth step is the construction of the semantic rule base [11, 12] i.e. determining the meaning of word and its description. Soil descriptions in the Table (I) describing the soils are classified in to six classes (i.e. soil types), as given in the Table (V) column 3, using soil classification reference Table (IV).

TABLE V: DISTINCT DESCRIPTORS AND SOIL TYPES/CLASSES

Sr. No.	Distinct Descriptors	Soil Type
1	हलकी	Light/ हलकी
2	बारडाची डोंगराळ	
3	भुशभुशीत	
4	डोंगर उताराची	
5	डोंगराळ	
6	हलकी	
7	हलकी मुरमाळ	
8	कमी नीचरियाची	
9	लाल हलकी	
10	लालसर	
11	मुरमाळ	
12	तांबळी भूरकट	
13	उथड हलकी	
14	हलकी मध्यम भारी	Light to Medium/ हलकी ते मध्यम
15	हलकी मध्यम	
16	मध्यम हल	
17	उथड मध्यम खोल	
18	मध्यम काळी	Medium/ मध्यम
19	मध्यम तोल	
20	मध्यम निचरा होणारी	
21	मध्यम निचरियची	
22	मध्यम उतम निचरियाची	
23	मध्यम हलकी उतम निचरियाची	
24	मध्यम	
25	मध्यम काळी	
26	Madhyam Khol/ मध्यम तोल	
27	मध्यम प्रतिची	
28	मध्यम मध्यम तोल	
29	मध्यम काळी भारी	Medium to Heavy/ मध्यम ते भारी
30	मध्यम काळी भारी काळी	
31	मध्यम काळी काळी	
32	मध्यम भारी	
33	मध्यम भारी खोल उतम निचरियाची	
34	मध्यम भारी निचरा होणारी जमीन	
35	मध्यम भारी सुपीक	
36	मध्यम भारी सुपीक उतम निचरियाची	
37	मध्यम खोल उतम निचरियाची	
38	भारी काळी	Heavy/ भारी
39	Bhari/ भारी	
40	भारी निचरियाची	
41	चांगल्या निचरायची	
42	काळी कसदार	
43	काळी	
44	कसदार	
45	तोल काळी	
46	खोल मातीची	
47	मध्यम निचरियाची खोल	
48	अमलयूक्त	Problematic/ समसयायूकत
49	चुनखडी	
50	चोपण	
51	दलदलीची	
52	गाळाची	
53	खारवट	
54	पानथड	
55	पोयटयाची	
56	रेताळ	
57	क्षारयूक्त	
58	विनलयूक्त	
59	वालूमिश्रित	
60	वालूमय	

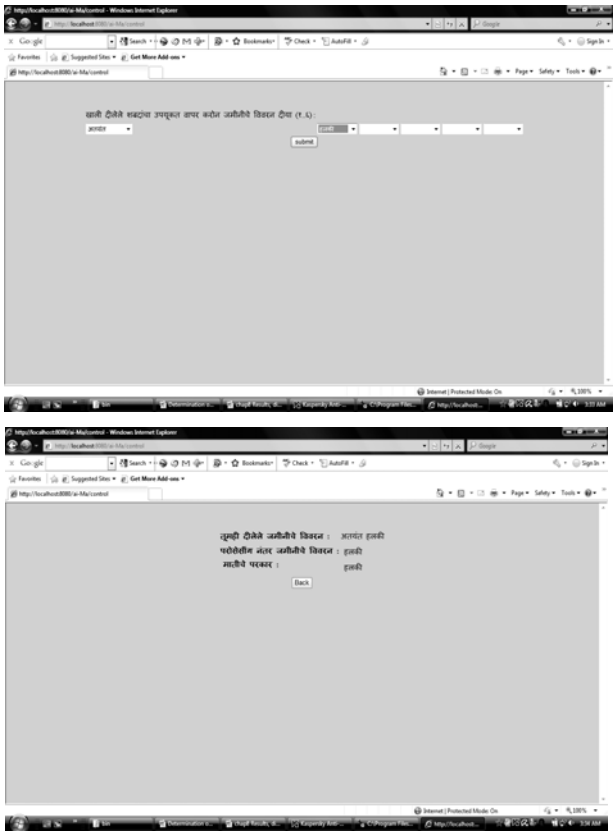


Fig. 1: Sample Screen of user Interface and Output of the System (Marathi)

VII. Mapping Procedure

User can give soil description by using drop down list boxes (Table II, column 2-6) or by typing in text boxes (Fig. 1). After accepting the user's soil description common, qualifier and connecting words listed in Table (III) are removed. Afterwards the processed soil description is matched with the distinct soil descriptors listed in the Table (V) column 2 [13]. If it matches completely then a corresponding soil type (Table V, column 3) is displayed. In case if it doesn't match then the user is requested to re-enter the soil description once again (Fig. 1). Other than sixty standard soil descriptions a farmer can choose a different soil description by using the same drop down list boxes and a meaningful soil type can be determined.

VIII. Soil_Type Determination Algorithm

1. Start
2. Set *Flag = False*
3. Initialize the *vocabulary_array* with the vocabulary words (section IV, Table II)
4. Initialize the *common_qualifire_connecting_words_array* with the common, qualifier and the connecting words (section 5, Table III)
5. Initialize the *soil_type_array[i][j]* with the *distinct descriptor* and *soil type* (section 6, Table V, column two and three))
6. Display the *vocabulary_array* (Table II) in the dropdown list box
7. Accept the *user_soil_description* from the user
8. 8. Match and remove the words in *common_qualifire_connecting_words_array* from the *user_soil_description*
9. Repeat step 10 for i = 1… n (i.e. for each distinct descriptor in the *soil_type_array*)
10. If *user_soil_description = = distinct_descriptor_array[i]* then

a. farmer *soil type = soil_type_array[i][j]* (i.e. corresponding soil type)
b. *Flag = True*
c. Go to step 12 (section 7)

11. If *Flag = = False* then go to step 6 (Ask user to reenter soil description as soil description is vague)
12. Display farmer *soil type*
13. Stop

IX. Results, Discussions and Validations

In this section, we discuss the results of the program developed for the algorithm ***Soil_Type*** given in previous section (i.e. section 8).

A. Experiment 1: using Standard soil Descriptions

Standard soil descriptions are used to determine the soil types here (Table VI).

TABLE VI: Results of Experiment Using Standard Soil Descriptions

Run	Label	Input, Processed, Result
1	**Entered Soil Description :**	*वारडाची डोंगराळ*
	Soil Description After Processing :	*वारडाची डोंगराळ*
	Soil Type:	Light/ *हलकी*
2	**Entered Soil Description :**	*मध्यम हलकी उतम निचरियाची जमीन*
	Soil Description After Processing :	*मध्यम हलकी उतम निचरियाची जमीन*
	Soil Type:	Medium/ *मध्यम*
3	**Entered Soil Description :**	*मध्यम ते भारी खोल उतम निचऱ्याची*
	Soil Description After Processing :	*मध्यम भारी खोल उतम निचऱ्याची*
	Soil Type:	*मध्यम ते भारी*
4	**Entered Soil Description :**	*तांबुस काळी*
	Soil Description After Processing :	*तांबुस काळी*
	Soil Type:	Heavy/ *भारी*
5	**Entered Soil Description :**	*अती अम्लायूक्त*
	Soil Description After Processing :	*अती अम्लायूक्त*
	Soil Type:	Problematic/ समसयायूकत

Identified soil types by the system for a given soil descriptions (experiment number 1) are in agreement found in literature and panel of experts.

B. Experiment 2: using other Soil Descriptions

Validation using non standard soil descriptions was carried out in this test (Table VII).

TABLE VII: RESULTS OF EXPERIMENT USING NON STANDARD SOIL DESCRIPTIONS

Run	Label	Input, Processed, Result
1	**Entered Soil Description :**	अत्यंत काळी
	Soil Description After Processing :	□ ळी
	Soil Type:	भारी
2	**Entered Soil Description :**	अती कसदार
	Soil Description After Processing :	□ सदार
	Soil Type:	भारी
3	**Entered Soil Description :**	अत्यंत मुरमाड
	Soil Description After Processing :	मुरमाड
	Soil Type:	Light/ हल□ी

Results of test runs in the above Table (VII) show if other soil descriptions (other than standard soil descriptions) were used then system identifies correct soil type.

X. CONCLUSION

A system has been developed to map soil descriptions used in the Marathwada region (Maharashtra, India) for a soil type. The system consists of standard sixty semantic rules. The system can be used to determine a soil type based on the soil description given by the farmer. The system is evaluated by comparing its results to those provided by the experts in the panel and farmers. In most of the cases studied, agreement was found between decisions by the system and the panel experts.

The developed system can be used in crop advisory systems [14].

ACKNOWLEDGMENT

Work presented in this paper is the part of research undertaken for pursuing Ph.D., submitted to the Swami Ramanand Teerth University, Vishnupuri, Nanded, Maharashtra, India.

REFERENCES

[1] Mohan, S. and Arumugam, N. (1994), CROPES: A rule based ES for crop selection in India, Transactions of ASAE vol. 37(3): pp. 1355–1363.

[2] Anonymous (2002), Soil Resource Inventory of Marathwada, Dept. of Soil Sciences, MAU Parbhani-431 402, MS, India.

[3] Anonymous (1995), Soils of Maharashtra for Optimising Land Use, ICAR Nagpur and Department of Agriculture Government of Maharashtra, Pune, MS, India. Pub. 54.

[4] Tomaž Erjavec (2007), Introduction to Human Language Technologies, Karl-Franzens-Universität Graz http://nl.ijs.si/et/teach/graz07/hlt/graz07-hlt-1-handout.pdf

[5] Anonymous (2003), Krishi Dainandini, Department of extension, MAU Parbhani–431 402, MS, India.

[6] Anonymous (2005), Krishi Darshani, Department of extension, MPKV Rahuri-413 722, Ahmad Nagar, MS, India.

[7] O. Challa, K. S. Gajbhiye and M. Velayutham (1999), Soils of Maharashtra for Optimising Land Use, NBBS pub. No. 54b, NBBS&LUP, Nagpur India, 112 p+6 sheet soil map (1:5000,000 scale) (46)

[8] Anonymous (1987), Handbook of Agriculture, Pub. Indian Council of Agric. Res., New Delhi.

[9] Wikipedia, (2010) Computational linguistics, Natural Language Processing. [Online]. Available at: http://en.wikipedia.org/wiki/Computational_linguistics , http://en.wikipedia.org/wiki/Natural_language_processing [Accessed: 24 September 2010].

[10] Joshua Zhexue Huang & Michael Ng. & Liping Jing (April 9, 2006), Text Clustering: Algorithms, Semantics and Systems, PAKDD06 Tutorial Singapore [Online]. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.3150&rep=rep1&type=pdf [Accessed: 24 September 2010].

[11] Rich, E. and Kavin, K. (1991), Artificial Intelligence, ISBN: 0074600818, ISBN-13: 9780074600818, 978-0074600818 Publisher: Tata Mcgraw Hill Publishing Company Limited

[12] Ronald, A. Waterman, (1985), A Guide to Expert System, Addison-Wesley Teknowledge Series In Knowledge Engineering, ISBN:0-201-08313-2

[13] Diego Mollá , Rolf Schwitter , Fabio Rinaldi , James Dowdall , Michael Hess, ExtrAns (July/August 2003), Extracting Answers from Technical Texts, IEEE Intelligent Systems, v.18 n.4, p.12–17 [Online]. Available at: http://web.science.mq.edu.au/~diego/publications/ieee03.pdf [Accessed: 24 September 2010].

[14] Syed Khizer, "Intelligent Support System for Generating Cropping Patterns of Agricultural Farming", Thesis Submitted for Ph.D. , Faculty of Computer Science, SRTMU, Nanded, September, 2008.

Application of Data Mining in Agriculture Portfolio Problem

Ratnmala Bhimanpallewar and Pravin Metkewar
Department of Information Technology, Pune University
JSPM's BSIOTR, Wagholi, Pune, Maharashtra, India
e-mail: ratnmalab@gmail.com, pravin_metkewar@rediffmail.com

***Abstract*—For Agriculture Portfolio problem (APP), funds provided by Indian Investment Authority for the development of Agriculture in Indian rural areas specific to the farmer and there is a need to maintain international competition with other countries. Under this portfolio loans will be provided to the farmers with various schemes like harvesting, agriculture Production and livestock.**

ID3 algorithms of data mining have been applied on APP for generating Decision Tree and Optimal Tree for optimum allocation of funds. ID3 algorithm helps to obtain Classes, Entropy and information gain.

Keywords: Agriculture Portfolio Problem, ID3 algorithm, Entropy, Information Gain, Decision tree.

I. INTRODUCTION

Agriculture is one of the globally competitive fields. The Agricultural Financing Portfolio is affiliated to the Investment Authority. An Investment Authority gives funds under Agriculture portfolio in order to provide loans to the farmers for agriculture. Banks will decide proper way to utilize funds and it will be given to the farmers under the heading of agriculture loan with subsidized rate of interest. While providing loans to the farmer risk is always there and it may be environmental risk or market risk.

The risk depends on cyclicality of revenues and earnings, economic recession, currency fluctuations, changing consumer tastes, economic health of consumers, extensive competition, weather conditions, quotas and governmental regulation and subsidies.

The banks finance various activities like agricultural production, raising cattle and poultry, purchase of machines and tools, establishing greenhouses, digging and rehabilitating irrigation canals, establishing, repairing and rehabilitating buildings, restoration of farms, digging wells, and other similar purposes.

Types of Loans are 1) Short term loans repayable within one year. 2) Long term loans with repayments not exceeding 10 years.

II. PROBLEM DEFINITION

The structure and conduct of agricultural lending has been changing rather dramatically over the past two decades. Rapid changes are occurring in technology embodied in inputs and management of resources and the environment.

With this changing face of lending, the decision making process has becoming much more complex . Quality management and risk management issues have cropped up.

The Basel II Capital Accords, scheduled to be implemented by the end of 2009, has implications for setting capital requirements, supervisory review, and market discipline at banking institutions. The measurement and management of credit risk, operational risk, and market risk lie at the heart of Basel II. While implementation will begin at the nation's largest banks, the more advanced approaches to calculating capital requirements and other management practices will have implications for other banks and non-bank lending institutions as well.

Traditionally, most financial institutions relied virtually exclusively on subjective analysis or the so-called banker expert system to assess the credit risk of borrowers. Bank loan officers used information on various borrower characteristics, which are called as the "5 Cs" of credit. They are (1) character of borrower (reputation), (2) capital (leverage), (3) capacity (volatility of earnings), (4) collateral, and (5) condition (macroeconomic cycle). However, this method may be inconsistent if it risk weights are also based on expert's opinion. The weights should be grounded based on the historical experiences. Accordingly, we have followed a statistical model approach which takes care of "5 Cs" subjectively and produce consistent forecast about the borrower's default probability. Bank can use such credit rating tool in the loan processing, credit monitoring, loan pricing, management decision-making, and in calculating inputs (Probability of default, loss given default, default correlation and risk contribution etc.) for portfolio credit risk model. The objective of this empirical research is to develop a credit risk model for an agricultural loan portfolio in India. This model takes into account the characteristics of the agricultural sector, attributes of agricultural loans and borrowers, and restrictions faced by commercial banks. The proposed model is also consistent with Basel II, including consideration given to forecasting accuracy

and applicability. We also suggest how such model would help the Indian Banks to mitigate risk in Agricultural lending.

III. Literature Survey

Arindam Bandyopadhyay (2007) have developed a credit scoring model for agricultural loan portfolio of a large Public Sector Bank in India and suggest how such model would help the Bank to mitigate risk in Agricultural lending. In this study, He has shown , with the help of a logistic model, how agricultural exposures are typically can be managed on a portfolio basis which will not only enable the bank to diversify the risk and optimize the profit in the business, but also will strengthen banker borrower relationship and enables the bank to expand its reach to farmers because of transparency in loan decision making process.

David McG. Squire (2004) have stated that the weather problem is a toy data set which we will use to understand how a decision tree is built. It comes from Quinlan (1986), a paper which discusses the ID3 algorithm introduced in Quinlan (1979). It is reproduced with slight modifications in Witten and Frank (1999), and concerns the conditions under which some hypothetical outdoor game may be played.

Markus Schmidt (2002) have documented "NUS Masterlist" containing 260 neglected and underutilized species, based on farmer interviews and literature review, and have given adequate analysis of the relevant economic scenario. After an initial pre-selection we conducted a trans-disciplinary The priority NUS selected in their study and the recommendations for their improved sustainable use, should help scientists to focus on the R&D of these species to overcome the current lack of knowledge, it should also help policy makers to enable suitable policy measures on the species level, and it can guide farmers to use alternative crop species to diversify the *agricultural portfolio* to increase food security, open up opportunities for income generation and market opportunities for farmers in China and southeast Asia.

Wei Peng, Juhua Chen and Haiping Zhou (Anonymous) have concluded that Decision tree learning algorithm can be successfully used in expert systems in capturing knowledge. They examine the decision tree learning algorithm ID3 and implement this algorithm using Java programming. First implemented basic ID3 in which dealt with the target function that has discrete output values and also extend the domain of ID3 to real-valued output, such as numeric data and discrete outcome rather than simply Boolean value. The Java applet provided at last section offers a simulation of decision-tree learning algorithm in various situations. Some shortcomings are discussed in this paper as well.

Carl Kingsford & Steven L Salzberg (2008) have observed that many scientific problems entail labeling data items with one of a given, finite set of classes based on features of the data items. Decision trees, such as C4.5, CART2 and newer variants, are classifiers that predict class labels for data items. They are then applied to classify previously unseen examples. If trained on high-quality data, decision trees can make very accurate predictions.

IV. ID3 Algorithm

For assumed training data, ID3 algorithm generates a decision tree (popular classifier). The decision tree is generated on the basis of calculated Entropy and Information gain of assumed training data. Using decision tree we can easily generate optimal tree in order to predict classification of new (unseen) training data.

The steps of computation are well documented in standard texts of Data mining.

V. Results Generated Through ID3 Algorithm

A. Data Provided

Here Indian Investment Authority gives some Units (Money) to the Bank under Agriculture Portfolio for giving loans. Here we assume that the given fund is 100Units. Bank utilizes these units for giving Loans to the Farmers. The amount of loan is decided on the basis of market risk, Duration and Interest rate. Final sum of loan distribution should not exceed than 100 Units. The fund would be distributed in fields like Agriculture production, livestock, harvesting etc. Here we assume that the fund allocated for each field is 100 Units. Bank utilizes these units for giving Loans to the Farmers in respective fields. The amount of loan is decided on the basis of market risk, Duration and Interest rate. Final sum of loan distribution in each field should not exceed than 100 Units.

B. ID3 Steps for APP

1) Identify input & output parameters

Inputs: are nothing but the parameters on which the amount of loan is based.

1. Market Risk (R):

R is categorized under three types: High, Marginal, Low.

2. Duaration (D):

D is nothing but the duration for which the bank is giving loan to the farmers.

Categories: Minimum- 0 to 6 months
Average- 7 to12 months
Maximum-13 to18 months

3. Interest Rate (I):

Categories: Minimum- 0% to 2.5%.
Average-2.6% to 5%.
Maximum- 5.1% to 7.5%.

2) Identification of classes (Subcategories of output)

Output: is a decision parameter which gives decision about
amount of loan to be given.
Here classes are-
Investment / Loan (L):
Low – 0 to 35 Units.
Marginal- 36 to 70 Units.
High-71 to 100 Units.

TABLE I: ASSUMED TRAINING DATA SET

Duration (D)	Interest Rate (I)	Market Risk (R)	Investment / Loan (L)
Maximum	Minimum	High	Low
Marginal	Minimum	High	Low
Maximum	Maximum	Low	High
Minimum	Minimum	Marginal	Low
Maximum	Maximum	Marginal	Marginal
Marginal	Maximum	Marginal	Marginal
Maximum	Marginal	Marginal	Marginal
Marginal	Marginal	Marginal	Marginal
Maximum	Maximum	Low	High
Maximum	Marginal	Marginal	Marginal

3) Calculate entropy (E)

Refer Table I

4) Entropy of output parameter investment

(i.e., (O)). Computed on above information
E(Inv)=1.4855Eq (1)

5) Entropy of input parameter (i.e.E(I,O))

Entropy of Input parameters (I) w.r.t. output parameter:
Here Let, I- Interest Rate (Int)

6) Calculate Ei

Make group of Output parameters having same value of Input parameter.
I ={ I0 ,I1 ,I2 ,…….I m }
Here I = {IMax ,IMin ,I Avg }
Let, we got m groups (i.e. on the basis of different m values of Inputs)
For each i = 0 to 2
E0(IMax,Inv)= - ∑j=0 to n [Pj* log2(Pj)]
Similarly,
E1(IMin, Inv)
E2(IAvg, Inv)

7) Entropy of input parameter (Int)

E(Int,Inv) = - ∑i=0 to m [Pi * E(Ii,O)]
= - [PMax * log2 (PMax) + PMin * log2 (PMin) + PAvg * log2(PAvg)]
=0.4000Eq (2)

8) Calculate information gain (G)

Here, G(Int) = E(Inv) - E(Int, Inv)
= 1.4855 – 0.4000
From Eq(1) & Eq (2)
=1.0855Eq (3)

Similarly, Calculate G for Market Risk(R), Durationv(Dur).
G(R)= 1.0954Eq (4)
G(D)=46Eq (5)

9) Generate decision tree

From Eq(3) , Eq(4), Eq(5)
G (IRsk) i.e. Information Gain of Market Risk is highest. So,
Market Risk will be the root node. Splited nodes are shown below

Fig. 1: Root Node Classification

Now Neglect the column Market Risk and Divide the Table 1.1 such that tuples in a sub_table is having same value of Risk,
Table II Belongs to Risk= High
Table III Belongs to Risk= Marginal
Table IV Belongs to Risk= Low

TABLE II: BELONGS TO RISK= HIGH

Duration (D)	Interest Rate (I)	Investment/ Loan (L)
Maximum	Minimum	Low
Marginal	Minimum	Low

TABLE III: BELONGS TO RISK=MARGINAL

Duration (D)	Interest Rate (I)	Investment / Loan (L)
Minimum	Minimum	Low
Maximum	Maximum	Marginal
Marginal	Maximum	Marginal
Maximum	Marginal	Marginal
Marginal	Marginal	Marginal
Maximum	Marginal	Marginal

TABLE IV: BELONGS TO RISK= LOW

Duration (D)	Interest Rate (I)	Investment / Loan (L)
Maximum	Maximum	High
Maximum	Maximum	High

Apply steps a. to e. to all sub_tables recursively.

10) Decision tree as an output

Obtained Decision tree is shown below:

Fig. 2: Decision Tree

VI. CONCLUSION

This paper presents the ID3 algorithm model for Agricultural portfolio problem that we have developed based on the training sample data for the purpose of Bank. Key issues in agriculture portfolio have been discussed. Since banks are balancing risk and return characteristics among alternative opportunities, banks cannot avoid risks. Credit risk is the largest risk faced by banks even in Agricultural loans. The most important implication of this paper is the argument that agricultural exposures are typically can be managed on a portfolio basis, and many exposures in the same portfolio have similar risk characteristics. This will enable the bank to diversify the risk and optimize the profit in the business which will ultimately enable them to comply for the Basel II requirements under the advanced approach. It is important to note that: entire exercise is based on a sample data. In order to have a robust model and robust tool for mitigating risk in agricultural loan which is perused as risky, for the entire bank, one has to enlarge the data sample and include other regions into the analysis. We have used sample data, applied ID3 algorithm for generating decision tree. As a pilot study, we have tried to demonstrate how this exercise can be done and it's utility to explore and expand the scope for further research.

REFERENCES

[1] Arindam Bandyopadhyay," Credit Risk Models for Managing Bank's Agricultural Loan Portfolio, ", National Institute of Bank Management, Pune, India,, 2007.

[2] David Mc G. Squire., CSE5230 Data Mining Tutorial: The ID3 Decision Tree National Institute of Bank Management, Pune, India Algorithm, MONASH UNIVERSITY, Faculty of Information Technology, 2004.

[3] S Markus Schmidt, "Integrating and Strengthening the European Research Area, specific measures in Support of International Cooperation (INCO)", Agrofolio: Benefiting from an Improved Agricultural Portfolio in Asia; Funded by the European Commission's sixth framework programme, FP6–2002–INCO–DEV/SSA–1–026293, 2002.

[4] Wei Peng, Juhua Chen and Haiping Zhou ;" An Implementation of ID3–Decision Tree Learning Algorithm", Project of Comp 9417: Machine Learning University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, Australia

[5] Carl Kingsford & Steven L Salzberg , "What are decision trees?;" nature biotechnology volume 26 number 9 september 2008

A Qualitative Analysis of Websites Providing Agriculture Related Information

Prof. Jawed S. Khan
Assistant Professor, MCA Department,
MCE'S Allana Institute of Management Sciences Camp Pune
E-mail: jawedkhan2@gmail.com

***Abstract*—In today's Information era agriculture and agriculture related information is available in abundance on Internet. Web portals and websites are helpful to farmers and agri-business. These vary considerably in quality and services offered. However there seem to be very few studies where these are subjected to Qualitative analysis with a view to help for better design. In this article a qualitative analysis of a set of randomly selected websites has been presented. This could be very useful to designers of websites in zeroing on a pattern or discover a inherent frame work that could form the core of such websites[3]. The choice of characteristics and the metrics for the measurements are all decided on adhoc basic at present.**

More than 200 websites were considered in the analysis. They were compared on the basis of selected parameters. Categorization schemes are proposed. There are very few websites which provide multilingual facility.

Among the services E-trading services appeared to be very poor in case of agriculture. The emphasis is on advisory are seen to high. Except for Metkewar and Acharya[1] there seem to hardly any efforts in this direction by research workers. Hence the importance of this work.

Keywords: Agriculture websites, Agribusiness, Data mining, farmers.

I. INTRODUCTION

Agriculture happens to be very important sector in the world of economy. World organizations like FAO have always been striving to provide quality information services to support the sector. The amount of such information is huge on internet. These websites and Web portals are helpful to farmers and agri-businessmen, Students, researchers, government organizations and private organizations globally. However there seems to be hardly any attempts to analyze these from the web designer's point of view.

As India is an predominantly agriculture country half the population still makes its living in agriculture.[5]. Changing the economic structure of agriculture and business practices of Indian farmers requires up-to-date, easily obtainable information. Currently, information on agricultural and economic developments in India is scattered and uncoordinated. Farmers can get the improved information and services through the creative use of the Information Technology. Agricultural issues are being covered by national media like Radio, TV and Newspapers only at macro level due to time constraint. But internet can go an extra mile by providing the information round the clock in local language, too.

In a developed country like U.S.A., most of the big farmers are using the internet to get information, to communicate and for buying inputs or selling outputs. Having agricultural and economic information available through websites expands its accessibility locally, regionally, and globally, helping avoid duplication of effort and waste of scarce resources.[6].

II. ROLE OF GOVERNMENT IN INFORMATION DISSEMINATION

Few websites and portals are run by State Agriculture Universities with assistance from NIC (National Informatics Center), a few run by organizations like Fertilizer corporation of India, Indian Council of agriculture Research, affiliated institutes and Statutory Agriculture Universities(SAUs) where as some web sites are run by private organization like AgricultreInformation.com. Research-based information is available on internet on a wide range of extension subjects including: agriculture, forestry, fishing, lawn and garden, environment, public policy, economics, and water quality.

Indian Council of Agricultural Research (ICAR), New Delhi, India is an autonomous organisation under the Department of Agricultural Research and Education, Ministry of Agriculture, Government of India. The Council is the apex body for coordinating, guiding and managing research and education in agriculture including horticulture, fisheries and animal sciences in the entire country. ICAR With over 90 ICAR institutes and 45 agricultural universities spread across India. This is one of the largest national agricultural systems in the world.[8]

The project Agricultural Research Information System (ARIS) is being implemented to bring information management culture to National Agricultural Research System (NARS) so that agricultural scientist can carry out research more effectively by having systematic access to research information available in India as well as in other countries, better project management of agricultural research, and modernization of the office tools.

The basic infrastructure required for linking all ICAR institutes has already been created. The E-mail connectivity has been established to 76 out of 90 ICAR institutes by linking through dial-up including six institutes with VSAT connectivity using NICNET and ERNET services. ARIS has four information modules namely Agricultural Research Personnel Information System (ARPIS); Agricultural Research Financial Information System (ARFIS); Agricultural Research Library Information System (ARLIS) and Agricultural Research Management Information System (ARMIS).

Status of AINET: The Agricultural Research Information System (ARIS) also known as AINET at the university level, aims at making all the departments/ laboratories/ offices share various resources (database) which are either located physically within the same building or located anywhere at remote location. The objective of the AINET is to generate database at the University level and exchange technology and information between ICAR, ICAR institutions and State Agriculture Universities (SAU's)[8].

III. Online Survey Conducted

We have analyzed more than 200 websites from various countries and agencies [6,9].

A. Identification of Parameters

The portals and websites are characterized by various parameters [3]. We have identified 50 such parameters as the once which provide distinctive characteristics to agro websites [Table 1].

B. Identification of Services

Any good Agro web portal should be providing following services, with importance indicated in the order in which they appear [Table2].

It is interesting to note that consultancy, real estate and agri-tourism dominates as services. Surprisingly agribusiness related services offered are very less. Technical services like laboratory and training are almost negligible. This speaks volumes about the relative importance of services [Table3].

Consultancy Consists of Landscaping, Gardening, Integrated farming, flowering, home gardens, terrace gardens, parks, private sponsored display landscapes at public places, rural development, Agri business profitability management, health monitoring and financial monitoring.

Real Estate consists of Sale, liaising of farm land, poultry farm Tracts of land for timber, farming, hunting, and development.

TABLE 1: IMPORTANT PARAMETERS CONSIDERED IN THE SURVEY

	Design Parameters		**Generic Parameters**		**Domain Specific Parameters**		**Business Specific Parameters**		**Advanced Parameters**
1	Accessibility(1-5)	1	Relevance(Most/moderate/poor)	1	Content	1	Products(a)*	1	Multiple language support
2	Appearance(1-5)	2	frequency of updating #b	2	Purpose(1-5)	2	Business Services provided	2	Support for Mobile devices
3	User Friendliness(1-5)	3	Location(Country)	3	Profit/Non-Profit	3	E-commerce (yes/No)	3	Link to Professional / Social networking
4	Authentication*C	4	Type of organization (Govt /Non Govt)	4	Market policy facilities	4	Advertisements (yes/no)	4	Blogs/post Technology facilities
5	Secondary links	5	Professional/Social	5	Government policy Facilities	5	Trade bodies	5	Precision Farming
6	Communication facilities	6	User Registration	6	Education / Extension	6	Payment Gateways	6	Research / Extension
7	FAQ/Inquiry	7	Site map	7	Services / Downloading Facilities	7	National/International	7	live chat facilities
8	help(1-5),	8	Search Effectiveness(1-5)	8	Weather info. (yes/no)	8	Classifieds	8	Translation Facilities
9	Web technology	9	Navigation (1-5)	9	National/International	9	Business Membership	9	MOBILE phone SMS facilities
10	Search Facilities (yeas/no)	10	Discussion Forum	10	Owner/Agency	10	Statistics	10	On line Database / E-Library Facilities

*a = I) Grain, Seeds, Insecticides II) Farm Equipments/tools III) Dairy, Fish,Meat *b= Quarterly /monthly/weekly/daily *C =Digital-signature/Certificate

TABLE 2: IDENTIFIED IMPORTANT SERVICE CATEGORY

Sr. No	**Service Name**	**Sr. No**	**Service Name**	**Sr. No**	**Service Name**
1	Aquaculture	11	Dairy Cattle	21	Weather and forecasting
2	Fisheries	12	Vegetables	22	Commodity Trading
3	Crops and Seeds	13	Animals	23	Consultancy
4	Floriculture	14	Communication	24	Trade Fairs
5	Food Manufacturers/processing industry	15	Ecommerce	25	Transport & package
6	Forestry	16	Promotion/boards/ Associations	26	Agric tourism
7	Organic Farming	17	Livestock	27	Publications
8	Gardening	18	Pest Management	28	Farm Industry
9	Dairy Products	19	Grassland & pastures	29	Agric research, Resources
10	Beef Cattle/Buffalo(Bison)	20	Soil Info	30	Production and Extension

Agric tourisms it is defined most broadly, involves any agriculturally-based operation or activity that brings visitors to a farm or ranch.. It is quite famous in the world. Most of the sites are international and very few sites are Indian.

Laboratory Services consists of Agricultural laboratory services, automated environmental controls and accessories, Interface for Agriculture Biotechnologies to increase the food production Mail Order Services: Mail order is a term which describes the buying of goods or services by mail delivery. It consists various kind of farm, gardening, food, botanical personal-and home-care products, elegant gift baskets, medicinal, poly houses, tunnels and associated equipment catalogs.

Training consists of MBA program related to agriculture at National and International levels. Like MBA International Agric-Food Management, MBA program for plantation management, Rural management programs, MBA in food and agriculture business.

These categorizations will be useful the information seekers to narrow down their requirements while searching information on the net[9].

C. Sample Websites Identified in the Survey

We have identified some of following websites which are an online community comprising of buyers, sellers and technical experts in agriculture. One can have access to post topics, communicate privately with other members, respond to polls, upload to the gallery, add links to their directory, and access to many other features.

International Portal: Agriculture.com
Public Portal: Icar.com
Private Portal: AgricultureInformation.com

TABLE 3: FREQUENCY DISTRIBUTION OF AGRIC WEBSITES AS PER SERVICES

Service Name	No. of Websites	Service Name	No. of Websites
Agric Tourism	149	Expositions	30
Auctions	92	Extension	13
Certification	13	Fencing	125
Classifieds	52	Finance	71
Commodity Trading	86	Grants	10
Consultancy	588	Insurance	25
Contract Farming	13	Lab Services	2
E-Commerce	15	Economics	36
Mail Order	03	Employment	73
Markets	47	Real Estate	177
Trade Fairs	69	Training	1
Transport & Packaging	1	Weather	76

D. Categorization of Agric-websites

Table number shows our concepts of categorization of agriculture websites mainly on the basis of services provided

TABLE 4: OUR CATEGORIZATION OF AGRIC WEBSITES

Sr. No	Service Name	Sr. No	Service Name
1	Agribusiness	9	Farming
2	Agricultural science	10	Organic farming
3	Agronomy	11	Permaculture
4	Animal husbandry	12	Sustainable Agriculture
5	Extensive farming	13	Urban agriculture
6	Factory farming	14	Intensive agriculture
7	Free range	15	Agric Library and databases
8	Industrial agriculture	16	Agric Culture health

TABLE 5: CATEGORIZATION OF SERVICES AS PER AGRICULTUREINFORMATION COM

Agric Services Agric Tourism, Auctions, Certification, Classifieds, Commodity Trading, Consultancy, Contract Farming, E-Commerce, Economics, Employment	**Crops** Cereals, Fiber Crops, Field Crops, Floriculture, Forestry, Gardening, Grassland & Pastures, Horticulture, Legume Crops, Medicinal plants & herbs	**Education & Research** Aquaculture & Fisheries, Agri Business, Agri Extension, Agricultural Economics, Agricultural Engineering, Agronomy, Animal Biology, Animal Science, Bio Science, Bioinformatics	**Farm Inputs** Biological Inputs, Crop Nutrients, Crop Protection, Farm Equipment, Farm Technology, Seeds & Planting Material
Livestock Aquaculture, Animal Feed & Supplements, Animal Welfare, Associations, Auction Facilities, Beef Cattle, Buffalo (Bison), Camelids, Dairy Cattle, Donkeys & Mules	**Processing Industry** Additives, Baby Food, Baked Food, Beverages, Brokers, Canned Food, Condiments and Seasonings, Confectionary, Dairy products, De-hydrated items	**Organisations** Cooperatives, Government, International Organisations, Museums, NGOs, Non Profit Org., Political Parties, Professional Associations, Super Markets, Trade bodies	**Useful Resources** Books, Conferences, Consumer Information, Databases, Discussion Forums, Journals, Magazines, Market News, Market Research, Newspapers

E. Farmers Guide, Libraries and Databases

This website has a compilation of nearly 2000 different useful links that are continually being updated and maintained.

http://www.rural.org/favorities.html

Agriculture Databases: This website contains links to a number of useful agriculture databases.

http://www.internets.com/sagrihtm

IV. The Mobile Revolution and its Impact

The interactive mobile-friendly format are becoming popular now a days, We can keep track of the latest markets, weather and news, receive timely reminders, and participate in polls and forums. GO Mobile component is populrise by Agriculture.com at m.agriculture.com. And it's free!

Ag Poll Join in polls that will give you a snapshot of what's happening with other farmers today.

Top Talk Interact with farmers and ranchers while on the go in mobile forums.

Ag Reminder Receive timely reminders of topics that will help you manage your farm business on a daily basis

Ag Alerts Receive timely market updates, cash bids, weather conditions and special reports via e-mail or as a text message. Just log in and customize the types of alerts you'd like to receive, and the format in which you'd like to see them [6].At international level this facility is available but in India it is not available.

V. International Agencies

FAO's Agriculture Department is helping countries achieve sustainable gains in agriculture to feed a growing world population, while respecting the natural environment, protecting public health and promoting social equity. The department helps farmers to diversify food production, reduce the drudgery of farming, market their products and conserve natural resources.FAO, which is the largest information system of its kind in the world.

The ARIC is also the national focal point for the SAARC Agricultural Information Centre (SAIC). It has published several directories in addition to a half-yearly Directory of conferences, seminars, symposia and workshops in agriculture. It is being upgraded to provide on line up linking and down linking facilities to the ICAR system, and to the agricultural information system of the entire world.

Their portal by the named Agricultural Outlook is active and brings various in to brings into the public domain information used and generated by collaborative work between the Organisation for Economic Co-operation and Development (OECD) and the United Nation's Food and Agricultural Organization (FAO) in the field of agricultural marketing. The final product of this collaboration is an annual publication presenting projections and related market analysis for some fifteen agricultural products over a ten year horizon[5].

The report analyses world commodity market trends and medium term prospects for the main agricultural products. It shows how these markets are influenced by economic developments and government policies and highlights some of the risks and uncertainties that may influence market outcomes.

In addition to highlights from the outlook publication, this website also provides the database that has been used in the analytical process. Detailed supply and use balances are available, as well as domestic and international commodity prices. The database also includes the detailed commodity and trade policy information where this was used in preparing the projections as well as the main underlying trends in key macro-economic variables and population. For OECD member-countries, the data is accompanied by detailed meta-data, where for non-member countries this documentation is still under development. In most cases the data is going back to 1970 and extended to the latest year in the projections[7].

VI. Agriculture Website Templates for Web Designers

Web templates are helpful are in designing websites, they reduce lot of designing time and efforts. Agriculture web templates are useful for personal, small-business or corporate website deigning. They come with.psd,.fla,.html source code. These Agriculture website templates can easily be integrated with backend programming done in PHP, ASP,.NET, ASP.NET, Ruby On Rails, etc.[11].

VII. Conclusion

In this paper attempts has been made to provide categorization of services in agri websites. This qualitative analysis will help web designer, farmers and the research workers in the area of information science. There are very few websites which provide multilingual facility. Among the services E-trading services appeared to be very poor in case of agriculture. The emphasis is on advisory are seen to high. In India ICAR, NIC and AgricultureInformation.com are playing very important role. At International level there number of portals which are providing many agriculture related services. Mobile friendly portals are available now a days and there is tremendous scope in this area.

REFERENCES

[1] P. S. Metkewar and Acharya H. S. (Year 2006), Agro Information Systems, Pub. IBDC, Lucknow India

[2] Antonio Mucherino, Petraq j. Papajorgji,Panos M. Pardalos (2009), Data Mining in Agriculture, Pub. Springer Science USA

[3] Shneiderman, Plaisant, Cohen And Jacobs (2010), Designing the User Interface, Pub. Pearson LPE

[4] www.Wikipedia.com

[5] http://www.foa.org

[6] http://www.agriculture.com

[7] http://www.agri-outlook.org

[8] www.icar.org

[9] http://www.agricultureinformation.com/dis/agriservies

[10] http://www.websitetemplates.com/Agriculture.html

Geographic Information System (GIS) Approach for the Assessment of Groundwater Quality Mapping in and Around Industrial Area Shirur Tehsil, District Pune, Maharashtra, India

Zeenat Nissa[1], Dr. S.W. Gaikwad[2], Dr. P.G. Saptarshi[3] and Anita Gokule[4]
[1]*Research Scholar, Department of Environmental Sciences, University of Pune, Pune–411007*
[2]*P.G. Department of Geography, S.P. College, University of Pune, Pune–411007*
[3]*Ex-Head, Department of Environmental Sciences, University of Pune, Pune–411007*
[4]*Professor, Department of Chemistry, A.T College, Bhor, Pune–412206*
e-mail: zeenatunnissa@gmail.com

***Abstract*—A geochemical assessment of groundwater quality in and around industrial area, Shirur tehsil of Pune district, Maharashtra was carried out by using a hydrochemical approach with GIS technique. Physico-chemical analysis was done to assess the seasonal variation in groundwater quality. Hence a GIS based groundwater quality mapping has been carried out in the region with the help of data generated from physico-chemical analysis of samples collected from 31 sampling stations. Comparison of the concentration of the chemical constituents with WHO (world health organization) drinking water standards of 2004 and Bureau of Indian standards (BIS) shows that the results of Total Dissolved Solids, Electrical conductivity, hardness and chlorides concentrations exceed the permissible limits for drinking water in some areas of the region. ArcGIS, Surfer and Global Mapper, GIS softwares were used for generation of various thematic maps and integration to produce the groundwater quality maps. The groundwater quality maps shows fragments pictorially representing groundwater zones that are desirable and undesirable for drinking purposes.**

Keywords: Pune. GIS. Groundwater. Quality. physico-chemical parameters

I. INTRODUCTION

India has 15% of the world's population, to be sustained with only 6% of the world's water resources and 2.5% of the world's land. Both resources must, therefore, be carefully managed in a sustainable manner. Groundwater is used for domestic and industrial water supply and irrigation all over the world. In the last few decades, there has been a tremendous increase in the demand for fresh water due to rapid growth of population and the accelerated pace of industrialization. According to WHO organization, about 80% of all the diseases in human beings are caused by water. Once the groundwater is contaminated, its quality cannot be restored by stopping the pollutants from the source. It therefore becomes imperative to regularly monitor the quality of groundwater and to device ways and means to protect it (Ramakrishnaiah et al 2008). Water quality analysis is one of the more important issues in groundwater studies. The hydrogeochemical study reveals the zones and quality of water that are suitable for drinking, agricultural and industrial purposes. Further, it is possible to understand the change in quality due to rock water interaction or any type of anthropogenic influence. Groundwater often consists of seven major chemical elements- Ca^{+2}, Mg^{+2}, Na^{+1}, K^{+1}, $Cl)^{-1}$, $HCO3^{-1}$ and $SO4^{-2}$.Hence, hydrogeochemical studies can be conducted by analyzing water samples based on these components (Anbazhagan et al. 2004).

The GIS techniques have been used in the present study, basically an integrated approach, including studies of groundwater contamination, for detecting the most polluted zone of ground water in the industrial area of Shirur, district Pune. We are sure that the data mining aspects of the study will interest the data miners, and introduce to them an interesting area where data mining will find adequate applications.

II. STUDY AREA

The Shirur Tehsil is located in the north-eastern part of the Pune district. The Headquarter of the Tehsil is 68 km away from Pune. The Shirur Tehsil is the drought prone area as decided by fact finding committee (FFC,1973).

The area is situated in the river basin of three rivers Bhima, Ghod and Vel, which offer water resource for agriculture and other activities. The other facts are

rainfall in 2006-07 is 733 cm.

latitude: 18^{O} 49' N to 19^{0} 14' N
longitude:74^{o} 22' E to 75^{o} 30' E
No of villages : 111 ,
Only one urban centre : Shirur

population 6, 48,179 [census (2001)]

Industries : 21 (11 are large scale, 02 medium and 08 small scale)

total quantity of effluent generation is 1200 CMD.

The industries are mainly automobile, electrical, electronics and chemical in nature. The people of the area are wholly dependent on groundwater as the only source of drinking water and the MIDC sector since fifteen years in this area is creating severe groundwater and other environmental problems.

Fig. 1: Map Showing Study Area

III. Methodology

The study is carried out with the help of three major components: input from Topographic sheets, GIS and data collected during field visits.

A Garmin global positioning system (GPS) was used for location and elevation readings and cross-checked against topographic sheets made available by the Survey of India (SOI). These data were used to select the representative wells and hand pumps for groundwater sampling.

Each of the groundwater samples was analyzed for physico-chemical parameters such as pH, electrical conductivity, TDS, bicarbonate and calcium using standard procedures recommended by (APHA 1995). Bicarbonates and calcium were analyzed by titration method using the standard procedure as given in APHA (1995). Total dissolved solids (TDS) were measured by evaporation and calculation methods (Hem 1991). EC and pH of water samples were measured in the field immediately after the collection of samples using a portable field kit. Except pH, which is expressed as dimensionless number, all other constituents are expressed in mg/l.

IV. Use of ArcGIS

Followed by water quality analysis, thematic maps were generated and digitized using Global mapper and ArcGIS GIS software, delineation of polluted zones were carried out for drinking water quality mapping in the area.

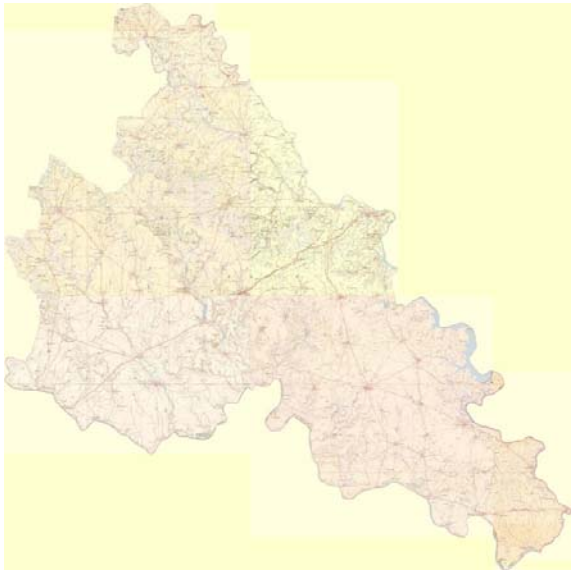


Fig. 2: Topomap of the Study Area (Shirur)

V. Results and Discussion

The quality of the groundwater samples has been analyzed (WHO 1998, standards) for drinking purposes.. It has been found that some samples show electrical conductivity, TDS and bicarbonate values above desirable limits. The values were plotted in the respective sample locations and contours were generated using the simple method of triangulation and interpolation techniques.

Water quality maps were generated for pH, TDS, electrical conductivity, chlorides and hardness in the study area showing areas falling under desirable limits and areas falling under undesirable limits. Integrating groundwater quality for drinking purposes can pictorially represent groundwater zones favorable for drinking purposes, Prioritization of zones on the basis of quality for drinking can be used for the planning and preservation of groundwater resources. The chemical analyses of the groundwater samples and concentrations of all parameters are presented in Fig. 1. It represents the GIS maps of these parameters which are measured and computed. It is found from the analysis, all the well water sample, almost all the parameters except pH, which is within the permissible limit of WHO (1998) in the study area are found to be high for almost all locations for pre-monsoon and determined to fall above the desirable limit of WHO specification. According to these results water can be classified as hard water.

VI. Chlorides

Chloride is a widely distributed element in all types of rocks in one or the other form. Its affinity towards sodium is high. Therefore, its concentration is high in ground waters, where the temperature is high and rainfall is less. The chloride ion is the most predominant natural form of the element chlorine and is extremely stable in water. As per WHO (1998) and Indian standards (ISI 1983) the desirable limit for chloride is 250 mg/l. For the study area it has been found that in certain locations the chloride concentration exceeds this limit for pre-monsoon samples. Chloride concentration at different locations was plotted and using the triangulation method and values were interpolated to generate contours. The contour map was digitized and imported into the GIS environment as a parameter for quality analysis. Areas with chloride concentrations above the desirable limit were delineated and differentiated from areas having values below the desirable limit (Fig. 3(a))

VII. Total Dissolved Solids

Chemical composition of groundwater is generally controlled by inputs through water/rock interaction and human activities. Variation in TDS in groundwater may be related to land use and also to pollution (Ellaway et al. 1999; Gillardet et al. 1999). Total dissolved solids (TDS) denote the various types of minerals present in water in the dissolved form. In natural waters, dissolved solids are composed of mainly carbonates, bicarbonates, chlorides, sulfate, phosphate, silica, calcium, magnesium, sodium and potassium. Concentrations of TDS are an important parameter in drinking water and other water quality standards. High values of TDS in groundwater are generally not harmful to human beings but high concentration of these may affect persons, who are suffering from kidney and heart diseases. Water containing high solids may cause laxative or constipation effects. In a majority of the water samples TDS in the study area falls within the permissible limits. However, in some locations the pre-monsoon values are above the desirable limit of WHO (1971) standards. Hence, the study area was delineated into two classes: desirable and undesirable.

VIII. Hardness

Water hardness is caused primarily by the presence of cations such as calcium and magnesium and anions such as carbonate, bicarbonate, chloride and sulfate in water. Water hardness has no known adverse effects; however, some evidence indicates its role in heart disease (Schroeder 1960). Hard water is unsuitable for domestic use. According to Sawyer and McCartly's (1967) classification for hardness, 10 samples fall under the moderately hard class and 21 samples fall under the hard class for post-monsoon water samples. The hardness values for the study area are found to be high for almost all locations for pre-monsoon and post-monsoon water samples and determined to fall above the desirable limit of WHO's specification (WHO 1971). Hence, the Indian standard given by Indian Standard Institution (ISI 1983) is taken into consideration for this parameter where the desirable limit is higher. The desirable limit for WHO is 100 mg/l (WHO 1971) while the Indian standard is 300 mg/l (ISI 1983). Contours were generated using the same procedure as for chloride to delineate areas of desirable hardness value from areas with the undesirable hardness value (Fig. 3(c)).

Fig. 3 GIS Maps of chemical parameters showing concentration of pollution.

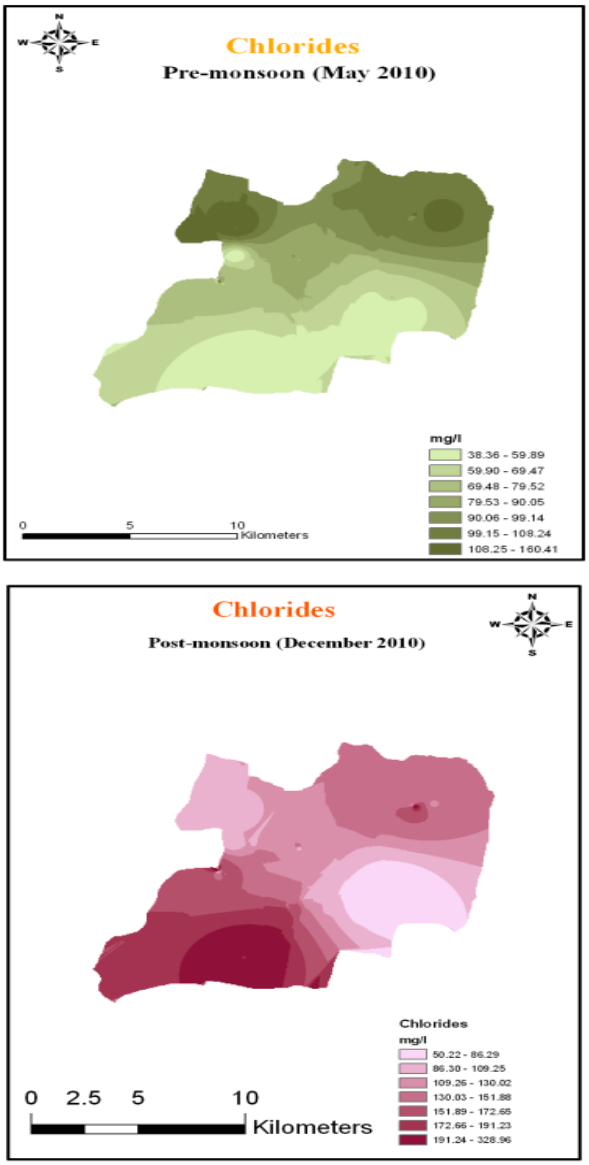
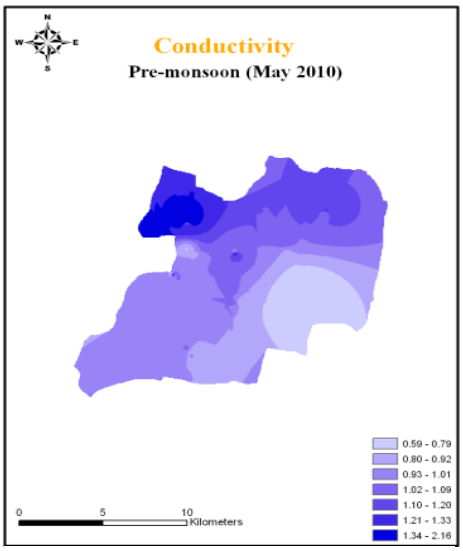


Fig. 3: (a) Map Showing Chloride Concentration in Pre-monsoon (May 2010) and Post-Monsoon (December 2010) Season



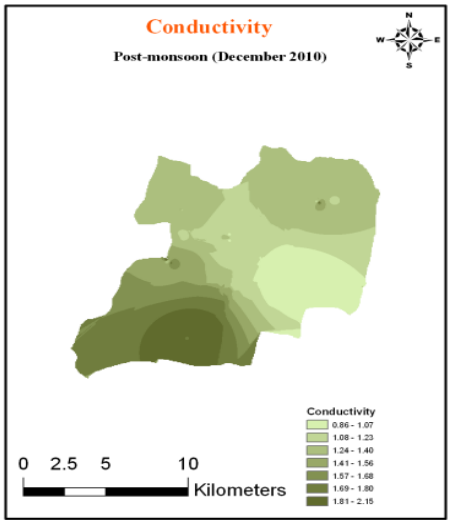
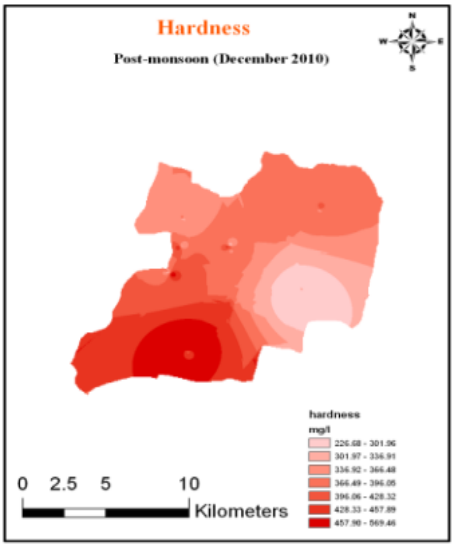


Fig. 3: (b) Map Showing Electrical Conductivity Concentration in Pre-monsoon (May 2010) and Post-Monsoon (December 2010) in the Study Area



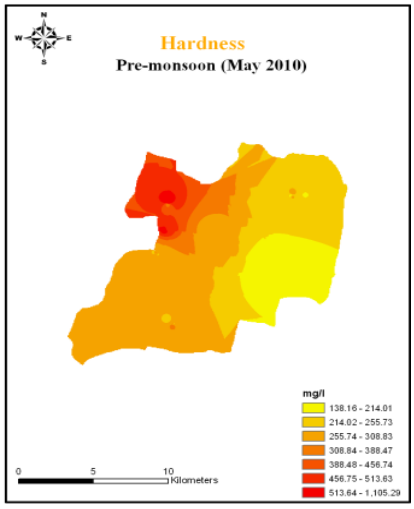
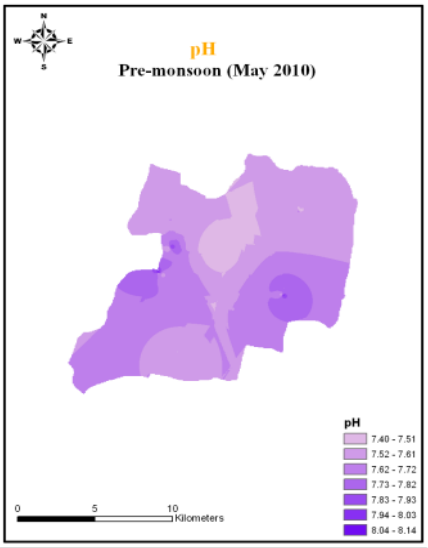


Fig. 3: (c) Map Showing Hardness Concentration in Pre-monsoon (May 2010) and Post-Monsoon (December 2010) in the Study Area



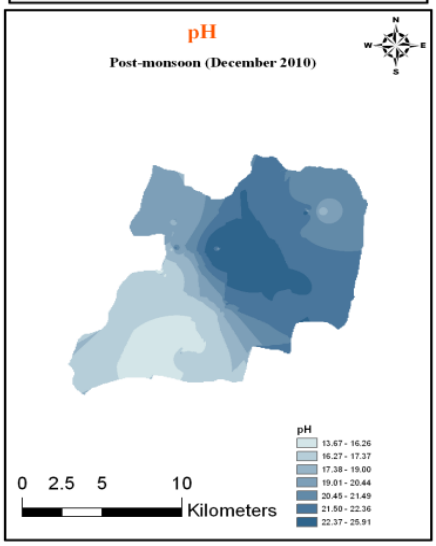


Fig. 3: (d) Map Showing pH Concentration in Pre-monsoon (May 2010) and Post-Monsoon (December 2010) in the Study Area

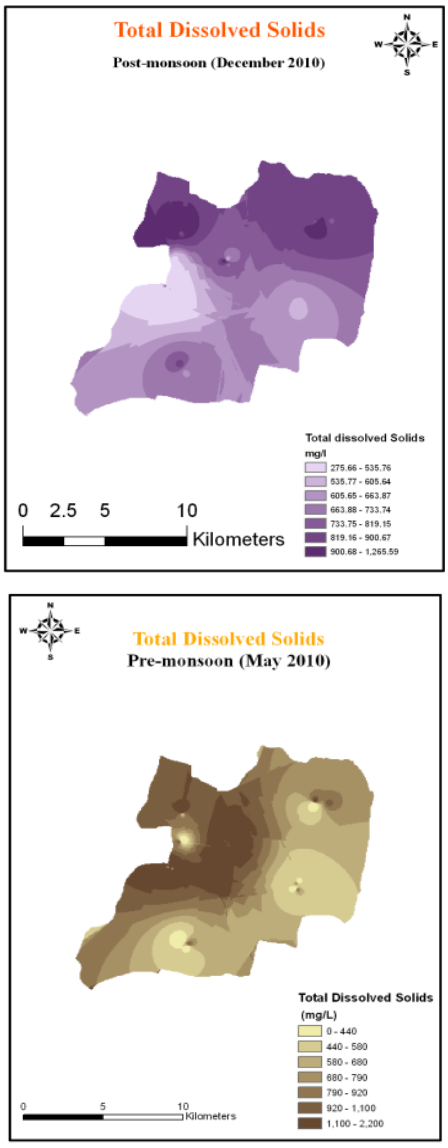


Fig. 3: (e) Map Showing Total Dissolved Solids Concentration in Pre-monsoon (May 2010) and Post-Monsoon (December 2010) in the Study Area

IX. Conclusions

In the present study, the GIS technique has successfully demonstrated its capability in groundwater quality mapping of the industrial area, shirur tehsil, district Pune, Maharashtra. ArcGIS, Surfer and Global Mapper-10, GIS softwares were used for generation of various thematic maps and integration to produce the groundwater quality maps in the study area. The final output has given the pictorial representation of groundwater quality suitable or unsuitable for drinking purposes in the area. Comparison of concentration of the chemical constituents with WHO (world health organization) drinking water standards of 2004 and Bureau of Indian standards (BIS) shows that the results of Total Dissolved Solids, chlorides and Total Hardness concentrations exceed the permissible limits for drinking water in some areas of the region. From the hydrogeochemical analysis, it is inferred that the excess concentration of chloride, TDS and hardness at some locations has determined an undesirable quality for drinking purposes. Similarly, considerable areas in the area are having high salinity hazards. Such zones require special care. The reasons for excess concentration of various elements and salinity levels require further detailed investigation. Thus GIS technique is the way to interpret the area to show fragments pictorially representing groundwater zones that are desirable and undesirable for drinking purposes.

References

[1] APHA, AWWA and WPCF, 1975. standard methods for the examination of water and waste water, 14th edition, American public health association, Washington D.C.:1193pp.

[2] WHO (1998) Guidelines for drinking water quality. Addendum to vol 2, 2nd edn. Health criteria and other supporting information (WHO/EOS/98.1), World Health Organization, Geneva.

[3] CPCB (2008), Status of groundwater quality in India, central pollution control board, Delhi, October 2008.

[4] Anbazhagan S, Archana MN (2004) Geographic Information System and groundwater quality mapping in Panvel Basin, Maharashtra, India. Environmental Geology 45:753–761.

[5] Ramakrishnaiah CR, Sadashivaiah C and Ranganna G (2008) Assessment of Water Quality Index for the Groundwater in Tumkur Taluk, Karnataka State, India. E-Journal of Chemistry 6(2) 523-530.

[6] Shrikant DL (2010) Review: Groundwater development and management in the Deccan Traps (basalts) of western India 18: 543–558.

[7] Hem JD (1991) Study and interpretation of the chemical characteristics of natural water, 3rd edn. Book 2254, Scientific Publ. Jodhpur, India.

[8] CGWB and CPCB (2000), Status of Ground Water Quality and Pollution Aspects in NCT- Delhi, January 2000.

[9] CGWB (2000), Ground Water in Urban Environment of India, Central Ground Water Board, Faridabad, December 2000.

[10] Ahn H-I, Chon H-T (1999) Assessment of groundwater contamination using geographic information systems. Environ Geochem Health 21:273–289.

[11] Elhatip H, Afsin M, Kuscu I, Dirik K, Kurmac Y, Kavurmaci M (2003) Influence of human activities and agricultural on groundwater quality of Kayseri-Incesu-Dokuzpinar springs, central Anatolian part of Turkey. Environ Geol 44:490–494.

[12] Melian R, Myrlian N, Gouriev A, Moraru C, Radstake F (1999) Groundwater quality and rural drinking-water supplies in the Republic of Moldova. Hydrogeol J 7:188–196

[13] Singh KP, Malik A, Mohan D, Singh VK, Sinha S (2006) Evaluation of groundwater quality in northern Indo-Gangetic alluvium region. Environ Monitor Assess 112:211–230

[14] Umar R, Ahmad MS (2000) Groundwater quality in parts of central Ganga Basin, India. Environ Geol (Cases and Solutions) 39(6)

Retail Management and Relationship Management in Agriculture

Prof. Akabarsaheb B. Nadaf and Abhijit Kadam
Bharati Vidyapeeth Deemed University, Pune
Institute of Management and Social Sciences, Bijapur Road, Solapur–413004(Maharashtra, India)
e-mail: nadafab@yahoo.com

***Abstract*—Retailing task in agriculture sector has been a prominent issue in the economic position of the farmers as well as the retailers. The retail outlets are active throughout the year and are responsible for family budgets of individuals. The prices of agriculture produce are highly dependent on the relationship between the middlemen and the retailers i.e. vegetable sellers. The research explains the behavior pattern of the middlemen and the retailers when the farmers come to market for selling their agriculture produce.**

The study is carried out at Solapur market yard where the questionnaires are filled up by two of the channel members. A model is proposed where Information Technology helps to dissolve the problems of existing distribution channels of agriculture produce. The suggested model will be useful in providing good prices to the farmers.

Keywords: E-Chaupal, Distribution system,

I. Introduction

A. Distribution Channels of Agricultural Marketing

Indian farmers are still following the old distribution channels for selling their produces. There are basically three channels

Channel I Farmer-Retailer—Consumer

Channel II Farmer-Markey yard(Apmc)-Retailer-consumer

Channel III Farmer-govt. Controlled market (govt Procuement)

Channel I and II involve middlemen hence result in high prices for consumers. In the first the farmers sell their produce through large retailers .

The third channel is completely controlled by the government and the producer is assured of minimum price. Further commodity is sold to the common people at subsidized rates.

II. Literature Overview

The E-Choupal has made significant development in supplying the information regarding the prices of the agriculture produce as on today to the farmers. This has been helping the farmers in getting at least minimum prices for their produce. SMS services of mobile service providers have been playing a key role in it. APMC's are also helping the farmers in bringing the buyers and the producers together in a place .This has been an assured place for the farmers for getting the buyers for their produce. Shetkari Bazzars are also emerged resulting in giving good prices to the farmers for their produce. These efforts are still not sufficient as there is a large flexibility in the supply, demand and the prices. The farmers are facing major difficulties as compared to other channel members.

B. Need of Data Integration and Abstraction in Agriculture Produce Distribution Channel

In absence of an effective information system support, farmers are dependent on information derived from their relation with the middlemen.

Objectives and Methodology of the study

1. To study the relationship of the farmers and the middlemen
2. To study the impact of relationship of the middlemen and the retailers
3. To propose a computerised system where the retailers and the farmers are brought together for selling and buying the agriculture produce.

The study is carried out in a market yard located in Solapur City of Maharashtra state which is assumed as the representative of all other market yards.

The primary data is collected on the basis of the questionnaire prepared for the farmers and the retailers. The sample size is taken as 50 farmer and 40 retailers on the basis of non-probability convenient sampling technique

C. Role of Middleman

The Agriculture Produce Marketing Committee gives all the permissions to the middlemen in making the selling of agriculture produce of the farmers. They have their shops in the Market Yard and hence are providing some of the basic facilities to the farmers. The retailers in one of the distribution channel of the agriculture produce are the vegetable sellers who generally keep special type of relationships with the middlemen. The middlemen plays very interesting roles such as

- Auctioning the agriculture produce on behalf of the farmers .
- Creating the awareness of auctioning amongst the buyers, generally the retailers
- Keeping the auction records and submitting it to the authorities of APMC
- Providing final bills and making payment to the farmers for their produce
- Fixing the base price of the agriculture produce depending upon the quality of the agriculture produce.
- Providing temporary funds to the retailers (small credits) on the basis of their relations with the retailers.

III. Role of the Retailers

The retailers (in this case the vegetable sellers) are the second last channel members and hence are playing the important roe in fixing the prices of the agriculture produce for the end consumers. Their major roles are :

- Buying the agriculture produce through auctioning process carried out by the middlemen.
- Making the payment immediately after buying the produce
- Transporting the produce to the point of selling, generally small market places
- Selling the produce to the end consumers by fixing the prices.

IV. Hypothesis

- Most of the farmer prefer sticking to particular middlemen for selling their produce.
- Most of the farmer give weight age to instant payment over relationship with the middleman

V. Data Collection and Analysis

TABLE 1: THE FLEXIBILITY OF THE FARMERS IN SELECTING THE MIDDLEMEN

Selection of middlemen	No. of farmers	% of farmers
Particular Middlemen	44	88
Other middlemen	06	12

Most of the farmers are sticking up to the particular type of middlemen every time they come to the market yard for selling their produce. Few of the farmers believe in changing the middlemen every time they come to the market yard.

TABLE 2: THE REASONS FOR SELECTING A PARTICULAR TYPE OF MIDDLEMEN

Reasons	No. of farmers	% farmers
Relations with particular middlemen	10	20%
Good base prices in auctioning process	14	28%
Instant payment facility	40	80%
Familiar Middlemen	28	56%

80% of the farmers choose particular middlemen as they get instant payment form them. 20 % of the farmers give weightage to relationship with the middlemen as one of the factor to go to the particular middlemen.

TABLE 3: THE FLEXIBILITY OF RETAILERS IN SELECTING THE MIDDLEMEN

Selection of middlemen	No. of retailers	% of retailers
Particular Middlemen	35	87.5`
Other middlemen at random	05	12.5

87.5 % of the retailers prefer selecting a particular middlemen where as 12.5 % of the retailers go to other middlemen on random basis.

TABLE 4: THE REASONS FOR BUYING AGRICULTURE PRODUCE FROM A PARTICULAR MIDDLEMEN BY THE RETAILERS (I.E. VEGETABLE SELLERS)

Reasons	No. of Retailers	% Retailers
Relations with particular middlemen (prominent middlemen)	26	65
low prices during auctioning process	35	87.5
credit facility with the middlemen	15	37.5
quality agriculture produce	12	30
No specific reason	02	05

The table shows that more than $3/4^{th}$ of the retailers buy agriculture produce from particular middlemen. 87.5 % of the retailers think that the middlemen are providing the produce at lower prices to them. Availability of credit facility with the middlemen attracts 37.5 % of the retailers. Only 30 % of the retailers prefer specific middlemen on the basis of availability of good quality of agriculture produce.

D. Major Findings

- More number of farmers prefer to sell their produce through a particular middlemen as they get money instantly. It is one of the reason that the farmers are being trapped by the middlemen.
- The farmers do not prefer a specific middlemen on the basis of their relationship with them.
- It is clear from the content of the tabular data that the relationship between the middlemen and the retailers plays an important role in fixing the prices of produce .
- Maximum number of retailers buys produce from particular middlemen as they get the produce at lower prices.
- Credit facility by the middlemen to the retailers also play an important role in auction process.
- The current distribution channel involves expenses of the middlemen to be borne by the farmers

E. Suggested Distribution System Model

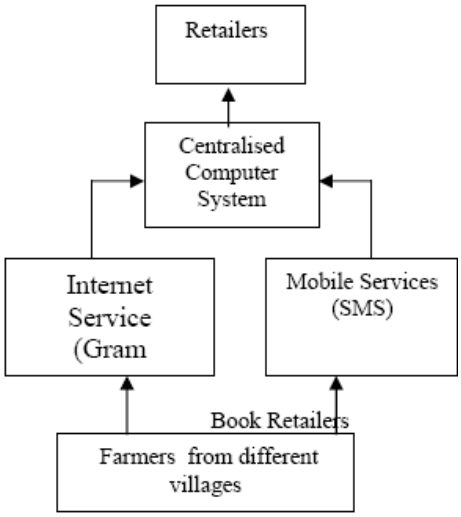


Fig. 1: Suggested Distribution System

F. Working of the System

The farmers will book the retailer on the basis of the capacity of the retailer to buy the quantity either by making use of SMS or by making use a tiny facilitating center. .The capacity is already stored in the centralized computer system

The Gram Panchyat Samiti's Office will provide on line retailer booking facility to the farmers through the internet Charging a trifling amount.

G. Centralised Computer System

The central computerised system will manage all the operations such as :

- Keeping the details of the registered retailers along with their daily requirement of different agriculture produce and the place of buying
- Keeping the details of the registered farmers and helping them in booking the retailers randomly and the place of selling
- Managing the data supplied through SMS .
- Booking the retailer on the basis of his requirement and the supply available with the particular farmer.
- It can be managed by the agriculture department of the government

H. Retailers

- The farmers should always give preference to the middlemen who are starting the auction with good base price..
- The farmers should not prefer particular middlemen in order to get instant payment as all other middlemen will also make the payment within a day.
- The middlemen should give at least good base prices during auction to the farmers on the basis of the quality .

VI. Advantages of the Model Suggested

The model will help in resolving the problem of middlemen commission, relationship with the middlemen and the fixing the base price by the middlemen which is generally low. The farmers will be having rights to fix the base price.

VII. Conclusion

It has been noticed that the relationship between the channel members affects the agriculture produce prices. The farmers are worrying about their payment and hence are losing the profit at a greater extent. The retailers gain through the auction process due to close relationship between the middlemen and the retailers. There is a need of creating awareness amongst the farmers in understanding the market trends. The farmers are to be assured for getting good prices irrespective of the kind of relationship they have with the middlemen. The model suggested will resolve most of the problems of the farmers and will be useful if implemented effectively.

References

[1] Acharya, S. S. and Agarwal, N. L (1999) . "Agriculture, Policies Marketing in India" Oxford & IBH publishing Co. Pvt. Ltd.; New Delhi

[2] Advances in Agricultutre(1996), Published by Agricultural Officers Association, Govt. of Goa , Assistant Office of the Registrar of Co-perative Societies, Annual Administrative Report on Agricultural Marketing from the period 1986–87 to 2005–06.

[3] Gupta S. P. and Rathere N. S. (1998). Marketing of Vegetables in Raipur District of Chhattisgarh State. An Economic Analysis. Indian Journal of Agricultural Economics 53 (3): 393.

[4] K.N.Selvaray and K.R.Sundaravradarajan(Oct-Dec1998) Performance and attitude towards regulated marketing in Tamil Nadu, published by "The Bihar Journal of Agricultural Marketing", Pant Bhavan , Patana.Vol.VI No.4.

[5] Sankaran, S. "28". Indian Economy: Problem sand Development. pp. 492–493.

[6] Sengupta, Somini (22 June 2008). "The Food Chain in Fertile India, Growth Outstrips Agriculture"

[7] Subrahmanyam K V (1986☺Economics of production and marketing of chrysanthemum flowers in Karnataka. Indian Journal of Agricultural Economics 41 (3): 286.

[8] Thakur D.S. and A.S.Shandil (1993), Steps to increase market arrivals and efficiency of regulated market."The Bihar Journal of Agricultural Marketing.

[9] Wani, M. H.; Mattoo, M. S. and Sofi, A. A.(1995); Resource use and economic efficiency of various marketing cost components in apple; Agricultural Marketing; 37(4) PP 38–40.

[10] http://www.isapindia.org/
http://www.indiatogether.org/agriculture/

Availability and Delivery of Health Related Information on the Internet in Different Medical Streams: A Quantitative and Qualitative Analysis

Sonal Khosla[1] and H.S. Acharya[2]
[1]*Research Scholar, Symbiosis International (Deemed University), Pune, India*
[2]*Professor, Allana Institute of Management Sciences, Pune*
e-mail: sonal.khosla@rediffmail.com, haridas.undri@gmail.com

***Abstract*—More than 70,000 websites disseminate health information. The most common way to access these information is through search engines. Although specialized search engines are available for specific searches like videos, images, web and search related to a particular domain, it was a general observation by the 2002 Pew Internet and American Life Project Poll that a consumer usually starts search at a generic search engine rather than a specialized one[2]. The efficiency of a search engine to find a website and index them depends on factors like coding used and the structure of a website, search keywords being used, the use of paid placements by the search engine and the criteria used by a search engine. Hence the results obtained through search engines vary a lot. In the current paper we attempt to study how generic search engines are able to disseminate health information to professionals and consumers.**

An experiment was performed by employing free text queries on search engines using different combinations of keywords, browsers and search engines. The hit statistics were noted for various medical streams like Ayurveda, Allopathy, Homeopathy, Unani, Sidha etc. The result was then analyzed using ANOVA and other data analysis tools. The primary objective was to test whether the information obtained is dependent on the choice of the browsers and search engines and how. The study has also attempted to analyze the trend of the figures obtained through various charts.

The analysis of the secondary data available for different medical streams on the Internet leads to discovery of a relationship between the search engine and the medical stream dominant on it. Thus we have been able to study the factors that influence the online health related searches and develop a quantitative and qualitative analysis and set a basis for future research and development to improve the results of health related searches.

Keywords: Internet, Ayurveda, Homeopathy, Allopathy, Sidha, Unani, Co-occurrence rate

I. INTRODUCTION

Many consumers are seeking health related information on the Internet as a first guide before consulting specialized physicians. In the olden days people would depend on experience of elders, neighbors and articles in newspapers and health magazines for such information and subsequent decision making. According to ComScore the number of people using the Internet has increased by 10% reaching 747 million in 2007. The most common way of accessing such information is through search engines rather than visiting websites directly. Very less of the information would have been available to consumers without the use of Information Retrieval search engines [5]. According to a study it was found that eight million American adults look for health information online on a typical day [9]. The study also says that a typical health information search starts at a search engine. 66% of the visitors began their online health query at a search engine and 27% at a health related website [9]. It shows that a success of an Internet searchers' search has a lot of confidence on search engines [9]. Now which website a consumer might access depends more on the efficiency of the search engine rather than the accuracy or reliability of the website or the knowledge of the existence of any such website. Hundreds of people search the web daily to gather information among which Google, Yahoo, MSN, Ask and AOL account for the majority of the searches [6]. There are also specialized search engines providing information about a particular domain as well. Some examples of these "vertical" search engines include: Healthline.com, Healia.com, Kosmix.com, Mammahealth.com, and Medstory.com. These search engines have been developed in such a way that it has built in knowledge of medical terminology and hence filters out the irrelevant data [7].

II. WEB SEARCH

A Web search is a text based information-retrieval system that searches the web and creates an indexed and ranked list of web pages. A Web search engine has four components: Web Crawling, Indexing, Querying and Ranking [7]. Web crawlers or spiders visits the web

pages, reads them and returns a link to these web pages to the master database of the search engines. An indexed database is created as a second step. It contains a copy of every web page found by its spiders/crawlers. Next step is querying the database to retrieve records relevant to the user search and ranks the pages based on their relevance [10]. Some search engines also use Boolean operators to specify the search query. Every search engine works differently. Page titles, body of the web page, Meta tags and other elements play a vital role in deciding the relevancy and ranking of each page. Some search engines also use intelligent agents to refine their search [10]. Some search engines like Google store all or a part of the web page in their master index while on the other hand Altavista store every word of the page they find.

The current study aims at comparing the retrieval performance of two-term search efficiency in Google, Altavista, Bing and AOL. Differential responses of the engines to search attempts specific to different medical steams like Ayurveda, Homeopathy etc have been used to indirectly quantify information availability in each of the streams.

III. Methods

A very simple experimental method on the lines of method suggested by Eisenach and Kohler [3] was used to gather basic data for the study. The schema is as shown in Table 1. Different Keywords were chosen as a part of our study based on different medical streams. In all, searches were performed using 26 keywords: Health, Health Ayurveda, Ayurveda, Health homeopathy, Homeopathy, Allopathy, Health Allopathy, Health Sidha, Sidha, Health Unani, Unani, Health education, Education, Health information, Medical Education, Medical Diagnostic software, Diagnostic Software, Cancer health, Cancer, virtual hospitals, hospitals and online medical services. The hit statistics were also observed for exact occurrence of keywords:"health wiki", "medicine wiki","virtual hospitals" and "sharable medical databases".

TABLE 1: The Data Model

Col. No	Fact/Dimension	Name	Options/ Variable Type
1	DIMENSION_1	Browser	Internet Explorer, Mozilla Fire fox, Google Crome
2	DIMENSION_2	Search Engine	Google, Yahoo, Bing, AOL
3	DIMENSION_3	Medicine	Alopathy, Ayurveda, Homeopathy, Unani, Sidha, General
4	DIMENSION_4	Time Span Set	Past week, Past Month, Past year, Free
5	DIMENSIONA_5	Language	English
6		Key Word Input	Words chosen for varied attempts of search.
7	FACT	Search Results	Integer count

The search attempts, performed on combinations of 4 different search engines , 3 browsers and 4 time span settings, using above keywords resulted in 625 rows of data, organized in flat multidimensional data base. Organization of the data, slicing and dicing requirements were all handled using DataPilot of Open Office Org. Analytical tools used were restricted to what is provided in the OOoStat package, an open source package which is an 'Add **on' to** the Open Office Calc. Searches resulting in no results were summarized in a separate group. These queries indicate that the content was out of scope and is a problem of content coverage which shows that either the content was not available or there is a mismatch between the user's expectations and the actual fact [4]. This could also be a limitation of the search engine. But when the search was performed with the keyword "sharable medical database" on three different search engines, it resulted in no results. There could be three reasons for the same: content coverage, systems functionality and user query formulation. It was also observed that although data was available for some content the result varied for different search engines.

A. Single Dimension Responses

1) Search engine response

A one-way ANOVA, using F-test was done to test the variability in the results obtained for different search engines. The F-value (Table 2.) observed fell in the critical region, and hence we had to reject the Null hypothesis, concluding that highly significant difference could be seen (at level of significance $p = 0.01$) between the search engines. Search results from Yahoo and Bing dominated and were at par, with Google and AOL giving significantly lower results.

TABLE II: One-Way Anova with Search Engines as the Dimension.

Groups	*Average*
AOL	3882.75
Google	33700.25
Yahoo	24650000
Bing	24291667

ANOVA

Source of Variation	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3	2.39E+15	6.122364	0.001424	2.816466
Within Groups	44	3.91E+14			
Total	47				

2) Browser wise response

Again a one- way ANOVA was used to test the variability in the output obtained through different browsers. The results showed that there was no significant difference amongst the responses of browsers.

3) Time span wise response

Since there is no significant difference in the hit statistics with different browsers we have analyzed the hit statistics for different medical streams based on how recent the web page is. We have taken three categories: "Past Week", "Past Month" and "Past Year". The Bing search engine does not have a date feature and Yahoo does not have a feature for "Past Week" and "Past Month". To maintain consistency in the analysis, Google and AOL search engines have been compared for different medical streams.

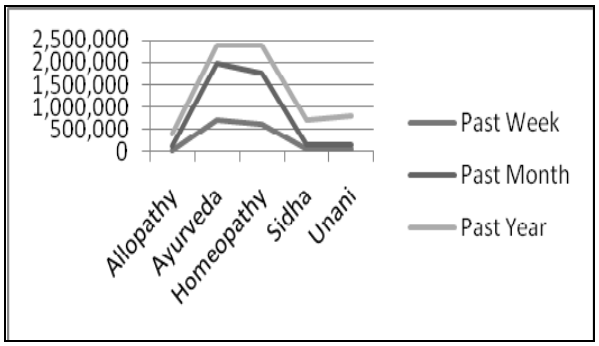


Fig. 3: Actual Search Results for Different Medical Streams on Google using the Browser Google Chrome in Various time Spans

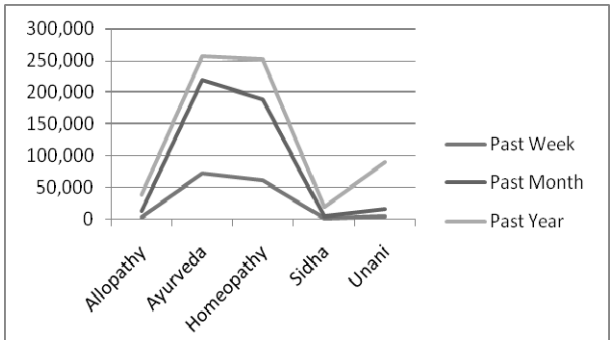


Fig. 4: Actual Search Results of Different Medical Streams on AOL using the Browser Google Chrome in Various time Spans

4) Stream wise response

Further to identify whether the statistics obtained by searching for different medical streams is actually related to health we calculated the co-occurrence rate for each stream on different search engines. The co-occurrence rate can be defined as the proportion of the web pages obtained with the search term AND the keyword "health" to the number of pages obtained with the keyword alone [3]. This metric shows how frequently the search keyword appears on the same web page with the word "health".

C = pages(keyword AND health) / pages(keyword)

Where C = co-occurrence rate

Keyword = medical stream

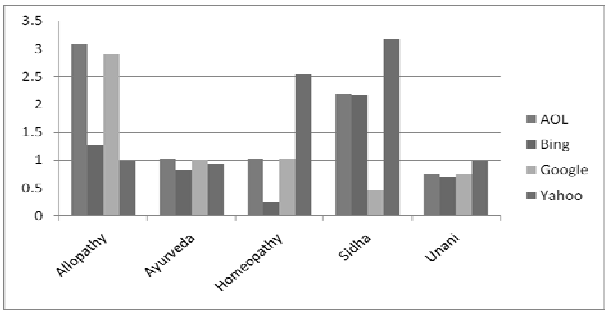


Fig. 1: Co-occurrence Rate of Different Keywords on Different Search Engines using the Browser Google Chrome in the Past on Year

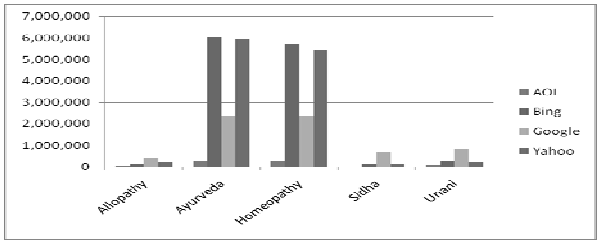


Fig. 2: Actual Search Results of Different Keywords on Different Search Engines Using the Browser Google Chrome in the Past One Year

B. *Multiple Dimension Responses*

A Two way ANOVA taking search engines and different medical streams as dimensions were also performed taking the hit statistics from the browser Google Chrome and taking the time dimension as one year. Since the F- value for rows is greater than F-crit and p-value is less that 0.05, we can say that the responses for different medical streams are significantly different. However when we see the columns the F-value is less than F-crit, we say that there is no significant difference in the responses from different search engines.

TABLE III: TWO FACTOR ANOVA

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Allopathy	4	761500	190375	2.4E+10
Ayurveda	4	14597000	3649250	8E+12
Homeopathy	4	13802000	3450500	6.86E+12
Sidha	4	1006800	251700	9.66E+10
Unani	4	1416000	354000	9.99E+10
AOL	5	655300	131060	1.34E+10
Bing	5	12279000	2455800	9.76E+12
Google	5	6668000	1333600	9.27E+11
Yahoo	5	11981000	2396200	9.1E+12

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	5.19163E+13	4	1.3E+13	5.711477	0.008229	3.259167
Columns	1.79667E+13	3	5.99E+12	2.635435	0.097567	3.490295
Error	2.72695E+13	12	2.27E+12			
Total	9.71525E+13	19				

C. Responses using Exact Occurrence of Keywords

Wiki is a content management tool that allows easy creation and editing of any number of interlinked web pages using any simplified editor like WYSIWYG text editor. WikiWikiWeb was the first Wiki created in English language. In simple terms it is a website that can be edited by any reader and thus helps in sharing knowledge or information.

Next the study has also analyzed the fact whether the Wiki's related to health and medicine has been updated and how recently. The search for the "Health Wiki" and "Medicine wiki" were done under advanced search category of the search engines and exact occurrences of the keywords only were noted for analysis. Again the data has been only taken for Google and AOL search engine using the browser Google Chrome.

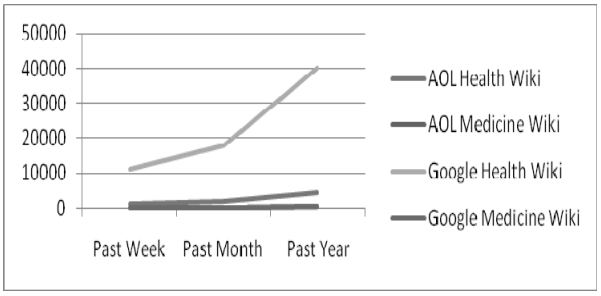


Fig. 5: Actual Search Results of "Health Wiki" and "Medicine Wiki

IV. Conclusion

Thus the study comes to a conclusion that the choice of a browser does not play any significant role while doing any health related search. But the choic of a search engine brings out significant results. It was found in the study that although the actual search results for Allopathy, Sidha and Unani were comparatively low, but they showed higher co-occurrence rate. This could be due to the high efficiency of the search engine. Also the difference could be due to the factor that all the search engines are designed and optimized differently. Also it could be seen that Ayurveda and Homeopathy websites were modified more frequently than the other streams. Even Health wiki had a higher modification value than Medicine wiki.

References

[1] Kristen Shuyler, Kristen Knight. What are patients seeking when they turn to the Internet? Qualitative Content analysis of Questions Asked by Visitors to an Orthopaedics Website. Journal of Medical Internet Research, 2004.

[2] Liza Greenberg, Guy D'Andrea, Dan Lorence. Setting the public agenda for online health search: A White Paper and Action Agenda. Journal of Medical Internet Research, 2004.

[3] G. Eysenbach, Ch. Kohler. What is the prevalence of health related searches on the World Wide Web? Qualitative and Quantitative analysis of search engine queries on the Internet. American Medical Informatics Association, 2003.

[4] Alexa T. McCray, Tony Tse. Understanding Search Failures in Consumer Health Information Systems. National Library of Medicine, Bethesda, MD.

[5] Pierre Jacquemart, Pierre Zweigenbaum, Towards a Medical Question-Answering system- A Feasibility Study.

[6] Mazin Gilbert and Junlan Feng. Speech and Language Processing over the Web: Changing the way people communicate and access information.

[7] Humbert H. Suarez, M.D Ph.D., Xiaolong Hao, Ifay F. Chang. Searching for Information on the Internet using UMLS and Medical World Search

[8] Andrea Haase, Markus Follman, Guido Skipka, Hanna Kirchener. Developing search strategies for clinical practice guidelines in SUMSearch and Google Scholar and assessing their retrieval performance.

[9] Most Internet users start at a search engine when looking for health information online. Very few check the source and the date of theInformation they find. Pew Internet and American Life Project. 202–419–4500. http://www.pewinternet.org/

[10] ServInt Free Net. http://servintfree.net/support/trainingdocs/searchengines.pdf

Control of NPA in Cooperative Banks using Data Mining Technique

Syed Azharuddin[1] and Bashir A. Hamza[2]
[1]*Associate Professor, Department of Commerce, Dr. B.A.M. University, Aurangabad, 431004, (M.S.) India*
[2]*Research Scholar, Dr.B.A.M. University, Aurangabad, 431004, (M.S.) India*
e-mail: Azharuddinsyed@ymail.com, Zakroos1@yahoo.com

***Abstract*—One of the main problems facing the banking sector today is the management and control of NPAs. Non Performing Asset means a loan or an account of borrower, which has been classified by a bank or financial institution as sub-standard, doubtful or loss asset, in accordance with the directions or guidelines relating to asset classification issued by RBI. Problem of NPAs in Cooperative Banks can be reduced by attaching it to core banking module. This paper attempts to provide a solution to reduce NPAs with the help of Data mining technique.**

Keywords: E-Commerce, NPA, Marketing, Risk management Fraud.

I. Introduction

The application of Data Mining in Banking Sector is evidenced in **Marketing:** Data mining carry various analyses on collected data to determine the consumer behavior with reference to product, price and distribution channel. **Risk Management:** Banks provide loan to its customers by verifying the various details relating to the loan such as amount of loan, lending rate, and repayment period, type of property mortgaged, demography, income, and credit history of the borrower. **Fraud detection:** Sometimes the given demographics and transaction history of the customers are likely to defraud the bank. Data mining technique helps to analyze such patterns and transactions that lead to fraud. Data mining can be very useful in controlling NPA. Controlling NPA especially in Cooperative Banks of Aurangabad can help in discourage defaults and irregularities in paying back by customers.

II. Objectives of the Study

1) To study Application of Data Mining in Banking sector
2) To study the effectiveness of Data Mining Technique in controlling NPAs in Cooperative Banks of Aurangabad.

III. Hypothesis

Data Mining can help in reducing NPAs in Cooperative Banks in Aurangabad.

IV. Methodology

- Primary data is collected through Interviews.
- Secondary data from Reports, Books and Journals.
- A 100% sample of all 21 Cooperative Banks in
- Aurangabad has been chosen for study.

The volume of business conducted electronically has grown very much due to widespread Internet usage. There are numerous innovations and transaction conducted in electronic funds transfer, internet banking, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems.

There are mounting difficulties in the growth of NPAs in Indian banking sector. New private banks' bad loans write-off grew four-fold in the last three years. On the other hand, old private banks' bad loans write-off had tripled in the last three financial years. In absolute terms, new private banks' bad loans write-off for 2009-10 stood at Rs 6,696 crore, a more than four-fold increase over Rs 1,581 crore in 2007-08. In the case of old private banks, the bad loan write-off in 2009-10 stood at Rs 1,331 crore, nearly a three-fold increase over Rs 453 crore in 2007-08. PSBs' bad loan write-offs grew 37 per cent to Rs 10,040 crore in 2009-10 from a level of Rs 7,347 crore in 2007-08.

Meanwhile, the gross NPA level of new private sector banks increased to Rs 13,772 crore in end March 2010 from a level of Rs 10,419 crore in end March 2008. The gross NPAs of old private sector banks stood at Rs 3,612 crore in end March 2010, higher than the level of Rs 2,557 crore in end March 2008.

The gross NPA level of PSBs stood at Rs 57,301 crore as of end March 2010, much higher than the level of Rs 39,749 crore in end March 2008.

The above picture is not pleasing at all. The fact is that Aurangabad banks are part of the Indian banking system and they follow the same credit policy of their head offices. It is also true that out of the massive NPAs in the country, Cooperative Banks of Aurangabad too have a large share in them. Notwithstanding other factors, a study in E-commerce implementation with special reference to use of data mining for control of

NPAs in Cooperative Banks of Aurangabad banks shall highlight the overall picture of the cooperative banking scenario in Aurangabad, its management and control of NPAs through data mining.

There are about 21 Nationalized banks, 17 private commercial banks and 21 cooperative banks, i.e. about 52 banks are operating in Aurangabad

The application of Data Mining in Banking Sector is evidenced in:

A. Marketing

Data mining carry various analyses on collected data to determine the consumer behavior with reference to product, price and distribution channel. The reaction of the customers for the existing and new products can also be known based on which banks will try to promote the product, improve quality of products and service and gain competitive advantage. Bank analysts can also analyze the past trends, determine the present demand and forecast the customer behavior of various products and services in order to grab more business opportunities and anticipate behavior patterns. Data mining technique also helps to identify profitable customers from non-profitable ones. Another major area of development in banking is Cross selling i.e banks makes an attractive offer to its customer by asking them to buy additional product or service. For example, Home loan with insurance facilities and so on. With the help of data mining technique, banks are able to analyze which products and service are availed by most of the customers in cross selling and which type of consumers prefer to purchase cross selling products and so on.

B. Risk Management

Banks provide loan to its customers by verifying the various details relating to the loan. Customers with bank for longer periods, with high income groups are likely to get loans very easily. Even though, banks are cautious while providing loan, there are chances for loan defaults by customers. Data mining technique helps to distinguish borrowers who repay loans promptly from those who don't. It also helps to predict when the borrower is at default, whether providing loan to a particular customer will result in bad loans etc. Bank executives by using Data mining technique can also analyze the behavior and reliability of the customers while selling credit cards too. It also helps to analyze whether the customer will make prompt or delay payment if the credit cards are sold to them.

C. Fraud detection

Data mining technique helps to analyze patterns and transactions that lead to fraud.

D. Customer Retention

Today in this competitive environment, customers have wide range of products and services provided by different banks. Hence, banks have to cater the needs of the customer by providing such products and services which they prefer. This will result in customer loyalty and customer retention. The importance of adequate knowledge and information in today's business is a factor not to be understated.

TABLE 1: NPA LEVAL IN COOPERATIVE BANKS OF AURANGABAD (M.S.)

S.No.	Co-operative banks	Banks using software for reference check	NPA level
1	Abhyudaya co-op. Bank Ltd.	NO	-
2	Bombay Marcantile Bank Ltd.	NO	3%
3	Development credit Bank	NO	5%
4	Janta sahkari Bank	NO	5%
5	Maharashtra sate co-op. Bank	NO	**30%**
6	Shamrao vitthal co-op. Bank	NO	2%
7	Rupee co-op. Bank	NO	**10%**
8	Akola urban co-op. Bank Ltd.	NO	6%
9	Punjab & Maharashtra co-op. Bank	NO	2%
10	Saraswat co-op. Bank	NO	0%
11	Cosmos Bank	NO	3%
12	Ka ichlkarnji janta sah Bank	NO	5%
13	S.T. co-op. Bank	NO	**10%-**
14	Jalgaon janta sahakari Bank	NO	3%
15	Mallkapur urban co-op. Bank	NO	6%
16	Sundarlal savji co-op. Bank	NO	5%
17	People co-op. Bank Ltd.	NO	7%
18	Aurangabad dist. Central co-op. Bank	NO	**30%**
19	Deogiri nagari sahakari Bank	NO	4%
20	Lok vikas nagar sahakari Bank	NO	2.62%
21	The vaidynath ur. Co-op. Bank	NO	8%

Source: Primary Data

Data mining could be the best solution to NPA management and control in banking sector. Data mining is becoming strategically important area for many businesses. It is a process of analyzing the data from various perspectives and summarizing it into valuable information. Data mining assists the banks to look for hidden pattern in a group and discover unknown relationship in the data**.** Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately

apparent to managers because the volume of data is too large or is generated too. Quickly to screen by experts. The managers of the banks may go a step further to find the sequences, episodes and periodicity of the transaction behavior of their customers which may help them in actually better segmenting, targeting, acquiring, retaining and maintaining a profitable customer base. Business Intelligence and data mining techniques can help them in identifying various classes of customers and come up with a class based product and/or pricing approach that may garner better revenue management as well. Data mining techniques helps to analyze the customers who are loyal from those who shift to other banks for better services. If the customer is shifting from his bank to another, reasons for such shifting and the last transaction performed before shifting can be known which will help the banks to perform better and retain its customers. It is in the interesting to find out and at least suggest ways as to how banks can reduce relying on reactive customer service techniques and conventional mass marketing. Especially in Cooperative Banks they should increase dependence on use knowledge and information which are the keys for solving NP As.

Performance of cooperative Banks understudy was not Satisfactory as compared to private commercial and public sector nationalizes banks. Cooperative Banks did not use the software which mines the data for reference check of prospective customers applying for loan. Hence the percentage of NPA persisted in virtually all banks and only few of the banks were found to have more than 10% limit recommended by the regularity authority. The NPA level in cooperative banks can be brought down considerably by applying the reference check method using software or rather sharing information of clients with other banks.

Hence a multi channel interface module including cooperative banks can be helpful to put brakes on the level of NPA in cooperative banks.

V. CONCLUSION

The cooperative banks are an important constituent of the Indian financial system judging by the role assigned to them, the expectations they are supposed to fulfill, their role in rural financing continuous to be important even today, and their business in the urban areas also has increased phenomenally in recent years. Cooperative banks have not been able to maintain the balance between lending and recovery hence they run into losses, this is primarily due to rise in NPA which is again due to faulty distribution of loans to costumers who are already indebted and apply for loans and even get it, because of faulty system in reference checking. Data Mining provides a solution to join cooperative banks with core banking system and help them functioning smoothly so that they earn reasonable profit and reduce non performing assets in cooperative banks.

REFERENCES

[1] Kr Srivats, *Business Line* (Business Daily from the Hindu group of publications, Aug 2010.

[2] Jaiwei Han/Micheline Kamber, *Data Mining Concept and Techniques"Integration of Data Mining System with Database"*, Morgan Kaufmann Publisher, Gurgaon, India (pg 34–40), 2009.

[3] Dr.Madan Lal Bhasin, *Data Mining: competitive Tool in The Banking and Retail Industry*," The Chartered Accountant", pg (588–594), 2006.

[4] http://www.iimahd.ernet.in/publications/data/

[5] http://www.rbi.org.in/home.aspx

[6] Kamlesh K. Bajaj Dbjan Nag, *E-commerce the cutting edge of business*, published by Tata McGraw Hill, New Delhi, India (pg 121–129), 2009.

Application of Data Mining Techniques for Journal Search Tool–A Case Study Specific to Information Search in Agriculture

N.M. Tamboli[1] H.S. Acharya[2] P.S. Metkewar[3]

***Abstract*—Due to convenient online access mechanism most of the reputed journals are available online. Though, ample amount of hard copy journal and back volumes are available in libraries, still access to these volumes is limited to name and publisher information through web opac. Essentially there is need of some mechanism to access the content of journal without its physical access. Search tools to search the required keywords in journal with help of electronic copy of content and index page serves the purpose of locating journal to certain extend. The mechanism for the said purposes is discussed. Further, In this study, applicability of text mining techniques is discussed for identification of journals based on the content.**

I. INTRODUCTION

Huge amount of hard copy journals or back volumes are available in Agricultural university and research institute libraries.

IT awareness in library users is increasing rapidly. A case study is presented here to illustrate how provision of information services can be made to increase proper utilization of recourses available in library. A corpus (large and structured set of text [5]) is utilized for experiment setup. This corpus consists of electronic copy of table of content (TOC) page of set of journals. Two methods can be used to search content based on user query. One of the common method adopted is keyword search in corpus and list all matching documents. This method will not take in account frequency count but will list out all journal in which keyword is found in TOC. The second method is use of data mining techniques, which not only searches the TOC but calculates the ***weights*** or ***probabilities*** to order the list with a heuristic called relevancy, hence would be more intelligent.

In this article we have proposed using TF-DF weights and text mining. Appropriate scripts are also given in Fig1 and Fig2 which could be of help to readers.

II. LITERATURE SURVEY

Debnath et.al [1] states that Text and data mining are two closely related approaches for information retrieval.

Panunzi et.al [2], Keti et. Al [3] presented a study using TF-IDF algorithm and reported that multi-term keywords increase the content identification with a 100% relative factor and that the adequacy is enhanced in 33% of cases. Keli et.al [3] have stated that in the domain of text categorization (TC), the TF (term frequency)*IDF (inverse document frequency) weighting algorithm and TF*IWF*IWF weighting algorithm are widely used.

Zhang [4] presented improved TF-IDF approach for text classification. In his article he mentioned that in text classification, a text document may partially match many categories. We need to find the best matching category for the text document. The term (word) frequency/inverse document frequency (TF-IDF) approach is commonly used to weigh each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories. Further he added, Text classification can be effected by various learning approaches of classifier, such as *k*-nearest neighbour, decision tree induction, naïve Bayesian, support vector machine and latent semantic index. Some of these techniques are based on, or correlated with, the TF-IDF approach representing text with vector space in which each feature in the text corresponds to a single word.

Wikipedia [5] defines a **corpus** or **text corpus** as a large and structured set of texts. In information retrieval, a **ranking function** is a function used by search engines to rank matching documents according to their relevance to a given search query. The tf-idf is example of simple ranking function

III. METHODOLOGY

The **TF–IDF** weight (term frequency–inverse document frequency) is a weight often used in information retrieval and text mining. A corpus of TOC page of 300 journals available in Marathwada Agricultural University library was built and used for experiment. Following procedure was followed for extraction of Journal name list based on keyword search.

```
<html> <body>
<H1><font color=BLUE  ALIGN=CENTER>  Journal Search  Tool </font></H1>  <HR>
<?php
$d = dir('cfiles/') or die($php_errormsg);
$jlc = file("jl.txt") or die($php_errormsg);
$tnofd = 332  ;
echo "<BR> Total Files : ";echo $tnofd;print "<BR>";
while (false !== ( $f = $d->read() ) )  {
if(('.' == $f) || ('..' == $f) ) { continue; }
if(preg_match('/^[a-zA-Z()-_]+$/',$f))
{
$file='cfiles/'.$f;
$content = strtolower(file_get_contents($file));
$wordArray = preg_split('/[^a-z]/', $content, -1,PREG_SPLIT_NO_EMPTY);
$wordFrequencyArray = array_count_values($wordArray);
arsort($wordFrequencyArray);
}
$cwf=count($wordFrequencyArray);
foreach ($wordFrequencyArray as $topWord => $frequency)
if((!preg_match("/^(|that|this|Dr.|a|at|and|the|this|an|in|or|of|is|for|to)$/",$topWord)) && (strlen($topWord) > 2))
{
if(!isset($klist[$topWord]))
{ $klist[$topWord] = array($frequency , $f); }
else
{ $klist[$topWord] =  array_merge($klist[$topWord],array($frequency ,$f)); }
}}
while(list($key,$value) = each($klist)) {
```

Fig. 1: Source code for Process Step 1

```
<html> <body>
<H1><font color=BLUE  ALIGN=CENTER>  Journal Search  Tool </font></H1>  <HR>
<?php
print "Please Enter Keywords : ";
print "<form action=\"$_SERVER[PHP_SELF]\" method=\"post\">";
print "<input type=hidden name=stage value=1> ";
print "<input type=text name=kword> ";
print "<input type=Submit name=go value=Go> <a href=$_SERVER[PHP_SELF]> Refresh </a>";
print "</form>";
$info = array($_POST[kword]);
$kw = $_POST[kword] ;
if($_POST[stage] > 0) {
print "Searchning for keyword $_POST[kword] - <BR> ";
$pat = "/$kw/";
$matchl = preg_grep($pat,file('keywordlist2.txt'));
$rc = 1;
$flist = array();
while ($rc<count($matchl)) {
foreach($matchl as $key=>$value) {
$val  = explode(',',$value);
unset($val[0]);unset($val[1]);
for($lcount=2;$lcount<count($val);$lcount=$lcount+2) {
$flist=array_merge($flist,array($val[$lcount+1]=>$val[$lcount])) ;}
}
$rc++ ;
} # while end
arsort($flist);
$jlc = file("jl.txt");
foreach($flist as $key=>$value)
{
for($c=0;$c<count($jlc);$c++)
{ $ccc = explode(',',$jlc[$c]);
if( $ccc[0]== substr($key,0,5) )
{ print( " <BR> $ccc[1] <a href=cfiles/$key.txt> < Text File > </a> <a href=cfiles/$key.pdf#search=\"$_POST[kword]\"> < PDF File > </a>
<Br>"); }
}
} # foreach $flist end
} # if POST end
?>
</html></body>
```

Fig. 2: Source code for Process Step 3

- Read each croup in the set. The stop words were eliminated to reduce vector space. Calculate and record - word, TOC text file (croup) name, and TF/IDF value for the word. This recorded data was the intermediate output (stored in say Reffile1) of the process which will be utilized for future The other index file (say Reffile2) was maintained for reference purpose which consist of pair Journal name and text filename(in the croup). The source code written in PHP is listed in Figure1.
- Present UI to accept keyword. The entered keyword is searched in recorded intermediate output. The Journal names are presented to user ordered by TF/IDF ranks. The source code written in PHP for this purpose is listed in Figure2.
- When a new journal is added to croup, a entry is added to Reffile2. Also the index Reffile1 is regenerated to accommodate keyword from newly arrived journal TOC.

Weka *Knowledge flow* provides an alternative to the Explorer graphical front end to Weka's core algorithms. Datasource TextDirectoryLoader is used to load directory containing text files used to store TOC pages of Journals. This datasource also converts the loaded data into arff (Attribute Relation File Format). Further stringtowordvector filter is applied to convert the loaded data into vector form followed by TextViewer available under Visualization tab of KnowldeFlow GUI interface to show the results of process.

The TOC text files needs to be classified based on content. The ARFF attribute for identification of Journal is to be included. Attempt is being made to implement using WEKA, which is expected to further reduce the search time.

IV. Conclusion

Merely scanning TOC page of journal and using its text for keyword search will generate unordered list of matching documents. Though it will help locating journal but would be inconvenient and time-consuming process. Our experience confirms that use of TF-IDF for weight measurement helps to extract ordered relevant documents. A open source tool like WEKA [6] would be more appropriate for such a experimental setup as it provides facility to read text directory, convert it into vector form and calculate TF/IDF weights. It also provides many more filters and converters for text and data mining. Work is in progress to implement the process using improved TF-IDF algorithm and java based tool WEKA.

References

[1] Bhattacharyya D, Das P, Ganguly D., Mitra K., Das Purnendu, Bandyopadhyay S.K., Kim T, Unstructured Document Categorization: A Study , Int. J of Signal Processing, Image Processing and Pattern Recognition, p51–56. Accessed on internet [www.sersc.org/ journals/IJSIP/vol1_no1/papers/07.pdf 07.pdf]

[2] Alessandro P., Marco F.,, Massimo M., Integrating Methods and LRs for Automatic Keyword Extraction from Open Domain Texts. Univ. of Florence, Italian Dept., Piazza Savonarola 1, Florence, Italy, [hnk.ffzg.hr/bibl/lrec2006/pdf/305_pdf.pdf]

[3] Keli Chen and Chengqing Zong klchen, cqzong, A NEW WEIGHTING ALGORITHM FOR LINEAR CLASSIFIER, National Laboratory of Pattern Recognition, Inst. of Automation, Chinese Academy of Sciences, Beijing 100080, China [http://www.nlpr.ia.ac.cn/english/cip/proceedings/A New Weighting Algorithm For Linear Classifier(Keli Chen).pdf]

[4] ZHANG Yun-tao et.al, An improved TF-IDF approach for text classification, J Zhejiang Univ SCI 2005 6A(1): 49–55 Accessed on net [A050108.pdf]

[5] http://en.wikipedia.org/wiki/ Ranking_function

[6] WEKA-http://www.cs.waikato.ac.nz/~ml/weka/

A Compression Algorithm for DNA Sequences Based on Palindrome Sequences with Information Security

Syed Mahamud Hossein[1] and B. Acharjee[2]
[1]*District Officer, Regional Office, Kolaghat, DVET, Govt. of West Bengal*
[2]*C.V.Raman College, Bhubneswar*
e-mail: mahamud123@gmail.com, bidisa.54@gmail.com

***Abstract*—A lossless compression algorithm, for genetic sequences, based on searching for exact palindromes is reported. The compression results obtained in the algorithm show that the exact palindromes are one of the main hidden regularities in DNA sequences. The proposed DNA sequence compression algorithm is based on genetic palindrome substring and creates online Library file acting as a Look Up Table. The genetic palindrome substring is replaced by corresponding ASCII character starting from 33(!). This substring length depends on user. Information security is the most challenging question to protect the data from unauthorized user. This proposed method may protect the data from hackers. It can provide the data security, by using ASCII code and on line Library file acting as a signature. Compressing the genome sequences will help to increase the efficiency of their uses. This algorithm is tested on benchmark DNA sequences, also on the reverse, the complement and the reverse complement benchmark DNA sequences, and on artificial DNA sequences. The algorithm can approach a compression rate of 3.851273 bit/base.**

Keyword: Data Compression

I. INTRODUCTION

Biological sequence compression is a useful tool to recover information from biological sequences. This was demonstrated for example in the construction of whole genome phylogenies [1]. With more and more complete genomes of prokaryotes and eukaryotes becoming available and the completion of human genome project in the horizon, fundamental questions regarding the characteristics of these sequences arise. We study one such basic question: the compressibility of DNA sequences. Life represents order. It is not chaotic or random [2]. Thus, we expect the DNA sequences that encode Life as nonrandom. Naturally they should be very compressible. There are also strong biological evidences in supporting this claim: It is well-known that DNA sequences, especially in higher eukaryotes, contain many genetic palindromes. It is also established that many essential genes (like rRNAs) have many copies. It is believed that there are only about a thousand basic protein folding patterns. Further it has been conjectured that genes duplicate themselves sometimes for evolutionary or simply for "selfish" purposes. These all concretly support that the DNA sequences should be reasonably compressible. It is well recognized that the compression of DNA sequences is a very difficult task [3, 4, 5, 6]. The DNA sequences only consist of 4 nucleotide bases {a, c, g, t}(note that t is replaced with u in the case of the RNA), 8 bits are enough to store each base. However, if one applies standard compression software such as the Unix "compress" and "compact" or the MS-DOS archive programs "pkzip" and "arj", they all expand the file with more than 8 bits per base, although all these compression software are universal compression algorithms. These software are designed for text compression [7], while the regularities in DNA sequences are much subtler. It is our purpose to study such subtleties in DNA sequences. We will present a DNA compression algorithm, based on exact matching that gives the best compression results on standard benchmark DNA sequences. However, searching for all exact palindromes in a very long DNA sequence is not a trivial task. These algorithms take a long time (essentially a quadratic time search or even more) in order to find approximate palindromes that are optimal for compression. Simultaneously achieving high speed and best compression ratio remains to be a challenging task. Proposed DNA sequences Compression achieves a better compression ratio and runs significantly faster than any existing compression program for benchmark DNA sequences, simultaneously. Proposed algorithm consists of two phases: i) find all exact genetic palindromes; and ii) encode exact genetic palindrome regions and non-genetic palindrome regions. We have developed for fast and sensitive homology search[8], as our exact genetic palindrome search engine. Compression of DNA sequences is a very challenging task. This can be seen by the fact that no commercial file-compression program achieves any compression on benchmark DNA sequences we use in this paper. Several compression algorithms specialized for DNA sequences have been developed in earlier studies elsewhere. We will present a DNA compression

algorithm, based on genetic palindrome substring and corresponding genetic palindrome substring is place in Library file , this genetic palindrome substring create a dynamic Look Up Table and place ASCII character in appropriate places on source file and that gives the best compression results on standard benchmark DNA sequences. We will discuss details of the algorithm, provide experimental results and compare the results with the one most effective compression algorithm for DNA sequence (gzip-9).

In Look up Table Encoding time "Sub-sequence size-1" base segment is remaining, (if at the end of file segment are not match exactly with pre-coded table). We cannot find any arrangement in look up Table, in these circumstances, we just write the original segment into destination file[9]. This method will help to minimize this problem and increase the probability of compaction. We find the compression ratio, compression rate result in other orientation such as the reverse, the complement and the reverse complement the input sequences. But experimental result showing no meaningful changes are found using other orientation taking as a input sequences. Also we can find the compression rate, compression ratio of randomly generated of equivalent length of artificial DNA sequence. Compare all result to each other.

In this paper, if not otherwise mentioned, we will use lower case letters u and v, to denote finite strings over the alphabet {a, c, g, t},|u| denotes the length of u, the number of characters in u. u_i is

the i-th character of u. $u_{i:j}$ is the substring of u from position i to position j. The first character of

u is u_1. Thus $u = u_{1:|u|-1}$. and |v| denotes the length of v, the number of characters in v. v_i is

the i-th character of v. $v_{i:j}$ is the another substring of v from position i to position j. $u_{i:j}$ match with $v_{i:j}$. The first character of v is v_1. Thus $v = v_{1:|v|-1}$. The minimum difference between u-v is of substring length. The palindrome found if $u_{i:j=}$ $v_{i:j}$ and count exact maximum palindrome of $u_{i:j..}$ We use ε to denote empty string and ε=0.

II. Methods

A. File Format

We will begin discussing file type is text file (file extension is dot txt) contain a series of successive four base pair (a,t,g and c) and end with blank space ahead the end of file. Text file is the basic element to which we consider in compression and decompression. The output file also text file, contains the information of both unmatch four base pair and a coded value of ASCII character.

The coded valued are located in the encoded section. The coded information is written into destination file byte by byte. The file size depends on number of base pair present in the input file and output file measured by byte, i.e. File size (in byte)= number of base pair in a file(in byte).As per example total number of base pair in a file is n, so the file size is n byte. ASCII character also required one byte for storing. On the basis of ASCII code availability, we can take input as a lower case letter of a,t,g and c.

B. Generating the Substring from Input Sequence

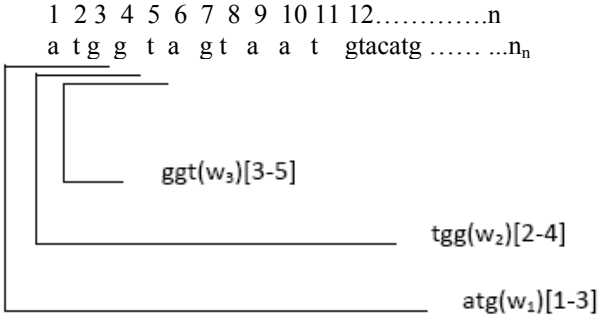


Fig. I: Substring Creation

From the pictorial representation of fig- I it is clear that for i^{th} substring W_i .

i, is the starting position of the substring and.

j= (i-1) + l, is end position of the substring; where l is the substring length (word size).

As for example if substring length is 3 then:

For: W_1 starting position (i)=1 and (end position)j= (1-1) + 3=3,

W_2 starting position (i)=2 and (end position)j= (2-1) + 3=4 and

W_3 starting position (i)=3 and (end position)j= (3-1) + 3=5 and so on.

The substring length is less than 3 (three) has no importance in matching context therefore we consider the substring size in the range: $3 \leq l \leq n$

Therefore range for i and j are as $1 \leq i \leq n-l+1$ and $1 \leq j \leq n$ respectively.

C. Searching for Exact Palindromes

Consider a finite sequence *s* over the DNA alphabet {a, c, g, t}. An exact palindrome is a substring in *s* that can be transformed from another substring in *s* with edit operations (palindrome, insertion). We only encode those exact palindromes that provide profits on overall compression.

This methods of compression is as below

- Run the program and output all exact palindromes into a list s in the order of descending scores;
- Extract a palindrome *r* with highest score from list s, then replace all r by corresponding ASCII code into another list o and place r in library file.
- Process each palindrome in s so that there's no overlap with the extracted palindrome *r* ;

- Goto step 2 if the highest score of palindromes in s is still higher than a pre-defined threshold; otherwise exit.

1) Example

Let s=atagattagatatacata……..n

{ata substring palindrome on three places, gat palindrome on two place, so, on. First replaced highest match score is atg by ASCII character and insert ASCII equivalent symbol in ith position]

B=!gatgat!tac! { B is intermediate encoding step}

o=!””!tac![where o is the compress output file]

All those extracted palindromes in list *B* then parse a DNA sequence into a mixture of regions with little structure and palindrome regions each of which can be replaced by a substring previously located.

D. Encoding palindromes

An exact palindrome can be presented as two kinds of triples. first is (l, m, p), where l means the palindrome substring length, m and p show the starting positions of two substrings in a palindrome, respectively, second Replace. This operation is expressed as (r; p; char) which means replacing the exact palindrome substring at position p by ASCII character char.

In order to recover an exact genetic palindrome correctly the following information must be encoded in the output data stream:

Encoding Analysis

m→

So, we can write s=atagattagatatacata……..n n>0 and 1≤i≤n-L+1

p→

Consider the sequence is defined by s, consider genetic palindrome substring store in S[m] and all match genetic palindrome substring are store in S[p]

After breaking the sequence(s) into substring of three bases lonng we can get the result as below.

So, we can get S[m]=S[1]……..S[n-2*l+1] 1≤m≤n-2*l+1 and

Genetic palindrome substring are S[p]=S[1]……S[n-l+1] 1≤p≤n-l+1

If the number of substring in S[m], total number of subsequence are generated by (n-2*l+1) and Number of mach genetic palindrome substring in S[p], total match genetic palindrome substring are (n-l+1)

As per above example s[m]→s[1]=ata and so on

And s[p] →s[1]=gat and so on.

This substring method are require to reduce the complexity of the programme execution.

E. Each Substring Match with all other Substring for Finding the Exact Maximum Genetic Palindrome Substring.

Match condition occur if S[m]=S[p] p=l+l

Step-I

S[1] match with S[p] to S[n-l+1] and count S[1]

{As per example S[1]=atg where substring size=3 and S[4]=gat, S[5]=gat….S[19]=ata

So, S[1] substring genetic palindrome at 3 places

Then m and p incremented by one

Step-2

Match S[2] match with S[p] to S[n-L+1] and count S[2]

[As per example S[2]=gat

and S[5]=tag ,S[6]=gat

So, S[2] substring genetic palindrome at one places

Then m and p incremented by one

Step-3

This method will continue to S[n-l+1]

So S[n-2*l+1] match with S[p] to S[n-2*l+1] and count S[n-2*l+1]

So, S[n-2*L+1] genetic palindrome only one place if mach occur.

Step-4:

Store all genetic palindrome count in descending order and find all exact maximum genetic palindrome count

Step-5:

Replace exact maximum palindromes substring by corresponding ASCII code and place genetic palindrome substring in library file, and create a on line look up Table.

Step-6:

Genetic palindrome Step-1 to step-5 excluding ASCII code

Step-7:

if the highest score of palindromes in s is still higher than a pre-defined threshold; otherwise exit.

As per above example: Now we find maximum palindrome probability. This substring are replace first. Here We can get S[2] = (gat) substring are genetic palindrome 3 times in this sequence.

This substring are place in Look up Table, corresponding ASCII[33(!)] character and replace all genetic palindrome substring by this ASCII character. Library file and create a on line Look up Table.

So, n=Length of the string = Total number of base pair in s = File size in byte

The Encoding procedure flow this rule and produce compression output file.

S[m] match with S[p] to S[n-L+1],place ASCII character in the output file ith position. Each matching cases the value of m is incremented by; m=number of unmatch character+ (number of sub-string match * substring length + 1)

Otherwise S[m]≠S[p] to S[n-L+1]place base pair in output file ith position. If unmatch occur , the value of m and p is incremented by one.

At the end, we can get the compress output file o which is contain the unmatch a,t,g and c and ASCII character set.

At the end we can gat the compress file, corresponding input sequence

So, O= !""!tac!…….n_1 where n_1 is the length of output file. Output file size is n_1 byte

And library file : !ata "gat

F. Decoding

Decoding time, first require on line Library file, which was created at the time of encoding the input file.

On this particular value , the encoded input string is decoded and produce the output original file.

LOOK UP TABLE-II

O= !""!tac!…….n_1 where n_1 is the length of output string ($n>n_1$).

At the time of decoding each ASCII character is replaced by corresponding base pair i,e O[M]=L[k] where O[M] is define by output sequence and L[k] is define by library file substring. If match occurs in between L[33] to L[256] with O[M], place ASCII equivalent substring in ith places in output file. The value of m is incremented by one. If unmatch found in between L[33] to L[256] with O[M], place base pair in ith position in output file. The value of M is incremented by one. This process will continue until $M=n_1$ position will appear.

The Decoding process mentioned this rule and produce original output string.

Match found if o[m]=L[33] to L[256] place ASCII character equivalent substring in i-th position. If match found, the value of m is incremented by one.

Otherwise o[m]≠L[33] to L[256] place base pair in i-th position in output file. If unmatch occurs , the value of m is incremented by one.

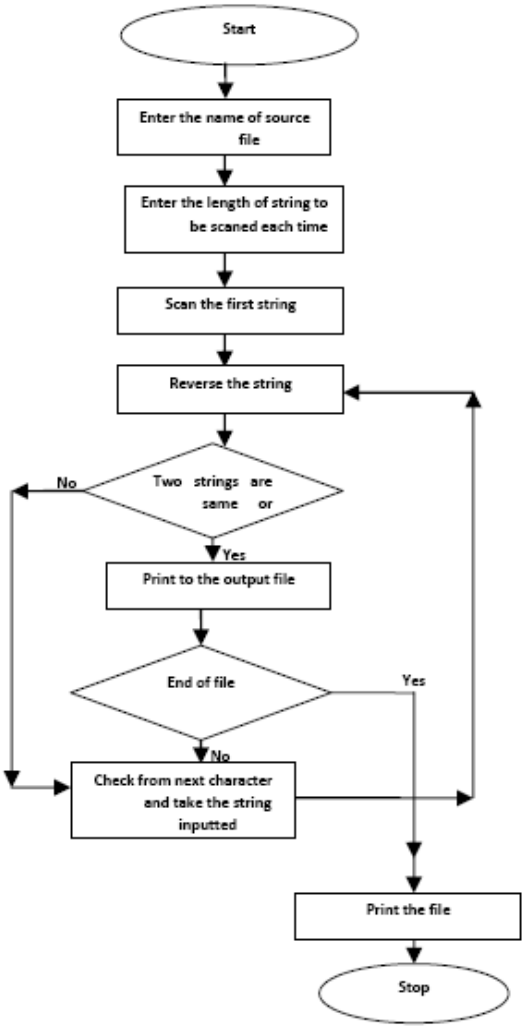
For easy implementation, characters a,t,g,c will no longer appear in pre-coded file and A,T,G,C will appear in pre-coded file. For instance, if a segment "atagattagatatacata……..n" has been read, in the destination file, we represent them as "!""!tac!……..n_1".Obviously, the destination file is case-sensitive

We know that each character require 1 byte (8 bit) for storing. In the above example string length = 18 that means 18 byte require for storing this string. After encoding on the basis of genetic palindrome techniques of 3 substring length, reduce string length is 8, require 8 byte for storing this string.

G. Algorithm

- Enter the name of the source file.
- Enter the name of the destination file where the palindrome will be printed.
- Enter the length of the string be taken input each time from the source file.
- Take the first string of the specified length.
- Reverse the string.
- Check whether the source and reverse string are same or not. If same write it to output file specifying the position.
- If palindrome found or not take the second string of specified length starting from second character of the source file. Continue steps 5, 6 & 7 till the end of the file.
- If the file is ended stop.

Flowchart



III. ALGORITHM EVALUATION

A. Accuracy

As to the DNA sequence storage, accuracy must be taken firstly in that even a single base mutation, insertion, deletion or SNP would result in huge change of phenotype as we see in the sicklemia. It is not tolerable that any mistake exists either in compression or in decompression. Although not yet proved mathematically, it could be infer from palindrome techniques that our algorithm is accurate, since every base arrangement uniquely corresponds to an ASCII character.

B. Efficiency

We can see that the internal palindrome algorithm can compress original file from substring length (l) into 1 characters for any DNA segment, and destination file uses less ASCII character to represent successive DNA bases than source file.

C. Space Occupation

Our algorithm reads characters from source file and writes them immediately into destination file. It costs very small memory space to store only a few characters. The space occupation is in constant level. In our experiments, the OS has no swap partition. All performance can be done in main memory which is only 512 MB on our PC.

IV. EXPERIMENTAL RESULTS

We tested palindrome techniques on standard benchmark data used in [10]. For testing purpose we use eight types of data.

These tests are performed on a computer whose CPU is Intel P-IV 3.0 GHz core 2 duo(1024FSB), Intel 946 original mother board, IGB DDR2 Hynix, 160GB SATA HDD Segate. Since the program to implement the technique have been written originally in the C++ language, (Windows XP platform, and TC compiler) it is possible to run in other microcomputers with small changes (depending on platform and Compiler used). The program runs on the IBM personal computer, requires 512K, without additional hardware except for disk drives and printer.

The definition of the compression ratio [11]; 1−(|O|/2| *I*|), where |*I*| is number of bases in the input DNA sequence and |O| is the length (number of bits) of the output sequence, The compression rate, which is defined as (|O|/| *I*|), where |*I*| is number of bases in the input DNA sequence and |O| is the length (number of bits) of the output sequence. Total reduce file size is equal to Compress file size plus Library file size in byte, i.e (T=C+L byte). The improvement[9] over gzip-9, which is defined as (Ratio_of_gzip-9 – Ratio_of_LUT-3)/Ratio_of_gzip-9 *100. The compression ratio and compression rate are presented in Table-I. Our result compared with gzip-9[12] in the same table, also this table shown , the reverse, the complement and the reverse complement sequences result.

V. RESULT DISCUSSION

Our algorithm is very useful in database storing. You can keep sequences as records in database instead of maintaining them as files. By just using the exact palindrome, users can obtain original sequences in a time that can't be felt. Additionally, our algorithm can be easily implemented.

From these experiments, we conclude that internal palindrome matching patter are same in all type of sources and Look up Table plays a key role in finding similarities or regularities in DNA sequences. Output file contain ASCII character with unmatch a,u,g and c so, it can provide information security which is very important for data protection over transmission point of view. This techniques provide the high security to protect nucleotide sequence in a particular source. Here we can get better security than static LUT.

TABLE-I

Sequence Size	Sequence Name	Base pair/ File size	Cellular DNA Sequences								Artificial sequences							
			Normal Sequences		Reverse Sequences		Complement Sequences		Reverse Complement Sequences		Normal Sequences		Reverse Sequences		Complement Sequences		Reverse Complement Sequences	
			Compression ratio	Compression rate (bits /base)	Compression ratio	Compression rate (bits /base)	Compression ratio	Compression rate (bits /base)	Compression ratio	Compression rate (bits /base)	Compression ratio	Compression rate (bits /base)	Compression ratio	Compression rate (bits /base)	Compression ratio	Compression rate (bits /base)	Compression ratio	Compression rate (bits /base)
Sub string Size 3	atatsgs	9647	-1.037939	4.075879	-1.070281	4.140562	-1.037939	4.075879	-1.070281	4.140562	-1.072354	4.144708	-1.07111	4.14222	-1.072354	4.144708	-1.07111	4.14222
	atef1a23	6022	-1.116905	4.233809	-1.114912	4.229824	-1.116904	4.233809	-1.114912	4.229824	-1.130854	4.261707	-1.13351	4.267021	-1.130854	4.261707	-1.13351	4.267021
	atrdnaf	10014	-1.017176	4.034352	-1.017975	4.03595	-1.017176	4.034352	-1.017975	4.03595	-1.045137	4.090274	-1.073098	4.146195	-1.045137	4.090274	-1.073098	4.146195
	atrdnai	5287	-1.073766	4.147532	-1.062417	4.124834	-1.073765	4.147532	-1.062417	4.124834	-1.166068	4.332135	-1.166068	4.332135	-1.166068	4.332135	-1.166068	4.332135
	celk07e12	58949	-0.917522	3.835044	-0.914536	3.829073	-0.917521	3.835044	-0.914536	3.829073	-0.969092	3.938184	-0.97167	3.943341	-0.969092	3.938184	-0.97167	3.943341
	hsg6pdgen	52173	-0.949591	3.899182	-0.956031	3.912062	-0.949590	3.899182	-0.956031	3.912062	-0.986238	3.972476	-0.983631	3.967263	-0.986238	3.972476	-0.983631	3.967263
	mmzp3g	10833	-1.023816	4.047632	-1.019385	4.03877	-1.023816	4.047632	-1.019385	4.03877	-1.052617	4.105234	-2.602326	7.204652	-1.052617	4.105234	-1.049663	4.099326
	xlxfg512	19338	-0.941876	3.883752	-0.950977	3.901955	-0.941876	3.883752	-0.950977	3.901955	-1.017168	4.034337	-0.148206	2.296411	-1.017168	4.034337	-1.017996	4.035991
Sub string Size 4	atatsgs	9647	-1.088525	4.17705	-1.074427	4.148855	-1.088525	4.17705	-1.074427	4.148855	-1.153208	4.306417	-1.136623	4.273246	-1.072354	4.144708	-1.136623	4.273246
	atef1a23	6022	-1.22916	4.458319	-1.219861	4.439721	-1.22916	4.458319	-1.219861	4.439721	-1.278313	4.556626	-1.286284	4.572567	-1.153208	4.306417	-1.286284	4.572567
	atrdnaf	10014	-1.12822	4.256441	-1.097064	4.194128	-1.12822	4.256441	-1.097064	4.194128	-1.148991	4.297983	-1.169363	4.338726	-1.278313	4.556626	-1.169363	4.338726
	atrdnai	5287	-1.256856	4.513713	-1.231133	4.462266	-1.256856	4.513713	-1.231133	4.462266	-1.354454	4.708909	-1.355967	4.711935	-1.148991	4.297983	-1.355967	4.711935
	celk07e12	58949	-0.799454	3.598908	-0.810039	3.620078	-0.810039	3.620078	-0.785543	3.571087	-0.852313	3.704626	-0.852788	3.705576	-1.354454	4.708909	-0.852788	3.705576
	hsg6pdgen	52173	-0.810822	3.621643	-0.819715	3.63943	-0.810822	3.621643	-0.819715	3.63943	-0.852299	3.704598	-0.820175	3.64035	-0.852313	3.704626	-0.857206	3.714412
	mmzp3g	10833	-1.038586	4.077172	-1.057786	4.115573	-1.038586	4.077172	-1.057786	4.115573	-1.13385	4.267701	-1.138281	4.276562	-0.852299	3.704598	-1.138281	4.276562
	xlxfg512	19338	-0.886441	3.772882	-0.879202	3.758403	-0.886441	3.772882	-0.879202	3.758403	-0.998345	3.99669	-0.99276	3.985521	-1.115388	4.230776	-0.99276	3.985521

VI. Conclusion

In this article, we discuss a new DNA compression algorithm whose key idea is internal genetic palindrome. This compression algorithm gives a good model for compressing DNA sequences that reveals the true characteristics of DNA sequences. The compression results of genetic palindrome DNA sequences also indicate that our method is more effective than many others. This method is able to detect more regularities in DNA sequences, such as mutation and crossover, and achieve the best compression results by using this observation. This method is fails to achieve higher compression ratio than others standard method, but it has provide very high information security.

Important observation are:

- Genetic palindrome substring length vary from 2 to 5 and no match found in case the substring length becoming six or more.
- The substring length is three of highly genetic palindromeed than substring length of four and five. That is why substring length of three is highly compressible over substring length of four and five.
- Normal sequence is highly compressible than reveres, complement and reverse complement sequences.
- Cellular DNA sequences compression rate and compression ratio are distingushable different due each sequence that come into different sources where as artificial DNA sequences compression rate and compression ratio are same in all time in all data sets.

VII. Future Work

We are developed to further research on as combination of palindrome, reverse ,repeat and genetic palindrome, this technique are also use to compression method. Also we try to reduce the time complexity

Acknowledgement

Authors are grateful to all their our colleagues/friends for their interest and constructive criticism of this study. The authors offer special thanks to Dr. S.Basu, West Bengal University of Technology, Kolkata.

References

[1] Li,M., Badger, J., Chen, X., Kwong, S., Kearney, P. and Zhang, H.(2001) An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics, 17, 149–154.

[2] M. Li and P. Vitányi, an Introduction to Kolmogorov Complexity and Its Applications, 2nd ed. New York: Springer-Verlag, 1997.

[3] Curnow, R. and Kirkwood, T., Statistical analysis of deoxyribonucleic acid sequence data {a review, J. Royal Statistical Society, 152: 199{220, 1989.

[4] Grumbach, S. and Tahi, F., A new challenge for compression algorithms: genetic sequences, J. Information Processing and Management, 30(6): 875–866, 1994.

[5] Lanctot, K., Li, M., and Yang, E.H., Estimating DNA sequence entropy, to appear in SODA '2000.

[6] Rivals, _E., Delahaye, J.-P., Dauchet, M., and Delgrange, O., A Guaranteed Compression Scheme for Repetitive DNA Sequences, LIFL Lille I University, technical report IT–285, 1995.

[7] Bell, T.C., Cleary, J.G., and Witten, I.H., Text Compression, Prentice Hall, 1990.

[8] Ma, B., Tromp, J. and Li, M. (2002) Pattern Hunter—faster and more sensitive homology search. Bioinformatics, 18, 440–445. 1698

[9] Md. Syed Mahamud Hossein, and Subhajit Das, "A Compression Algorithms for DNA Sequences based on two LOOK UP TABLE," in BVCOM'07, national conference, Dec, 7–8, 2007.

[10] S. Grumbach and F. Tahi, "A new challenge for compression algorithms: Genetic sequences," J. Inform. Process. Manage., vol. 30, no. 6, pp. 875–866, 1994.

[11] Xin Chen, San Kwong and Mine Li, "A Compression Algorithm for DNA Sequences Using Approximate Matching for Better Compression Ratio to Reveal the True Characteristics of DNA", IEEE Engineering in Medicine and Biology, pp 61–66, July/August 2001.

[12] T. Matsumoto, K. Sadakame and H. Imani," Biological sequence compression algorithm", Genome Informatics 11: 43–52 (2000).

[13] ASCII code. [Online]. Available: http://www.asciitable.com

[14] National Center for Biotechnology Information,http://www.ncbi.nlm.nih.gov

A Comparative Study of MSA Tools Based on Sequence Alignment Features and Platform Independency to Select the Appropriate Tool Desired

Dr. Sayyed Iliyas[1] and Ms. Farhana S. Sarkhawas[2]*
[1]*Department of Botany, Poona College of Arts, Science & Commerce, Camp, Pune-1*
*[2]*MCA Dept., Allana Institute of Management Sciences, Camp.Pune-411001*
E-mail: sayyed_iliyas@yahoo.com, farhana.ap@gmail.com

***Abstract*—Multiple sequence alignment is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignments are at the core of bioinformatical analysis. There is wide range of available MSA software for performing multiple sequence alignment. Since, there are many differences in the functionality and accuracy of these software's which makes it difficult for biologists and bioinformaticians to decide which program is best suited for a given purpose. There are many comparative studies done in the past research which used benchmarking tools like BALIBASE and Homstard for alignment accuracy and Friedman test for comparing scores. The aim of this research is to make a comparative study of eight multiple sequence alignment tools i.e. Geneious, ClustalW, DNAMAN, Strap ,GoCore, MUSCLE, HMMER and SequenceAnalysis based on the criteria of availability, execution, sequencing algorithms and statistical parameters which would help bioinformaticians and biologists to select the correct tool for performing multiple sequence alignment. Finally, based on the comparative study some deem parameters were highlighted to choose the required MSA tool as proposed in the study.**

Keywords: multiple sequence alignment, pairwise alignment, windows or linux, molecular weight, score , identity

I. INTRODUCTION

Multiple sequence alignment is perhaps the most commonly applied bioinformatics technique. The "sequences" to be compared can be DNA or proteins. Multiple alignments are often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related.

The goal of multiple sequence alignment is to generate a concise, information-rich summary of sequence data in order to inform decision-making on the relatedness of the sequences to a gene family. The drawback with using MSA is that it requires an enormous amount of both computer time and memory to align more than a few distantly related sequences.

Asieh and her team [17] evaluated well-known MSA tools used by biologists and bioinformaticians in order to select the proper software which corresponds best to their specific needs. Alignment results were compared to the BALiBASE benchmark output while scorecons server was employed to achieve scorecons score (SCS) as a new method to assess MSA tools.

The first systematic study of the most commonly used alignment programs like multalign, pileup, clustalx using balibase benchmark tool was done by Thompson [9].

Paulo [15] results indicated that employing Simprot's simulated sequences allows the creation of a more flexible and broader range of alignment classes than the usual methods for alignment accuracy assessment.

Diamantis [18] made a comparison study of ten multiple sequence alignment programs based on the following criterias web support, max sequences, global or local alignment, algorithm url and calculated the mean for tcoffee, clustalW and Dialign.

There have been many algorithms and software programs implemented for the inference of multiple sequence alignments of protein and DNA sequences.

Specialized alignment tools are almost always needed for long, genome sized DNA or protein sequences. Selecting the specialized tool is a tough job for biologists to execute their requirement.

The current study signifies this critical issue in relation to MSA algorithm by systematically comparing the differences of the latest and upcoming multiple sequence alignment tools (that are open source):

- Geneious
- SequenceAnalysis
- DNAMAN
- GoCore [14]
- MUSCLE [12]
- HMMER [8]
- STRAP [11]
- ClustalX [7]

The reason for choosing the above mentioned tools are as most of them are upcoming tools having many useful features related to multiple sequence alignment.

In this study, different comparisons are made on various factors like the algorithms, file formats supported, execution time, memory requirement of the alignments performed, representation techniques of the alignments and the score calculated from the alignment.

The results obtained are compared with the previous research work.

II. Experimental Setup

We have downloaded eight open source multiple sequence alignment software. Table-1 shows some information related to the eight software. HMMER and MUSCLE are the two software which have been executed on Linux platform which requires a lot of manual intervention as these software work using command line arguments. STRAP, DNAMAN, Geneious and SequenceAnalysis software requires java runtime environment to execute without which it will not run. Anyone who can use excel can easily use GoCore. It operates as an Excel Add-In, creating additional menu item that performs the bioinformatical analysis and uses Excel's convenient user interface to provide useful visualization. Geneious, ClustalX and DNAMAN are available as exe files and run on windows as well as linux platform.

We have downloaded five protein sequences from Protein Databank (PDB) database *(www.pdb.org).* The PID's of these protein sequences are >1BQ6, >1FM8, >2GAS, >1CGK and >1EYQ. These protein sequences are saved in fasta format.

The eight software were executed on a desktop computer with Pentium IV processor and 256MB ram with a dual boot operating system i.e. WindowsXP and Fedora4.

III. Results & Discussion

In this study eight most upcoming and widely used MSA tools have been tested under the windows and/or linux platform to perform the multiple sequence alignment. The procedure to perform multiple sequence alignment from any of the eight MSA tools is shown in Fig.1. The procedure is as follows:

- Downloaded five protein sequences from the public database i.e. PDB (Protein Data Bank) (see Table-2)
- The five sequences are given as input to the Geneious software .
- Set the parameters required to perform the MSA as alignment algorithm and scoring matrix
- Click on Alignment tab to generate the MSA.

Fig.1: Procedure to Perform MSA

All the eight alignment programs used algorithms to perform pairwise and multiple sequence alignment. Table-3 shows various algorithms used by the eight MSA software to perform pairwise and multiple sequence alignment.

It is really difficult to know which MSA software support which file format unless it is tried on the software and the error is generated saying file format not supported. Table-4 gives the list of various file format's supported by the MSA software.

The following experiments were performed:

IV. Experiment 1

Table-5 shows the step-wise progress in performing the multiple sequence alignment by MUSCLE tool using the following command:

muscle –in seq.fa –out seq.afa

MUSCLE is the only MSA tools gives the memory requirement to perform the sequence alignment as well the time elapsed.

V. Experiment 2

The SequenceAnalysis tool calculates the score and identity of the pairwise alignment depending on the Scoring matrices and alignment algorithm by clicking on the "Do-Alignment" button in the Pairwise tab of the tool (Table-7).

VI. Experiment 3

The Geneious tool finds score and identity related to the alignment by choosing the Alignment Option in the menu bar of the tool window (Table-8).

VII. Experiment 4

The HMMER tool calculates the standard deviation, the minimum, maximum and average score by typing the following command (Table-6)

> hmmbuild seq.hmm seq.cln

Few strengths and weaknesses of the MSA tools are shown in (Table-9) which will help the bioinformaticians to choose the desired tool.

VIII. Conclusion

The aim of this study was to evaluate the most upcoming MSA tools which are easily available over the internet.

Various output results from MSA software can help the bioinformatician to solve their desired requirements. The deem parameters listed about the tools can help the bioinformaticians choose the right tool accordingly.

TABLE 1: GENERAL INFORMATION OF THE MULTIPLE SEQUENCE ALIGNMENT SOFTWARE

Tools	Description	Sequence Type*	Alignment Type**	Link	Year	Author
Geneious	Progressive/ Iterative alignment;	Both	Both	www.geneious.com	2008	A.J. Drummond
Sequence Analysis	Pairwise Alignment	Both	Both	http://www.informagen.com /SA/	2006	Will Gilbert
GoCore	Sequence Alignment Tool	Protein	Both	www.helsinki.fi/project/ritvos/GoCore/	2005	Luke Jeffery
DNAMAN	Multiple Sequence Alignment	Both	Both	www.lynnon.com	2005	Huang & Zhang
MUSCLE	Progressive/iterative alignment	Both	Both	www.drive5.com/muscle	2004	Robert Edgar
HMMER	Hidden Markov profile search	Both	Both	www.hmmer.wustl.edu/	2003	Ewan
Strap	Multiple Sequence Alignment Tool	Both	Both	www.charite.de/bioinf/strap	2001	Christoph Gille
ClustalX	Progressive Alignment	Both	Both	www.clustal.org	1994	Julie Thompson Toby Gibson

**-Protein and nucleotide, **-local and global*

TABLE 2: STATISTICAL PROPERTIES OF PROTEIN SEQUENCES BY DNAMAN TOOL

Sequence Name	Sequence	Length	Isoelectric Point (pI)	Molecular Weight (Daltons)
>1BQ6	VSVSEIRKAQRAEGPATILAIGTANPANCVEQSTYPDFYFKITN SEHKTELKEKFQRMCDKSMIKRRYMYLTEEILKENPNVCEYM APSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIV CTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGT VLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLV GQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGA IDGHLREAGLTFHLLKDVPGIVSKNITKALVEAFEPLGISDYNS IFWIAHPGGPAILDQVEQKLALKPEKMNATREVLSEYGNMSS ACVLFILDEMRKKSTQNGLKTTGEGLEWGVLFGFGPGLTIET VVLRSVAI	388	6.27	42567.4D
>1FM8	MAASITAITVENLEYPAVVTSPVTGKSYFLGGAGERGLTIEGN FIKFTAIGVYLEDIAVASLAAKWKGKSSEELLETLDFYRDIISG PFEKLIRGSKIRELSGPEYSRKVMENCVAHLKSVGTYGDAEAE AMQKFAEAFKPVNFPPGASVFYRQSPDGILGLSFSPDTSIPEKE AALIENKAVSSAVLETMIGEHAVSPDLKRCLAARLPALLNEG AFKIGN	222	4.98	23821.9D
>2GAS	TENKILILGPTGAIGRHIVWASIKAGNPTYALVRKTITAANPET KEELIDNYQSLGVILLEGDINDHETLVKAIKQVDIVICAAGRLL IEDQVKIIKAIKEAGNVKKFFPSEFGLDVDRHDAVEPVRQVFE EKASIRRVIEAEGVPYTYLCCHAFTGYFLRNLAQLDATDPPRD KVVILGDGNVKGAYVTEADVGTFTIRAANDPNTLNKAVHIRL PKNYLTQNEVIALWEKKIGKTLEKTYVSEEQVLKDIQESSFPH NYLLALYHSQQIKGDAVYEIDPAKDIEASEAYPDVTYTTADE YLNQFV	307	5.01	34253.9D
>1EYQ	MAASITAITVENLEYPAVVTSPVTGKSYFLGGAGERGLTIEGN FIKFTAIGVYLEDIAVASLAAKWKGKSSEELLETLDFYRDIISG PFEKLIRGSKIRELSGPEYSRKVMENCVAHLKSVGTYGDAEAE AMQKFAEAFKPVNFPPGASVFYRQSPDGILGLSFSPDTSIPEKE AALIENKAVSSAVLETMIGEHAVSPDLKRCLAARLPALLNEG AFKIGN	222	4.98	23821.9D

>1CGK	MVSVSEIRKAQRAEGPATILAIGTANPANCVEQSTYPDFYFKI TNSEHKTELKEKFQRMCDKSMIKRRYMYLTEEILKENPNVCE YMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITH LIVCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAG GTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSL VGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEG AIDGHLREAGLTFHLLKDVPGIVSKNITKALVEAFEPLGISDYN SIFWIAHPGGPAILDQVEQKLALKPEKMNATREVLSEYGNMS SACVLFILDEMRKKSTQNGLKTTGEGLEWGVLFGFGPGLTIET VVLRSVAI	389	6.27	42698.6D

TABLE 3: ALGORITHMS SUPPORTED BY MSA TOOLS TO PERFORM MULTIPLE SEQUENCE ALIGNMENT

Tools	**Algorithm**		
	Pairwise		**Multiple**
	Local	**Global**	
Geneious	Smith-Waterman [2]	Needleman-Wunsch [1]	Feng and Doolittle & Needleman-Wunsch
Clustal X	Clustal [4]	Clustal	Clustal
GoCore	ND	ND	T_Coffee [10]
STRAP	NeoBio, JAligner	ND	ClustalW [6] , T_Coffee, Kalign and MUSCLE
MUSCLE	Muscle [13]	Muscle	Muscle
HMMER	ND	Plan7 Model [5]	Plan7 Model
Sequence Analysis	Smith-Waterman and Crochemore, Landau & Ziv-Ukelson	Needleman-Wunsch and Crochemore, Landau & Ziv-Ukelson	ND
DNAMAN	Smith-Waterman	Myer's & Miller's and Needleman-Wunsch	Feng and Doolittle and Wibur and Lipman[3]

ND – Not Done

TABLE 4: VARIOUS FILE FORMATS SUPPORTED BY THE MSA TOOLS

File Formats	**Tools**							
	Geneious	Clustal X	GoCore	Strap	MUSCLE	Hmmer	Sequence Analysis	DNAMAN
Clustal (*.cln)	√	√	X	√	√	√	X	X
DnaStar (*.seq) &(*.pro)	√	X	X	√	X	X	√	X
DnaStrider (*.str)	√	X	X		X	X	√	√
Embl/Uni Prot (*.embl) & (*.swp)	√	X	X	√	X	X	X	X
Fasta (*.fasta)	√	√	√	√	√	√	√	√
Gen Bank (*.gb) &(*.xml)	√	√	√	X	X	X	X	√
PDB (*.pdb)	√	√	X	√	X	√	√	√
PIR/NBRF (*.pir)	√	√	X	X	X	X	√	X
GCG/MSF (*.seq)	√	√	X	X	X	√	√	X
GDE (*.gde)	X	√	X	X	X	X	√	X
Embl/Swiss Prot (*.swp)	X	√	X	√	X	X	X	X

TABLE 5: PROGRESS MESSAGES SHOWN BY THE MUSCLE SOFTWARE WHEN PERFORMING THE PAIRWISE AND MULTIPLE ALIGNMENTS

Alignment Type	**Name of Sequences**	**Elapsed Time**	**Memory**	**Iterations**	**Time Completed**	**Pass**
Pairwise Multiple	1BQ6 & 1FM8	00:00:00	9 MB (3%)	1	100.00%	Pass 1
		00:00:00	9 MB (3%)	1	100.00%	Pass 2
		00:00:01	11 MB (3%)	1	100.00%	Align Node
		00:00:01	11 MB (3%)	1	100.00%	Root Alignment
Mutiple Pairwise	1BQ6, 1FM8 2GAS, 1EYQ & 1CGK	00:00:00	9 MB (3%)	1	100.00%	Pass 1
		00:00:00	9 MB (3%)	1	100.00%	Pass 2
		00:00:00	12 MB (3%)	1	100.00%	Align Node
		00:00:00	12 MB (3%)	1	100.00%	Root Alignment
		00:00:00	12 MB (3%)	2	100.00%	Root Alignment
		00:00:00	12 MB (3%)	3	100.00%	Refine Biparts

TABLE 6: STATISTICAL PARAMETERS OF THE PAIRWISE AND MULTIPLE ALIGNMENTS BY THE HMMER PACKAGE

Alignment	**Number of sequences**	**Name of sequences**	**Number of columns**	**Average score**	**Min. score**	**Max. score**	**Standard Deviation**
Pairwise	2	>1BQ6 & >1FM8	396	765.60 bits	432.01 bits	1099.20 bits	471.77 bits
Multiple	5	>1BQ6,>1FM 8, >1EYQ, >2GAS & >1CGK	403	640.14 bits	354.26 bits	926.18 bits	285.96 bits

TABLE 7: STATISTICS RELATED TO SCORE AND IDENTITY OF PAIR WISE SEQUENCE ALIGNMENT BY SEQUENCE ANALYSIS TOOL

Algorithm	Scoring Scheme		Scoring Matrices			
	Match (1), Mismatch (-1) and Gap (-1)		PAM 250		BLOSUM 80	
	Score	Identity	Score	Identity	Score	Identity
Smith & Waterman (local)	5	100	31	100	39	36.3
Needleman & Wunsch (global)	-195	25.8	-1401	22.1	-563	22.9
Crochemore, Landau & Ziv-Ukelson (Local)	5	100	31	100	39	36.3
Crochemore, Landau & Ziv-Ukelson (Global)	-195	25.8	-1401	22.1	-563	22.9
Smith & Waterman (local)	5	100	31	100	39	36.3

TABLE 8: STATISTICS RELATED TO THE PAIRWISE AND MULTIPLE ALIGNMENTS BY GENEIOUS TOOL

Alignment	Length	Number of sequences	Pairwise % Identity	Identical sites	Molecular Weight (mean)	Isoelectric Point (mean)
Pairwise	388	2	16.2	63	33.201kDa	5.62
Multiple	891	5	88.1	0	33.439kDa	5.50

TABLE 9: FEW DEEM PARAMETERS RELATED TO THE MSA TOOLS

Tool	Strengths	Weaknesses
Geneious	• Support online database connectivity • Free public API, plugin can be shared	• Less versatile as compared to DNAMAN • Requires java runtime environment
ClustalX	• One of the oldest of MSA tools • No limitation on the length of sequences	• Less accurate as compared to modern tools • Support sequences in fasta format
GoCore	• Built-in sample data available • Coloured visualization	• Well versed with MS-Excel • Support only fasta sequences
Strap	• Support online database connectivity • Aligns proteins by sequence and 3D-structures	• Requires java runtime environment • Requires online access to the website for first MSA generation
HMMER	• Use Probalistic basis • Profile method to search databases using MSA	• Less skill more manual intervention • Make poor models of RNA's
MUSCLE	• There is no limitation on input sequences • Support large size sequences	• Support only fasta sequences • More manual intervention
Sequence Analysis	• Allows conversion between nucleotide sequences • Multithreaded	• Requires java runtime environment
DNAMAN	• A very versatile tool for sequence analysis • Provides an integrated Web-Browser to access the internet	• Protein sequences need to be confirmed

REFERENCES

[1] Needleman, S.B. and Wunsch, C.D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol., 48 (3): pp. 443–53.

[2] Smith, T.F. and Waterman, M.S. (1981). "Identification of common molecular subsequences.", Journal of Molecular Biology, 147 : pp. 195–197.

[3] Wilbur, W.J. and Lipman, D.J. (1984). "On the statistical significance of nucleic acid similarities.", Nucleic Acids Research, 12 (1) : pp. 215–226.

[4] Higgins, D.G. and Sharp, P.M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS 5,151–153.

[5] Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). "Hidden Markov models in computational biology: Applications to protein modeling.", J. Mol. Biol., 235: pp. 1501–1531.

[6] Higgins, D. G., Thompson, J. D. and Gibson, T. J. (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol., 266, 383-402.

[7] Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., and Higgins D.G. (1997). "The clustal x windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Res., 25 (24): pp. 4876–4882.

[8] Eddy, S. R. (1998). "Profile hidden Markov models". Bioinformatics, 14, pp. 755–763.

[9] Thompson Jukie D., Plewniak F. and Poch O. (1999). "A Comprehensive Comparison of multiple sequence alignment programs". Nucleic Acids Res., 27 (13): pp. 2682–2690.

[10] Notredame, C., Higgins, D.G., Heringa, J. (2000). "T-Coffee: a novel method for fast and accurate multiple sequence alignment". J. Mol. Bio., 302: pp. 205–217

[11] Christoph, G. and Cornelius, F. (2001). "STRAP : Editor for STRructural Alignments of Protein.", Institute of Biochemistry, 17 (4): pp. 377–378.

[12] Edger, C. R. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput.", Nucleic Acids Research, 32 (5) : pp. 1792–1797.

[13] Edgar, Robert C (2004), MUSCLE: A Multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5(1): 113.

[14] Gilchrist RB, Ritter LJ, Cranfield M, Jeffery LA, Amato F, Scott SJ, Myllymaa S, Kaivo-Oja N, Lankinen H, Mottershead DG, Groome NP, Ritvos O (2004)."Immunoneutralization of growth differentiation factor 9 reveals it partially accounts for mouse oocyte mitogenic activity". Biol Reprod. 2004 Sep;71 (3): 732–739

[15] Paulo, A.S.N., Zhouzhi, W. and Elisabeth, R.M.T. (2006). "The accuracy of several multiple sequence alignment programs for proteins", BMC Bioinformatics.

[16] Rastogi, S.C., Mendiratta, N. and Rastogi, P. (2006). "Bioinformatics Methods and Applications", Prentice Hall India, New Delhi.

[17] Asieh S.,Rodziah Binti Atan, KhairinaTajul Arifin, Masrah Azrifah Binti Azmi Murad (2009)."Comparison and Evaluation of Multiple Sequence Alignment Tools In Bininformatics",IJCSNS, 9(7) : pp. 51–56

[18] http://www.stanford.edu/~dsellis/papers/MSAComparison.pdf

Targeted Drug Discovery using Open Source Public Tools

Afreen Sayed
Department of Microbiology, Abeda Inamdar Sr College, Pune
e-mail: meet_afreen@rediffmail.com

***Abstract*—Objective To identify and annotate the different glycan modifying enzymes in Mycobacterium tuberculosis. Identify an appropriate reference strain in prokaryotes or eukaryotes which can be used as a standard to compare the sequence and identify the glycan modifying enzymes. Finding glycan modifying enzymes in Mtb using reference strain. Identify the nature of enzyme. Use combination of bioinformatics tool to identify domains and patterns in the enzyme.**

Research and Design method: The glycosylation related genes need to be listed from the different reference genomes i.e. S.cerevisiae, C.jejuni, N.gonorrorhoea. BLAST against the TB genome available in TBDB, Tuberculist site. The product name was assigned based on the Blast hit based the blast bit score value or E-value. The optimum E-value of the range between 1×10-52 to 1×10-2 used for choosing the best hit result.

In this article we have used the public data base like the UniProt, CaZY and Tuberculist along with the Open source mining tools PsiBlast, Pfam, Interproscan, Prodom, COGnitor,CDD and THMM to extract records and process them to perform annotations of the genetic sequences. The output of the mining tools would be used to detect pattern and motifs of the proteins for targeted drug designing.

Results: Genes encoding for glycan modifying enzymes were annotated. The proteins coded by these genes were further used in Phase -2 for further evaluation and in-vitro studies.

Keywords: Glycan modifying enzyme, BLAST, Uniprot, Mycobacterium tuberculosis.

I. INTRODUCTION

The recent advances in genomics and proteomics have lead to the generation of a large amount of data. But many of the data are annotated as ‘‘uncharacterized,’’ ‘‘hypothetical,’’ or ‘‘unknown function’’[1]. Many databases are available which have such kind of data. To reduce such data which has no meaning we need to re-annotate the proteins to replace the un-meaningful data in the database with meaningful data. In bioinformatics and computational biology many tools are available which aid in annotation of proteins in-silico[2]. There are many public databases for genomic sequences and protein sequences which are accessible on the internet at the click of a button. Also various bioinformatics tools are available freely online for use [1, 2].

Though five decades are over since the introduction of first antibiotic against TB, no great developments have been achieved in finding an active compound or drug against this dreadful disease. There is a need for a new drug against tuberculosis, because the current treatment is prolonged and treatment with short duration needs to be found, there are emergence of multi drug resistant tubercle bacilli thus new drug for combating these strains is the need of the hour and also many latent infections are prevalent which need to be treated effectively. Thus there is a need for new, novel drugs for tuberculosis. The discovery of the mycobacterial glycoconjugates which are recognized by the immune system mannose-specific lectins, constitutes a major challenge to gain insight into the host-pathogen molecular interaction. The glycan processing enzymes have been relatively understudied in *Mycobacterium tuberculosis*[3] .

In the case study we try to re-annotate the genome of *Mycobacterium tuberculosis* to identify glycan modifying enzymes, for targeted drug discovery. The proteins would be annotated and then structurally compared to reveal motif’s and patterns in them which would be used for designing drug against the protein. We focus in this case study on the use of open source databases like Uniprot, cAZy, Tuberculist and the various bioinformatics tools like Psi-BLAST (Position specific iterative BLAST) , Cognitor (Clusters of Orthologous Groups of proteins), CDD (Conserved domains databases) used for finding the conserved domains in protein, TMHMM (Trans membrane helicases of Micobacterium) to detect the number of transmembrane helices in proteins, SCANPROSITE for High level of annotation and PRODOM (Protein domain families) for finding the major domains across the protein family and Pfam (Protein families) to detect the family for the given protein[3]. All these tools help in one dimensional annotation of protein in-silico. Our objective is to identify and annotated the different glycan modifying enzymes in *Mycobacterium tuberculosis* genome. Further two dimensional analyses with the structure can be done to find motif’s and patterns which can then be targeted by designing drugs against them.

Tools and Resources used for analysis and annotation[3]:

The Standard operative practices were made available by the OSDD and was followed for the annotation. The reference strains *Compylobacter jejuni*, *Saccharomyces cerevisiae*, *Streptomyces*, *Pseudomonas* were used to find the glycome modifying enzymes from Cazy (Carbohydrate-Active enzymes Database) which gave the Uniprot Id's of these protein the sequence of these proteins was taken from Uniprot. These enzyme sequences were then compared with the enzyme sequences of *Mycobacterium tuberculosis* in the database Tuberculist. The best hit with large BIT score and optimum E-value chosen for further analysis of finding the protein domain and family by comparing the protein sequences in many different bioinformatics tools. The database and various tools used with their URLs are given in figure1.

Step 1: The first public Database used is cAZy (Carbohydrate-Active enzymes Database). The glycoside modyfying enzymes in the refernce strains Compylobacter jejuni, Saccharomyces cerevisiae, Streptomyces, Pseudomonas are extracted from this site and the uniprot identification number is got. The CAZy database describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. This database has been online since 1998, CAZy specializes to display and analysis of genomic, structural and biochemical information on Carbohydrate-Active Enzymes (CAZymes)[3].

The database is publicly accessible either by browsing sequence-based families or by browsing the content of genomes in carbohydrate-active enzymes. New genomes are added regularly shortly after they appear in the daily releases of GenBank[4]. New families are created based on published evidence for the activity of at least one member of the family and all families are regularly updated, both in content and in description.The CAZy database covers all carbohydrate-active enzymes across organisms and across subfields of glycosciences.The enzyme classes coverd in CAZY **are** Glycoside Hydrolases (GHs)**,** GlycosylTransferases (GTs) , Polysaccharide Lyases (PLs) and Carbohydrate Esterases[4] . The website homepage is shown in figure 2. The result of a search is shown in figure 3.

Step 2: UNIPROT (Universal Protein Resource) http://www.uniprot.org. The Uniprot id's got from CAZY database are then searched in the Uniprot homepage in UniProtKB for getting the protein sequence [3].

TABLE 1: TABLE OF THE DATABASES OR TOOLS USED AND THEIR IMPORTANCE

TOOLS/ DATABASES	**IMPORTANCE**
CAZy- http://www.cazy.com	Carbohydrate-Active enzymes Database.
UNIPROT-http://www.uniprot.org	Universal Protein Resource
TUBERCULIST-http://genolist.pasteur.fr/TubercuList/	Database on Mycobacterium tuberculosis genetics.
PSI-BLAST-http://www.ncbi.nlm.nih.gov	Position specific iterative BLAST .
COGnitor - http://www.ncbi.nlm.nih.gov/COG/	Clusters of Orthologous Groups of proteins (COGs)
CDD-http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	Conserved domains databases
TMHMM- http://www.cbs.dtu.dk/services/TMHMM/	Trans membrane helicases of Micobacterium.
SCANPROSITE- http://expasy.org/prosite	High level of annotation

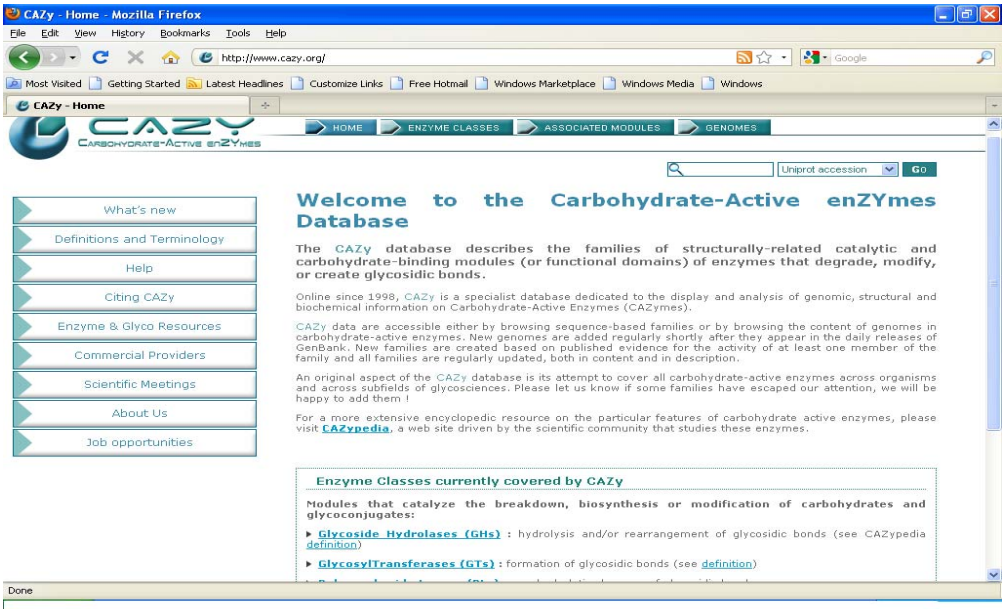


Fig. 2: Hompage of CAZY

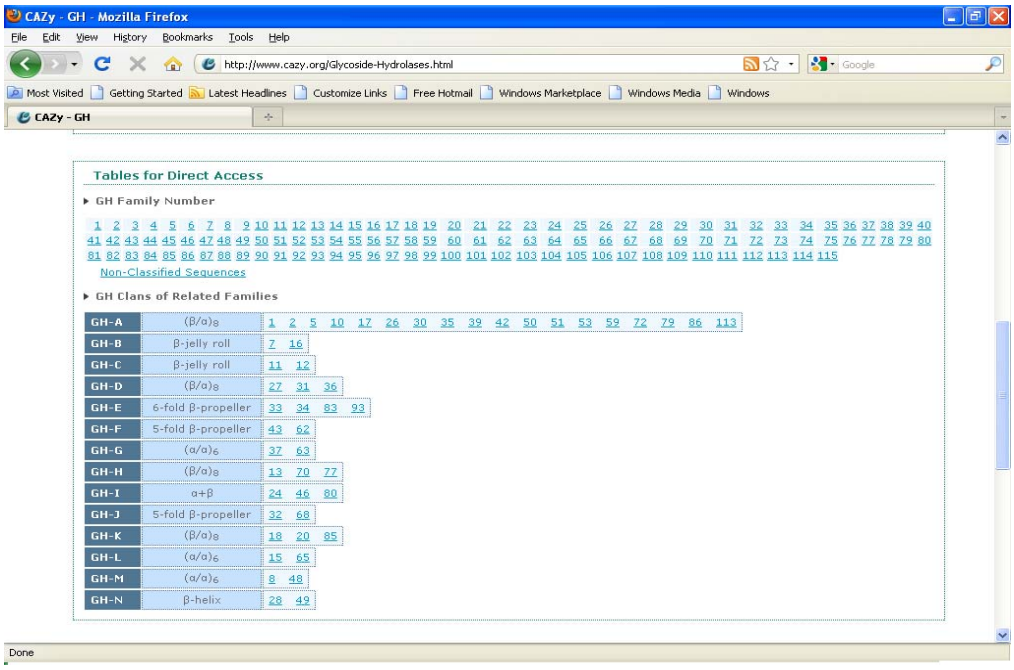


Fig 3: Search Result for Glycoside Hydrolase Enzyme Showing all the Family Number and Class of Related Proteins

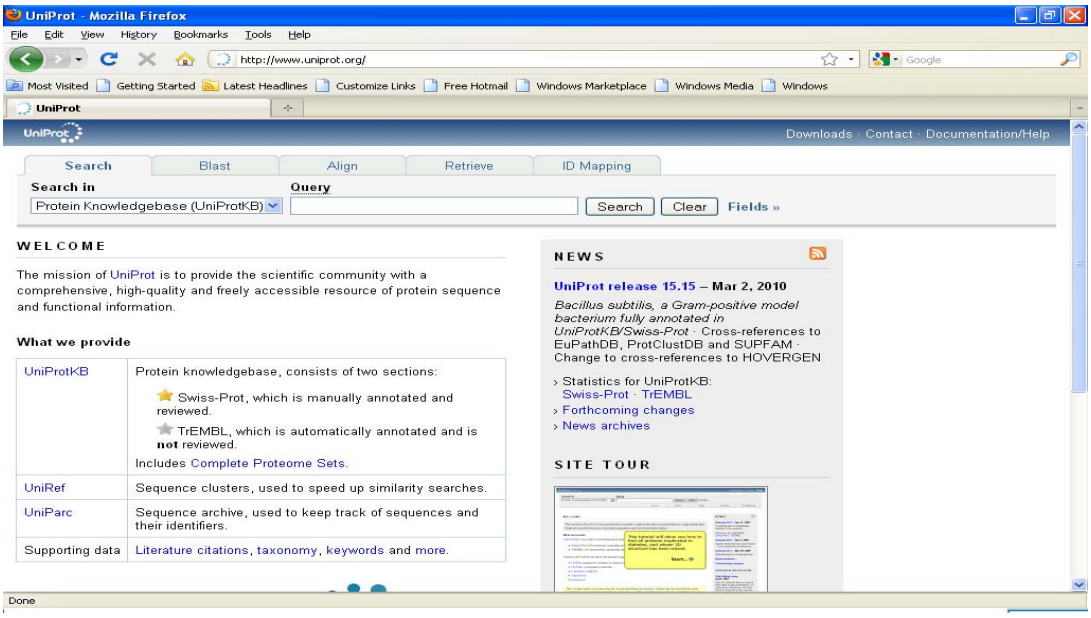


Fig. 4: Homepage of Uniprot

UniProt provides the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. UniProt is the Universal Protein resource, it is a central repository of protein data which was created by combining the Swiss-Prot, TrEMBL and PIR-PSD databases. UniProt is based on protein sequences, derived from genome sequencing projects. The UniProt Consortium comprises the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR)[5]. EBI located at the Wellcome Trust Genome Campus in Hinxton, UK, hosts a large resource of bioinformatics databases and services. SIB, located in Geneva, Switzerland, maintains the ExPASy (Expert Protein Analysis System) servers that are a central resource for proteomics tools and databases[6].

UniProt provides four core databases: UniProt Knowledgebase (UniProtKB) is a protein database that is curated by experts, consisting of two sections. UniProtKB/Swiss-Prot (containing reviewed, manually annotated entries) and UniProtKB/TrEMBL (containing unreviewed, automatically annotated entries. UniProt Archive (UniParc) is a comprehensive and non-

redundant database, which contains all the protein sequences from the main, publicly available protein sequence databases. UniProt Reference Clusters (UniRef) consist of three databases of clustered sets of protein sequences from UniProtKB and selected UniParc records .Finally UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for metagenomic and environmental data[5,6]. Figure 4 shows the homepage of Uniprot, while figure 5 shows the result of the search where protein sequence is displayed.

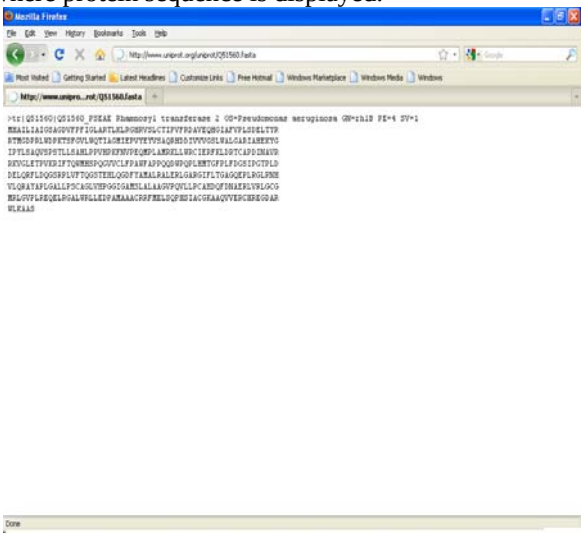


Fig. 5: The Result of the Search in Uniprot Where Protein Sequence is Displayed

Step 3: After retrieval of sequence the sequence is searched with the *Mycobacterium tuberculosis* genome by doing a protein BLAST in the Tuberculist web page- http://genolist.pasteur.fr/TubercuList/. Here the protein sequence is aligned with *Mycobacterium tuberculosis* protein sequence and Rvid and sequence of the glycoside modyfyiong enzyme in *Mycobacterium tuberculosis* was obtained. The protein with best hit, large BIT score and optimum E-value was chosen for further analysis[3].

The Tuberculist is a server which is constructed around a database, that dedicated to the analysis of the genomes of the tubercle bacilli. The main aim of this database is to collect and integrate various aspects of the genomic information from *M. africanum*, *M. bovis*, *M. bovis BCG*, *M. canetti*, *M. microti*, and *M. tuberculosis*[7]. TubercuList provides a complete database of DNA and protein sequences of the strain *M. tuberculosis* H37Rv, which are linked to the relevant annotations and functional assignments. Browsing is easy and the data can be viewed and retried for information, using various criteria (gene names, location, keywords, etc.)[8].The server is provided by Institut Pasteur.

In figure 6, we have the Tuberculist web page and figure 7. shows the result of BLAST in Tuberculist server with the protein sequence got from Uniprot. The best hit is selected and the sequence of that hit is taken from Tuberculist site (Figure 8).

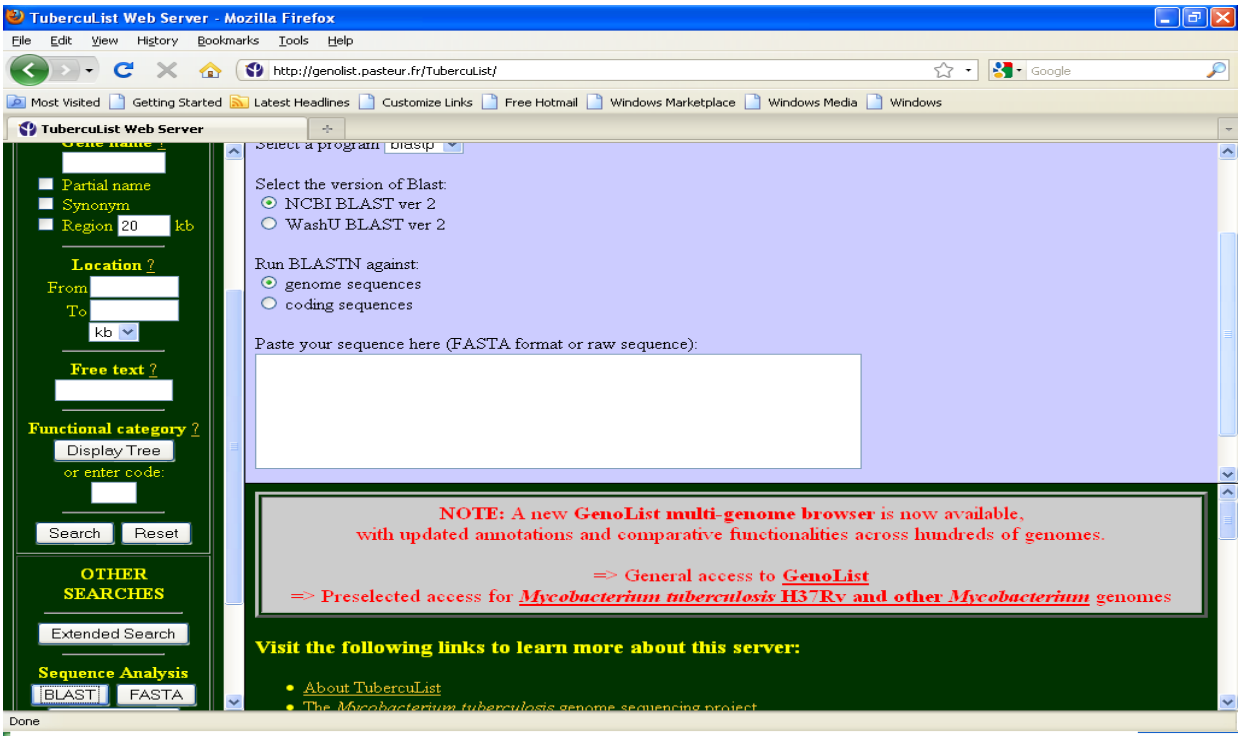


Fig. 6: We have the Tuberculist Web Page

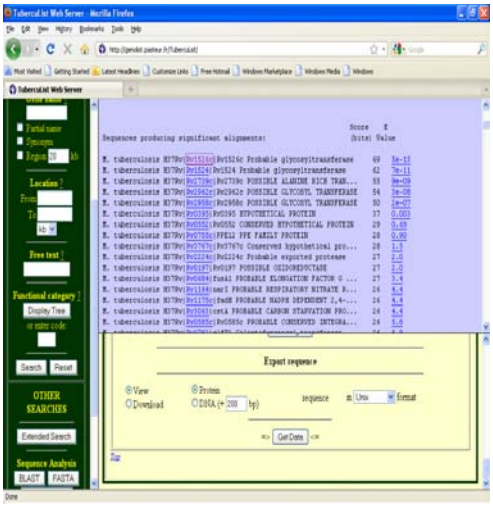


Fig. 7: Shows the Result of BLAST in Tuberculist Server with the Protein Sequence Got from Uniprot

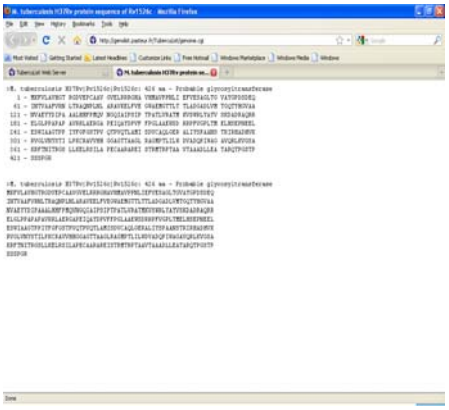


Fig. 8: The Sequence of the Best Hit from Tuberculist Site

Step 4: NCBI's BLAST stands for Basic Local Alignment Search Tool.One of the programs in BLAST is the PSI-BLAST i.e. Position Specific Iterative – BLAST. The sequence of protein got from Tuberculist of Mycobacterium tuberculosis is then aligned with the sequences in NCBI using PSI-BLAST, the search reveals a number of probable homologs.

Position specific iterative BLAST (PSI-BLAST) refers to a feature of BLAST 2.0 in which a profile (or position specific scoring matrix, PSSM) is automatically constructed from a multiple alignment of the highest scoring hits got by an initial BLAST search. The PSSM is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions in the sequence receive high scores and weakly conserved positions receive low scores nearing zero [9]. The profile is used to perform a second sequence alingnment and so on [9]. BLAST search and the results of each consecutive step is used to refine the profile. This iterative searching strategy results in increased sensitivity[10]. BLAST searches to identify even weak homologies to annotated entries in the database. It is used to find evolutionary related sequences. PSI-BLAST is an important tool for predicting both biochemical activities and function from sequence relationships. Its algorithm is designed to conduct further iterations of the search and to extend the search to distantly related homologues[10,11].

The figure 9,shows homepage of NCBI's BLAST. The result of the psi-blast is shown in figure10,the result show enzyme Glycosyl Transferase.

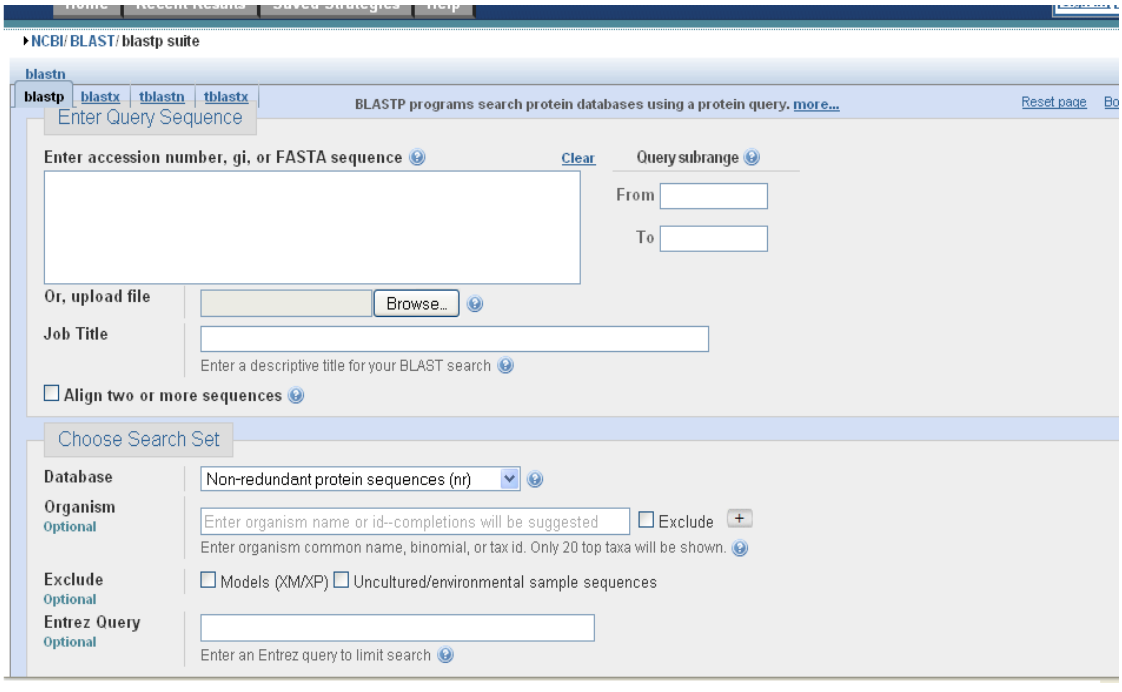
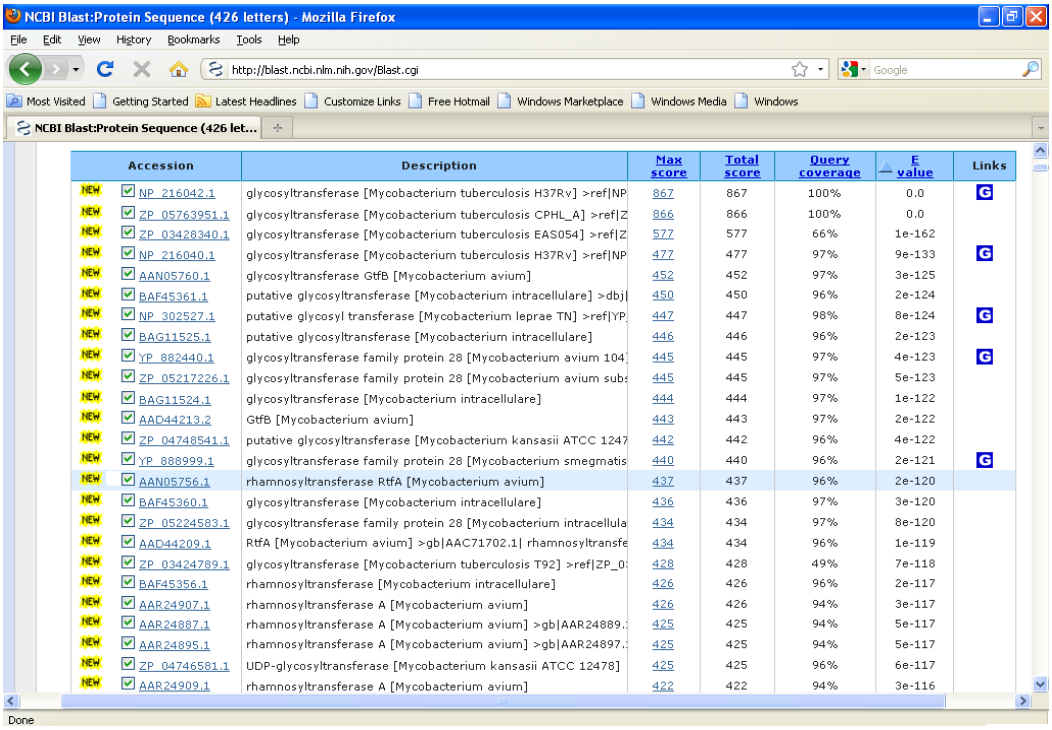


Fig. 9: Homepage of NCBI's BLAST



Accession	Description	Max score	Total score	Query coverage	E value	Links
NP_216042.1	glycosyltransferase [Mycobacterium tuberculosis H37Rv] >ref|NP	867	867	100%	0.0	G
ZP_05763951.1	glycosyltransferase [Mycobacterium tuberculosis CPHL_A] >ref|Z	866	866	100%	0.0	
ZP_03428340.1	glycosyltransferase [Mycobacterium tuberculosis EAS054] >ref|Z	577	577	66%	1e-162	
NP_216040.1	glycosyltransferase [Mycobacterium tuberculosis H37Rv] >ref|NP	477	477	97%	9e-133	G
AAN05760.1	glycosyltransferase GtfB [Mycobacterium avium]	452	452	97%	3e-125	
BAF45361.1	putative glycosyltransferase [Mycobacterium intracellulare] >dbj|	450	450	96%	2e-124	
NP_302527.1	putative glycosyl transferase [Mycobacterium leprae TN] >ref|YP	447	447	98%	8e-124	G
BAG11525.1	putative glycosyltransferase [Mycobacterium intracellulare]	446	446	96%	2e-123	
YP_882440.1	glycosyltransferase family protein 28 [Mycobacterium avium 104]	445	445	97%	4e-123	G
ZP_05217226.1	glycosyltransferase family protein 28 [Mycobacterium avium subs	445	445	97%	5e-123	
BAG11524.1	glycosyltransferase [Mycobacterium intracellulare]	444	444	97%	1e-122	
AAD44213.2	GtfB [Mycobacterium avium]	443	443	97%	2e-122	
ZP_04748541.1	putative glycosyltransferase [Mycobacterium kansasii ATCC 1247	442	442	96%	4e-122	
YP_888999.1	glycosyltransferase family protein 28 [Mycobacterium smegmatis	440	440	96%	2e-121	G
AAN05756.1	rhamnosyltransferase RtfA [Mycobacterium avium]	437	437	96%	2e-120	
BAF45360.1	glycosyltransferase [Mycobacterium intracellulare]	436	436	97%	3e-120	
ZP_05224583.1	glycosyltransferase family protein 28 [Mycobacterium intracellula	434	434	97%	8e-120	
AAD44209.1	RtfA [Mycobacterium avium] >gb|AAC71702.1| rhamnosyltransfe	434	434	96%	1e-119	
ZP_03424789.1	glycosyltransferase [Mycobacterium tuberculosis T92] >ref|ZP_0	428	428	49%	7e-118	
BAF45356.1	rhamnosyltransferase [Mycobacterium intracellulare]	426	426	96%	2e-117	
AAR24907.1	rhamnosyltransferase A [Mycobacterium avium]	426	426	94%	3e-117	
AAR24887.1	rhamnosyltransferase A [Mycobacterium avium] >gb|AAR24889.	425	425	94%	5e-117	
AAR24895.1	rhamnosyltransferase A [Mycobacterium avium] >gb|AAR24897.	425	425	94%	5e-117	
ZP_04746581.1	UDP-glycosyltransferase [Mycobacterium kansasii ATCC 12478]	425	425	96%	6e-117	
AAR24909.1	rhamnosyltransferase A [Mycobacterium avium]	422	422	94%	3e-116	

Fig.10: The Result of the Psi-blast is Shown, the Result Shows Enzyme Glycosyl Transferase

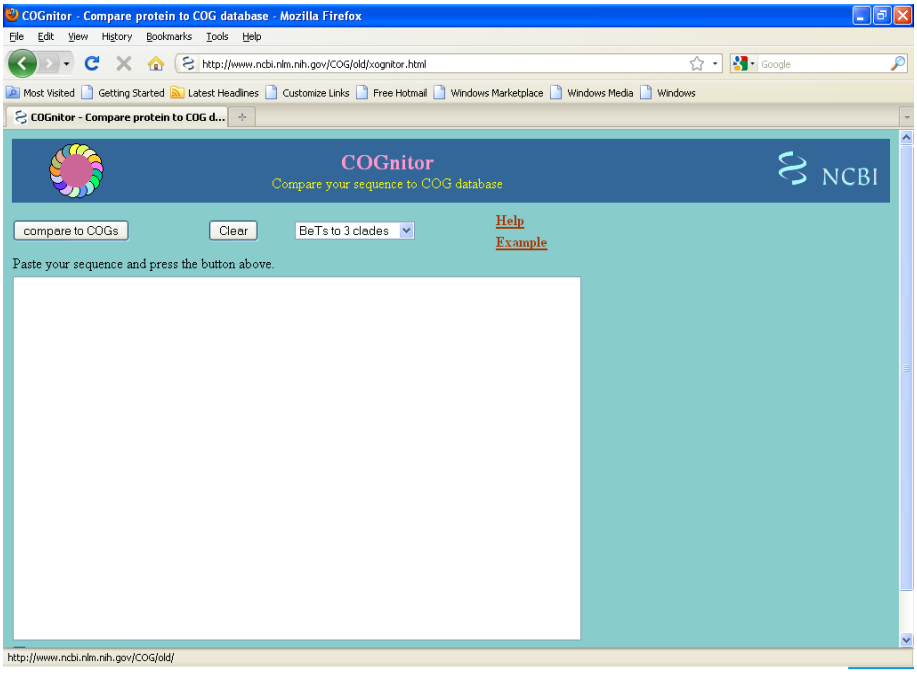


Fig. 11: The Homepage of COGnitor Where Sequence can be Pasted

Step 5: COGnitor i.e. Clusters of Orthologous Groups of proteins (COGs) - http://www.ncbi.nlm.nih.gov/COG/. The sequence of protein is now searched using COGnitor after psi-blast to find phylogenetic lineage and conserved sequence[3

NCBI's COGnito-Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.COGnitor is an attempt of classifying protyein encoded by 21complete genomes of bacteria, archaea and eukaryotes [12]. The program is used to fit new proteins in the COGs and also applied to find functional and Phylogenetic annotation of new

genomes [13]. Orthologs are direct evolutionary counterparts which are got by vertical descent thus it is necessary to study them in finding evolutionary related proteins. The orthologous proteins have similar structure of domain and function, thus by using well characterized COGs of organisms and finding protein function in orthologous organisms that are less studied can be used in annotation of proteins[12,13].

The figure 11 shows the Homepage of COGnitor where sequence can be pasted. The result is seen in figure 12.

Step 6: CDD- Conserved domains databases The URL is as follows:

http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml . The sequence of protein is now searched using CDD to find the conserved domain in the protein[3]

NCBI's CDD Conserved domains databases is a protein annotation resource consisting of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. The Database includes NCBI-curated domains, which use 3D-structure information to explicitly define domain boundaries and provide insights into the sequence, structure and functional relationships between proteins. Domain models imported from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAM) are also available[14]. The Database has tools for Conserved domain search where the query sequence can be searched using RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs). The results of CD-Search are presented as an annotation of protein domains on the user query sequence. Other tools include Conserved Domain Architecture Retrieval Tool (CDART) which performs similarity searches of the Entrez Protein database based on domain architecture, defined as the sequential order of conserved domains in protein queries and the CDTree is a helper application for your web browser allowing user to interactively view and examine conserved domain hierarchies curated at NCBI[14].

The figure 13 shows Homepage of CDD where sequence can be pasted in CD-Search. The results got are shown in figure 14 showind domains of glycosyl transferase enzyme.

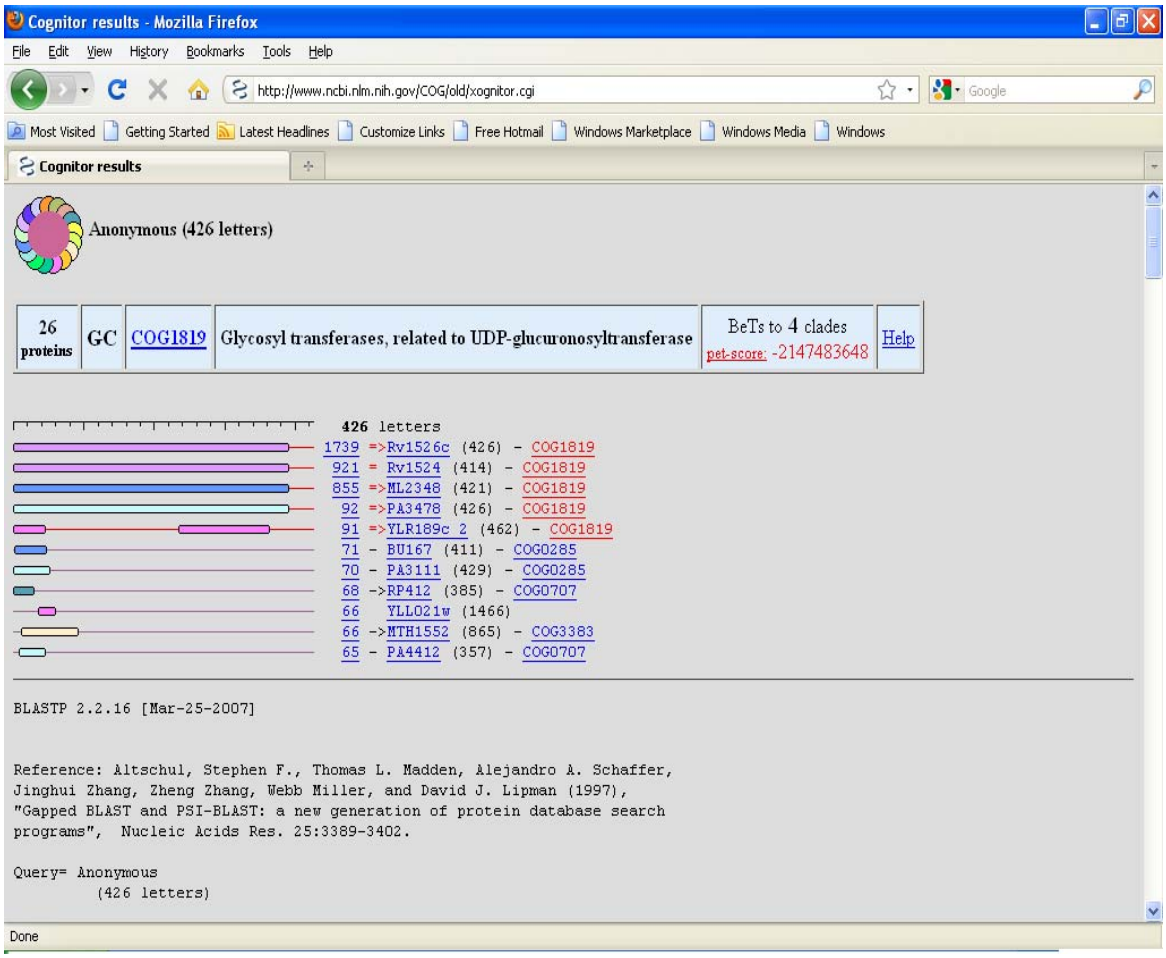


Fig. 12: Showing the Result as Glycosl Transferase Enzyme

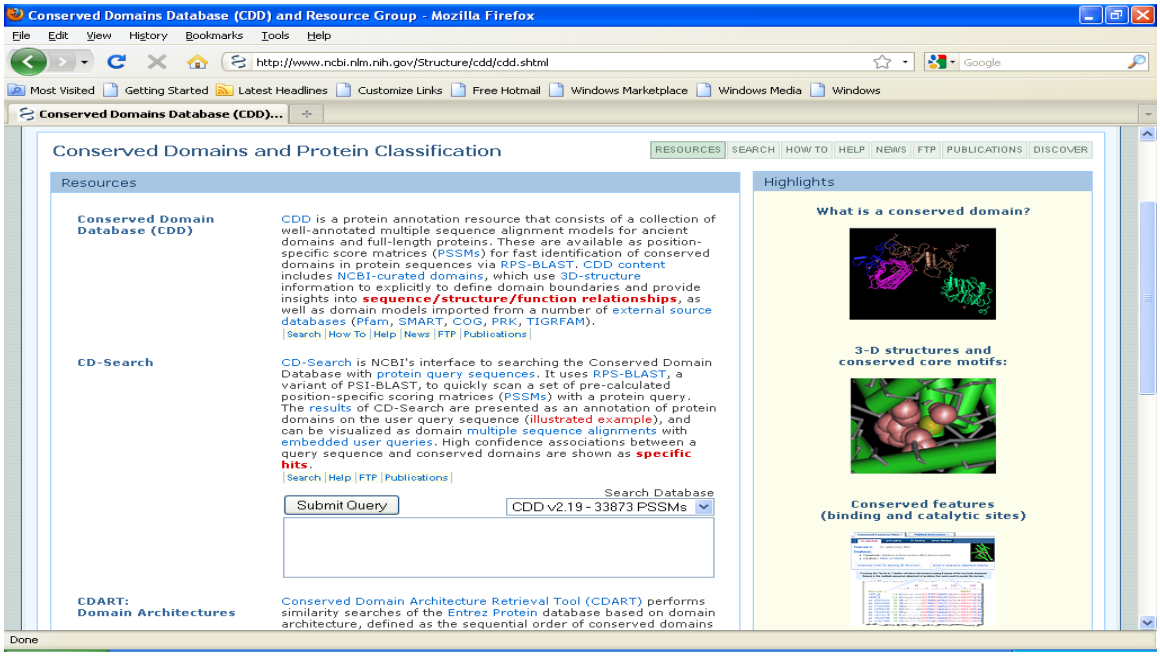


Fig. 13: Homepage of CDD where Sequence can be Pasted in CD-Search

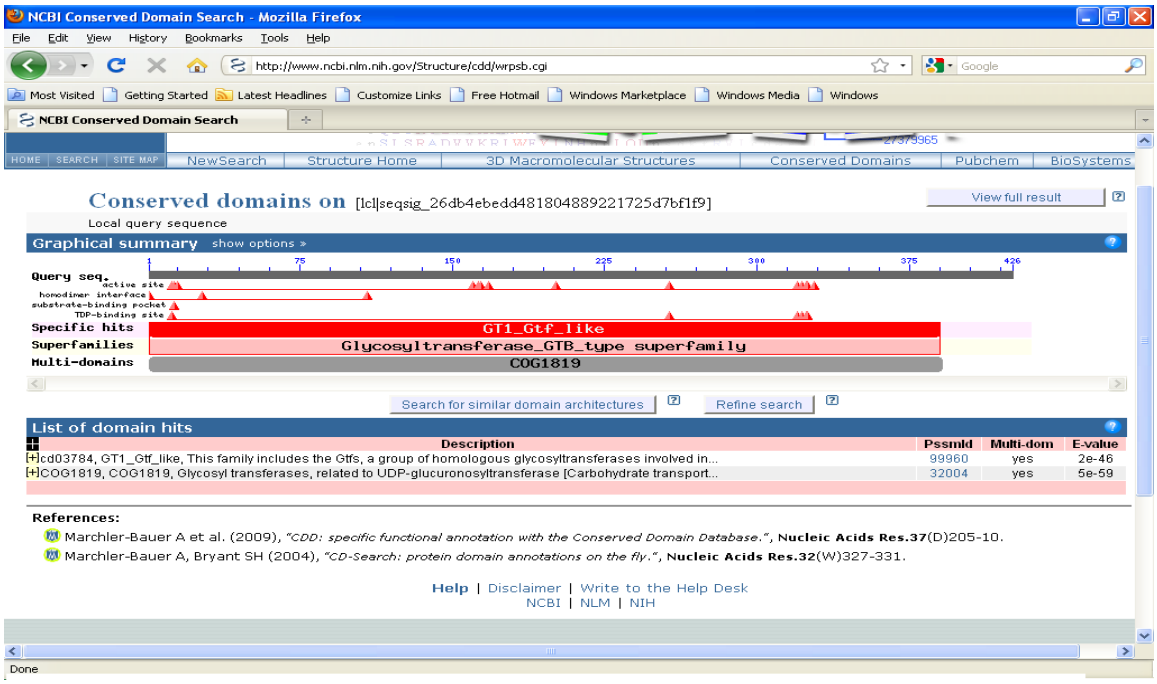


Fig. 14: The Results Got are Seen Showing Domains of Glycosyl Transferase Enzyme

Step 7: TMHMM the Transmembrane helices in proteins available at the website: http://www.cbs.dtu.dk/services/TMHMM/. The protein sequence is pasted in the search box for finding out if the protein contains any transmembrane helices.

TMHMM is one of CBS Prediction Servers for predicting the Transmembrane helices in proteins. TMHMM is based on a hidden Markov model [16]. TMHMM's performance shows that it correctly predicts 97–98 % of the transmembrane helices. TMHMM can discriminate between soluble and membrane proteins having specificity and sensitivity better than 99 %, which may decrease when signal peptides are present [16]. TMHMM Server v. 2.0 can be downloaded from the CBS Software Package or it can be run on the server site itself [15].

The figure 15 shows the Homepage where sequence can be pasted for search. The results are seen in figure 16 where the search shows no Transmembrane helices in the query sequence.

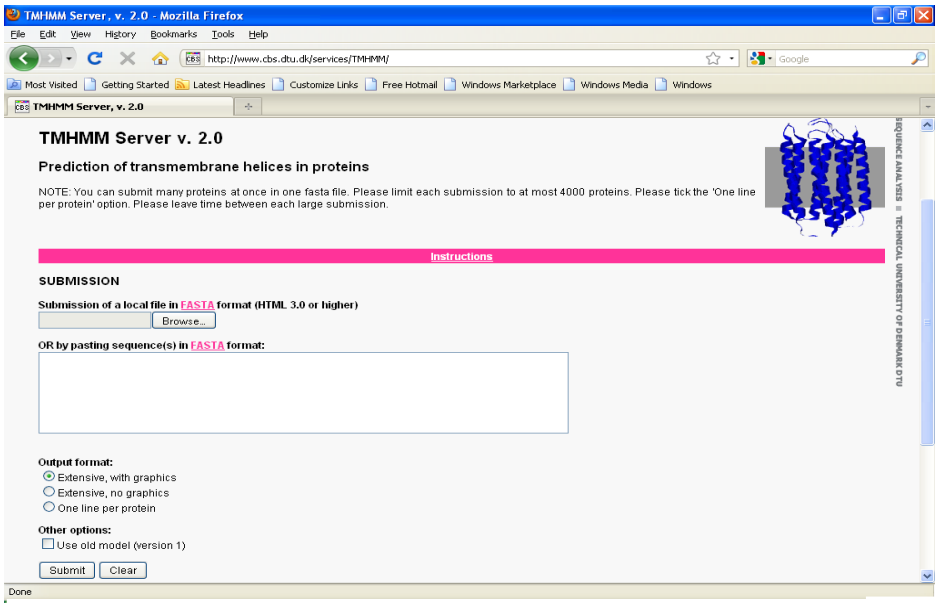


Fig. 15: Shows the Homepage where Sequence can be Pasted for Search

Fig. 16: The Search Result Shows no Transmembrane Helices in the Query Sequence

Step 8: PRODOM(Protein Domain Families) is available at the website–

http://prodom.prabi.fr/prodom/current/html/home.php. In the PRODOM hompepage the protein sequence is pasted in the box provided to compare the sequence withPRODOM.The program choosen is NCBI-BLAST and method is multiple alignment. The result gives the best hit for the closest domain in the family against the query sequence[3].

ProDom is a protein domain family database, generated automatically by clustering homologous protein segments. The ProDom building procedure uses MKDOM2 program which is based on recursive PSI-BLAST searches[17,18]. The source protein sequences are in the form of non-fragmentary sequences which are derived from UniProtKB (Swiss-Prot and TrEMBL databases)[17] . ProDom was first established in 1993 and maintained by the Laboratoire de Génétique Cellulaire and the Laboratoire de Interactions Plantes-Microorganismes (INRA/CNRS) in Toulouse. It is now maintained by the PRABI (bioinformatics center of Rhone-Alpes)[17]. The ProDom database consists of domain family entries.

Figure 17 shows the Homepage of PRODOM where the query sequence can be pasted. The result showing the closest domain in the family is seen in figure 18.

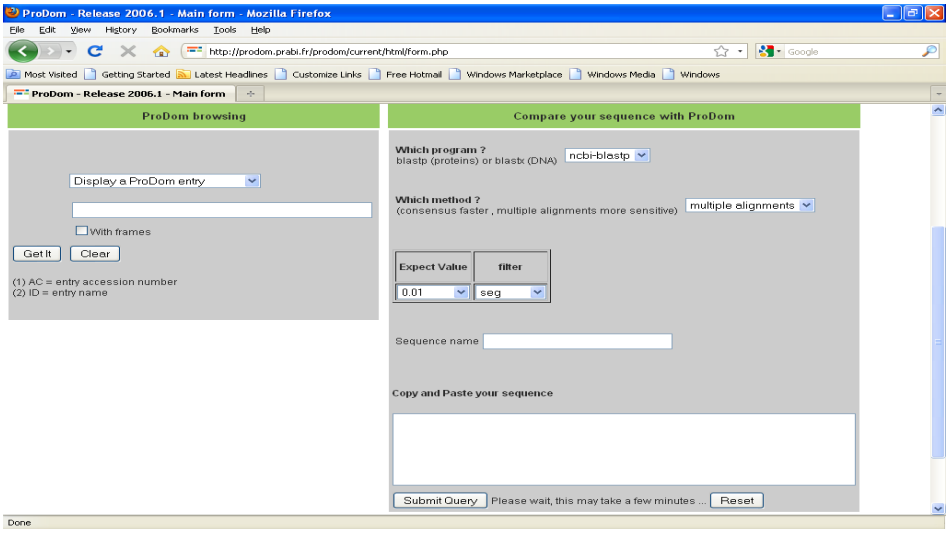


Fig. 17: The Homepage of PRODOM where the Query Sequence can be Pasted

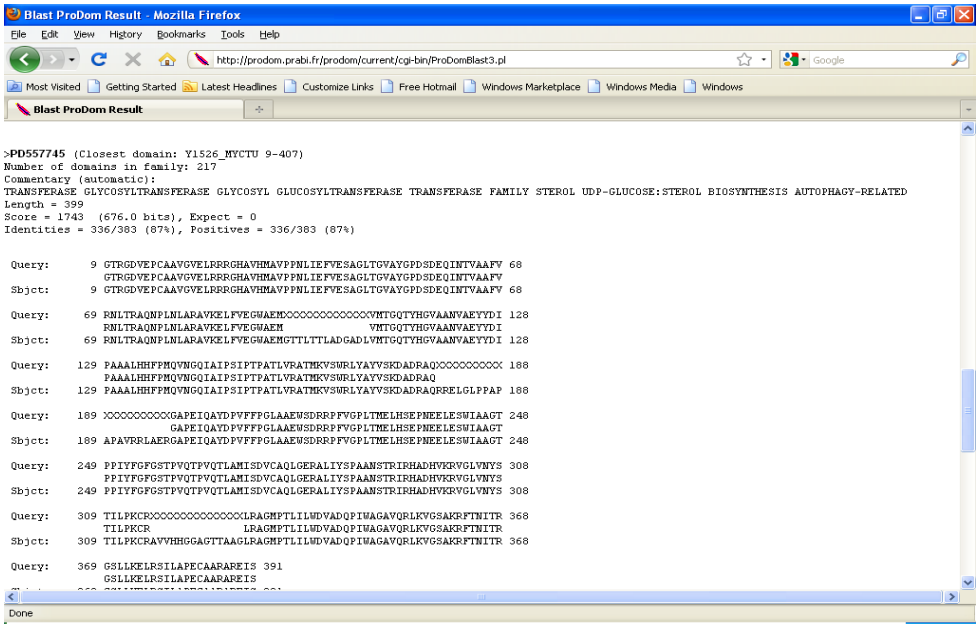


Fig. 18: The Result Showing the Closest Domain in the Family

Step 9: Pfam Protein families the website is: http://www.pfam.sanger.ac.uk. The query sequence is pated in the homepage of Pfam to find the family of the protein[3].

Pfam is a collection of Multiple sequence analysis of proteins and profile hidden Markov Models of proteins. It comprises 75 per cent of known proteins to form a library of protein families. The database is open access resource which was established at the Wellcome Trust Sanger Institute in 1998, provide a tool for experimental, computational and evolutionary biologists to classify protein sequences and find evolutionary relationship [19]. . The database comprises two main collections of information as Pfam-A which comprises high-quality entries which have been curated manually and Pfam-B which contains automatically curated entries that are of a lower quality but add valuable coverage for regions that are not yet curated and stored in Pfam-A. Pfam database contains nearly 12,000 curated protein families. The latest version of Pfam-6.6 contains 3017 families which match with the Swis-prot and Tremble database with 68% similarity [20].

The figure 19 shows the Homepage of Pfam where the query sequence can be pasted in sequence search option. The result showing the family for the query search is seen with the significant and insignificant hit in figure 20.

Step 10: Espasy's ScanProsite the website is http://expasy.org/prosite. Espasy's Scanprosite is used for high level of annotation of the query sequence.

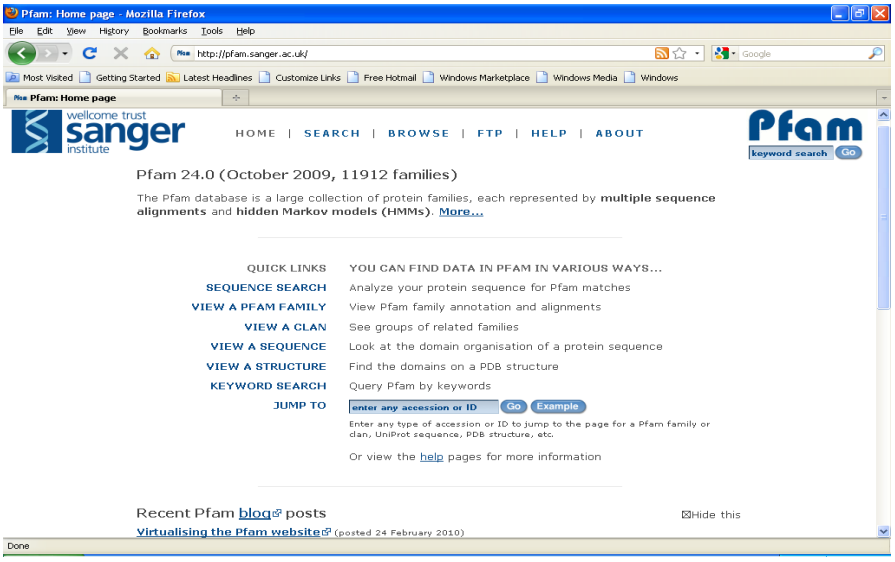
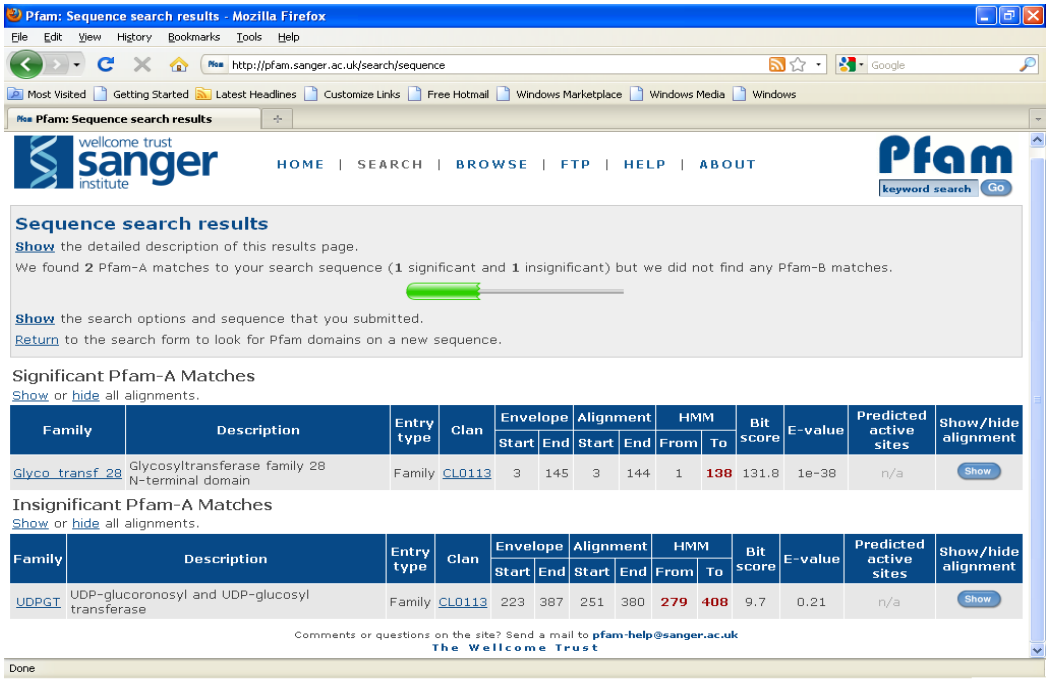


Fig. 19: The Homepage of Pfam where the Query Sequence can be Pasted in Sequence Search Option



Family	Description	Entry type	Clan	Envelope Start	Envelope End	Alignment Start	Alignment End	HMM From	HMM To	Bit score	E-value	Predicted active sites	Show/hide alignment
Glyco_transf_28	Glycosyltransferase family 28 N-terminal domain	Family	CL0113	3	145	3	144	1	138	131.8	1e-38	n/a	Show

Family	Description	Entry type	Clan	Envelope Start	Envelope End	Alignment Start	Alignment End	HMM From	HMM To	Bit score	E-value	Predicted active sites	Show/hide alignment
UDPGT	UDP-glucoronosyl and UDP-glucosyl transferase	Family	CL0113	223	387	251	380	279	408	9.7	0.21	n/a	Show

Fig. 20: The Result Showing the Family for the Query Search is Seen with the Significant and Insignificant hit

The ScanProsite tool is used for detecting functional and structural intra-domain residues helpful for function prediction of the protein. The PROSITE is a database consisting of a large collection of biologically meaningful signatures describing protein domains, families and function, also sites and associated patterns and profiles to identify them. It is used for short motif detection, or for detecting domains [21]. ScanProsite provides a web interface for identification of protein matches against signatures from the PROSITE database.It scans a sequence against PROSITE or a pattern against the UniProt Knowledgebase (Swiss-Prot and TrEMBL [21, 22].

The sequence to be scanned is pasted in the box at the homepage of scanprosite and the search is given to find any motifs present in the protein.

II. Result

The analysis of the protein sequence derived from Tuberculist is subjected to various alignment and analysis in the different bioinformatics tools mentioned above to find the Protein function, its closet Domain and Family. Thus we have done one dimensional annotation where the protein and its domain and family can be assigned. Further three dimensional analyses can

be done to find the Motif and Patterns in protein structures. These motifs and patterns can then be targeted for drug designing.

As seen in the process of protein annotation we come to know the various databases which are available freely. Open access to databases and various tools available on the internet can help in using these databases for various purposes like alignment, finding domain and family of protein or in our case protein annotation. By using the various available databases we can perform structural and functional analysis of genomic and proteomic data to find useful information (like annotation) from the huge amount data that is available in the databases.

ACKNOWLEDGMENT

The work was performed as a project for Conect2Decode under OSDD (Open Source Drug Discovery, a CSIR initiative. under the title- Glycomics of Mycobacterium tuberculosis: Annotating the glycan modifying enzyme from Mycobacterium tuberculosis.I would like to thank Dr Zakir Thomas Project director OSDD. Project Co-ordinator Dr Sulagna Banerjee, Molecular pathology lab, AU-KBC Research center MIT campus, Chennai for her guidance. Principal of Abeda Inamdar Senior college Dr E.M.Khan and Dr (Mrs)D.R.Majumder for their support.

REFERENCES

[1] *Raja Mazumder, Sona Vasudevan*, Structure-Guided Comparative Analysis of Proteins: Principles, Tools, and Applications for Predicting Function, September 2008, Volume 4 , Issue 9 ,e1000151

[2] *Khalid Raza* ,APPLICATION OF DATA MINING IN BIOINFORMATICS, Indian Journal of Computer Science and Engineering, Vol 1 No 2, 114–118

[3] Glycomics of *Mycobacterium tuberculosis:* Annotating the glycan modifying enzyme from *Mycobacterium tuberculosis*. SOP, OSDD: Summer research Program.Connect2Decode 2010.

[4] www.cazy.org/

[5] http://www.uniprot.org

[6] http://www.wikipedia//uniprot

[7] http://genolist.pasteur.fr/TubercuList/

[8] http://genolist.pasteur.fr/TubercuList/help/about.html

[9] http://www.ncbi.nlm.nih.gov

[10] http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html

[11] http://biochem.uthscsa.edu/~hs_lab/frames/molgen/tutor/psi.html

[12] http://www.ncbi.nlm.nih.gov/COG/

[13] *Roman L.Tatusov, et al*,The COG Database:a tool for genome-scale analysis of protein functions and evolution, Nucleic Acid Research 2000, Vol 28, No 1.

[14] http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

[15] http://www.cbs.dtu.dk/services/TMHMM/

[16] Krogh A, Larsson B, von Heijne G, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001 Jan 19; 305(3):567–80.

[17] http://prodom.prabi.fr/prodom/current/html/home.php.

[18] The ProDom database of protein domain families: more emphasis on 3D, Nucleic Acids Research, 2005, Vol. 33, Database issue doi:10.1093/nar/gki034

[19] http://www.pfam.sanger.ac.uk

[20] *Catherine Bru, Emmanuel Courcelle, Se´bastien Carre` re, Yoann Beausse, Sandrine Dalmar and Daniel Kahn*,The Pfam Protein families database, Alex Bateman, Ewan Birney, et al, Nucleic Acids Research, 2002, Vol. 30 No 1, 276–280.

[21] *Edouard de Castro1, Christian J. A. Sigrist1, Alexandre Gattiker, Virginie Bulliard1,Petra S. Langendijk-Genevaux,et al,* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins, Nucleic Acids Research, 2006, Vol. 34, doi:10.1093/nar/gkl124

[22] http://expasy.org/prosite

Image Mining to Identify Characteristics of Leaf using LAM

Shaikh Ashfak Ibrahim
MCA Department, Allana Institute of Management Sciences, Pune
e-mail: ashfaque.it@gmail.com

***Abstract*—This paper focuses on the identification of properties of betel leaf which can be used for the classification of leaves. LAM (Leaf attribute Miner) is a image processing tool for identification of various parameters of a leaf. Identification of the characteristics of leaf is possible by analyzing the internal arrangements of pixels. Classification of betel into classes 'healthy', 'with-black spots' and 'physically damaged' is possible on the basis of color attributes. Basic characteristics considered are height (in pixel), width(in pixel), height-width ratio, area, color composition (Red, green, blue) and infected area(in %) of leaf image for diagnostics.**

Keywords: Object identification, betel leaf, image mining, and characteristics of leaf, identification of disease.

I. INTRODUCTION

Image mining is extraction of data, knowledge or some other patterns stored in images, which are not explicitly visible. Typically an image mining process involves processing, transformations, feature extraction, interpretation and obtaining the final knowledge. The fundamental challenge in image mining is to determine the internal arrangement of the pixels in an image to find the properties of an image.

In agriculture image mining of leaf specially can help in discovering viral infections, identifying nutritional deficiencies in plant, and classifying them as diseased.

A tool named, Leaf Attribute miner (LAM) is developed using VB 6.0 as front end and MS access as back end, which takes a image as input and analyze them, as a part of this study. As part of experimentation 500 betel leaves were procured and images were obtained with black background using a digital camera. These images were converted to .jpeg format, before inputting to LAM to determine the attributes of leaves. Results were stored as a simple multidimensional database attributes as dimensions and diagnosed diseases or deficiency as the class value (fact) . Totally 300 leaves were used to train LAM and 200 leaves were use to validate classification output by LAM.

II. EDGE DETECTION

Edge detection is a technique to locate the edges of objects in the scene. This helps in locating the horizon, the corner of an object, and in determining the shape of an object. The algorithm is quite simple [1]:

- Go through the image matrix pixel by pixel
- For each pixel, analyze each of the 8 pixels surrounding it.
- Record the value of the darkest pixel, and the lightest pixel
- If (darkest_pixel_value –lightest_pixel_value) > threshold)
  then rewrite that pixel as 1;
  else rewrite that pixel as 0;

The challenge here lies in choosing a good threshold. Effects of varies choices of thresh hold on the extracted images are compared in Figs. 1-A to 1-C. You will need to run an image heuristics program to make the algorithm work properly.

III. LAM ENVIRONMENT

The input image must be in JPEG format with black background.

A. Obtaining the Leaf Pixel Count

Since the background of leaf is black, pixel with color value other than black will belong to the leaf. This is the first step tool LAM follows.

```
For i = 0 To nSize
' for each pixel across
For j = 0 To n1Size
'Get the main pixel color
PixCol=GetPixel(Picture1.hdc, i, j)
'Count pixels of leaf
If (pixCol<>VbBlack) then
totalPixel = totalPixel + 1
End If
Next j
Next i
```

Fig. 2: Calculation of Total Pixel Count

Where nSize = width of image in pixel
n1Size = height of image in pixels
totalPixel= Total number of pixels in leaf.

B. Find Threshold Value

In the Next step used by LAM is finding the threshold (fig 3)

thresholdValue= totalGreen/TotalPixels

This simple formula for threshold value has given satisfactory result as can be seen later. The final step is to detect edges.

```
For i = 0 To nSize
" for each pixel across
For j = 0 To n1Size
'Get the main pixel color
PixCol= GetPixel(Picture1.hdc, i, j)
'Convert to RGB main pixel
r = PixCol Mod 256
b = Int(PixCol / 65536)
g = (PixCol - (b * 65536) - r) / 256
gtotal = gtotal + g
Next j
Next i
```

Fig. 3: Computing Total Green Value of a Leaf

C. *Detecting the Edges in Leaf*

Edge detection means getting image as shown in Fig.1 from the given input (original) image.

The steps are

- Go through the image matrix pixel by pixel
- For each pixel, analyze each of the 8 pixels surrounding it.
- Record the value of the darkest pixel, and the lightest pixel
- If (darkest_pixel_value - lightest_pixel_value) >threshold)
  then rewrite that pixel as 1; else rewrite that pixel as 0
- The source code for the same is.

```
For i = 0 To nSize
' for each pixel across
For j = 0 To n1Size
'Get the main pixel color
pixCol0=GetPixel(Picture1.hdc, i, j)
'Convert to RGB main pixel
r0 = pixCol0 Mod 256
b0 = Int(pixCol0 / 65536)
g0 = (pixCol0 - (b0 * 65536) - r0) / 256
'Get the pixel1 color
pixCol1 = GetPixel(Picture1.hdc, i - 1, j - 1)
'Convert to RGB  pixel1
r1 = pixCol1 Mod 256
b1 = Int(pixCol1 / 65536)
g1 = (pixCol1 - (b1 * 65536) - r1) / 256
'Get the pixel2 color
pixCol2 = GetPixel(Picture1.hdc, i - 1, j)
'Convert to RGB  pixel2
r2 = pixCol2 Mod 256
b2 = Int(pixCol2 / 65536)
g2 = (pixCol2 - (b2 * 65536) - r2) / 256
.
.
.
.
'Get the pixel8 color
pixCol8 = GetPixel(Picture1.hdc, i + 1, j + 1)
'Convert to RGB  pixel8
r8 = pixCol8 Mod 256
b8 = Int(pixCol8 / 65536)
g8 = (pixCol8 - (b8 * 65536) - r8) / 256
maxGreen = g1
minGreen = g1
If maxGreen <= g2
Then maxGreen = g2
If maxGreen <= g3
Then maxGreen = g3
.
.
.
.
.
.
If minGreen >= g8
Then minGreen = g8
If ((maxGreen - minGreen)> gthreshold) Then
SetPixelV Picture2.hdc, i, j, vbBlack
Else
SetPixelV Picture2.hdc, i, j, vbGreen
End If
Next j
Next i
```

Fig. 4: Code snippet to Detecting Edges in Image

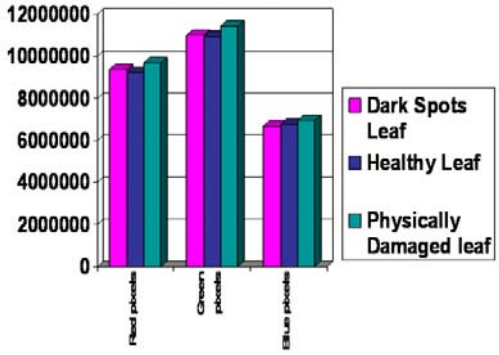


Fig. 1: Average Red-Green-Blue (in pixels) of Betel Leaves (Dark Spots, Healthy and Physically Damage

TABLE 1: TYPICAL MDB RESULTING FROM THE TEST RUNS

Database Structure

Image ID	Leaf Type	Height (in pixel)	Width (in pixel)	H:W Ratio	RED	GREEN	BLUE	Area (in pixel)	Infected Pixel	Infected area (%)
4	Dark Spots	494	259	1.90 : 1	8712648	10087580	6044247	73103	4961	6.78
5	Dark Spots	513	202	2.53 : 1	6305185	7499440	4126855	58080	3955	6.80
9	Dark Spots	545	208	2.62: 1	7669149	8663634	4606095	64710	12117	18.72
166	Healthy	553	252	2.19 : 1	10011394	11853368	7622710	80917	1056	1.30
223	Healthy	547	224	2.44 : 1	7931201	9535236	5880592	68218	1114	1.63
225	Healthy	490	235	2.08 : 1	8289906	9827388	6269522	66892	667	0.99
2	Physically Damaged	541	195	2.77 : 1	5742720	6889674	3815509	54693	2208	4.03
12	Physically Damaged	541	308	1.75 : 1	11845090	13550422	7790917	95987	8315	8.66
13	Physically Damaged	389	237	1.64 : 1	6606668	8029842	4488977	61662	725	1.17
14	Physically Damaged	508	309	1.64 : 1	12124509	13877751	7875576	98745	8431	8.53

TABLE 2: ELEMENTARY STATISTICS PARAMETERS USING DATA ON TEST RUNS (200 LEAVES)

Leaf Type	Avg Height (in pixel)	Avg. Width (in Pixels)	Avg. Red (in pixels)	Avg. Green (in pixels)	Avg. Blue (in pixels)	Avg. Area (in pixels)	Avg. Infected Pixel	Avg. Infected area
Dark Spots	539.08	253.27	9382768.78	10989591	6686784.3	78864.03	4652.8	5.92
Healthy	533.86	246.57	9214963.64	10963023	6778844.2	75708.96	1272.2	1.67
Physically Damaged	518.89	260.37	9686600.10	11422118	6958777	81150.05	3200.4	3.78

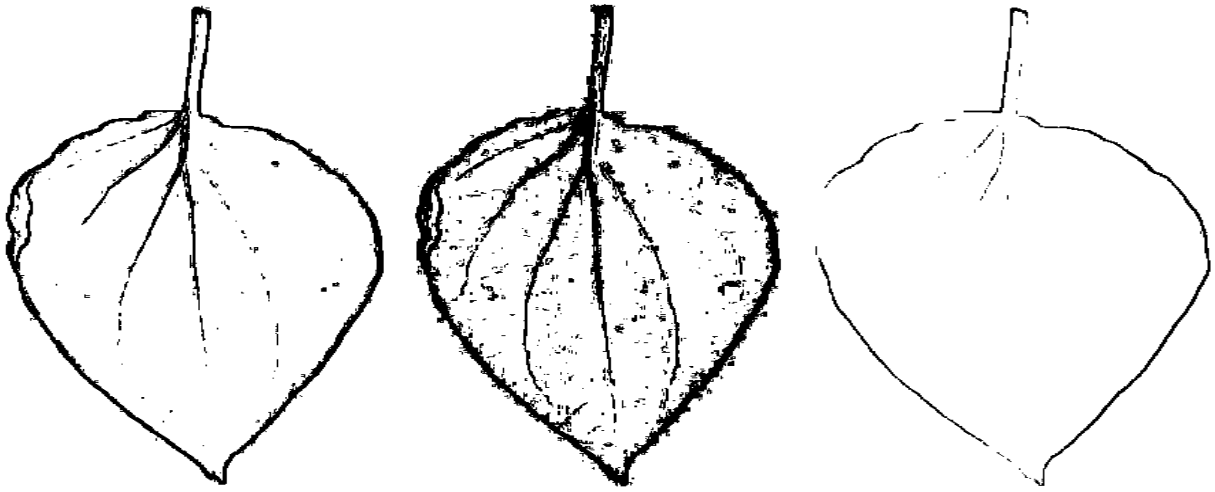


Fig. 1-A: Edges in original Image (betel leaf)

Fig. 1-B: Edge Detection (high threshold value)

Fig. 1-C: Edge detection (low threshold value)

Fig. 1 : Edge Detected under Different Threshold

IV. RESULTS OF THE TRAINING AND TEST RUNS

The LAM tool can be useful in agricultural field in identifying the characteristics of a leaf which can help us to identify the diseases and state of a leaf. DTREG tool can be used for classification of leaf.

REFERENCES

[1] http://www.societyofrobots.com/programming_computer_vision_tutorial_pt3.shtml#edge_detection

[2] Chia-Ling Lee and Shu-Yuan Chen, *"Classification for Leaf Images"*, 16th IPPR Conference on Computer Vision, Graphics and Image Processing, 2003.

[3] Milan Sonka, V. Halvac and R. Boyle, *"Image Processing, Analysis, and Machine Vision"*, Thomson Brooks/Cole, 2004.

[4] Rafael C. Gonzalez, Richard E. Woods, *"Digital Image Processing"*, Pearson, 2003.

[5] Ze-Nian Li, Mark S. Drew, *"Fundamentals of Multimedia"*, Pearson, 2006.

[6] Evangelos P., *"Mastering Visual Basic 6.0"*, bpb, 2002.

[7] Steven Holzner, *"Visual Basic 6 Programming Black Book"*, Dreamtecch Press, 2007.

Usage of Social Networking amongst Health-Care Professional for Dissemination of Medical Knowledge and Community Service

Dr. Manik S. Kadam[1*] and Prof. Murtaza. M. Junaid Forooque[2#]
[1]*JSPM Institute of Management and Reserch, Pune, India*
[2]*Allana Institute of Management Sciences, Pune, India*
*e-mail: *msk1612@hotmail.com, #binhasham@gmail.com*

***Abstract*—Social networks are beginning to be adopted by healthcare professionals as a means to manage institutional knowledge, disseminate peer to peer knowledge and to highlight individual physicians and institutions. There are social networks created to help its members with various physical and mental ailments. For people suffering from life altering diseases, and people who are alcoholic, addict, obese and suffering from disabilities.**

The present Paper studies the level of usage of online social networking sites by Medical professional and its potential benefit to them.

The study is based on a survey conducted involving practicing doctors of various medical traits. The study indicates that Physicians rarely use social networking site. Most of them doubt the reliability and accuracy of information for medical practice. However they agree it can be used for medical education, research and advertisement.

Keywords: Social network, physicians, Medical Traits, Medical Education, Practice, Research

I. INTRODUCTION

"A social networking site can be defined as web-based services that allow individuals to

1. Construct a public or semi-public profile within a bounded system,
2. Articulate a list of other users with whom they share a connection, and
3. View and traverse their list of connections and those made by others within the system". [1]

Christian Fuchs defines Integrated social networking sites (ISNS) as a web-based platforms that integrate different media, information and communication technologies, that allow at least the generation of profiles that display information that describes the users, the display of connections (connection list), the establishment of connections between users that are displayed on their connection lists, and the communication between users"[2]

A social network service is an online service, platform, or site that focuses on building and reflecting of social networks or social relations among people, e.g., who share interests and/or activities. A social network service essentially consists of a representation of each user (often a profile), his/her social links, and a variety of additional services. Most social network services are web based and provide means for users to interact over the internet, such as e-mail and instant messaging. Although online community services are sometimes considered as a social network service in a broader sense, social network service usually means an individual-centered service [10]

II. SOCIAL NETWORK USAGE BY MEDICAL PROFESSIONALS

Physicians are using world wide web to build their reputation and as a tool to advertise their own practices and research. They are participating in medical wikis, becoming member of medical communities. And maintaining their own medical blogs. The medical blogs contains more contents and opinions then a medical research paper.[3]

A. But there Can be an Issue of Reliability and Accuracy of Information

Web provides significant potentials like structural and "Just in time "learning enabling students to collaborate. This may be challenging for both academia and students, and will require a different perception of roles. Emergence of new knowledge generation and distribution media like wikipedia brings issues of assessment of the reliability or accuracy of resources [5] and may challenge the role of academic journals, by providing open access and easy availability. There can be other challenges like detection of plagiarism etc.

The potential offered by Web 2.0 technologies in the education of healthcare professionals, is potentially significant, however these developments need to be balanced with the inherent risks and challenges. [4]

Social computing platform have considerable potential to be used for research on rare diseases because of sustainability and profitability for pharmaceutical industry and the society at large. As the body of knowledge on Rare Dieses has developed very slowly and is still largely an "'uncharted territory.

Research on specific rare diseases through the application of social computing is worth - socially, clinically and economically.[5]

Clinicians collect a large amount of data from clinical trials like the disease history, symptoms, treatment details, and reactions of all participants in the trial etc. similar type of data is also available at data-based health social networks like PatientsLikeMe, MedHelp, and CureTogether etc. This data can be useful for Pharma or biotech companies.

B. Patients Like Me Makes it Clear That They

Sell the information in an anonymous, aggregated and individual format to the companies that can use it to improve or understand products or the disease market." [6]

Medico-legal expert, *Steven I Kern*, suggest s that, embracing social network may be challenging for medical a professional. . Because patient reaction to social networking efforts may vary dramatically from patient to patient and also may depend on the specialty of the physician sending the message, careful thought must be given. He gives following suggestions

- Consider your patient population and your specialty before you decide to enter the **social networking** market.
- Ensure that the content of your communication is appropriate.
- Maintain patient confidentiality
- Do not merge your personal social site and your professional site.[7]

Medical and health-related examples of social networking services include the LibraryThing Medicine Group (http://www.librarything.com/groups/medicine), a library social network site promoting social interactions, book recommendations, self-classification, and monitoring of new books, and the MySpace 'CURE DiABETES group' (http://groups.myspace.com/cureDiABETES) run by patients and supporters. Some social networking services combine or bundle several Web 2.0 tools/features together, e.g. instant messaging, social bookmarking, blogs and podcasts. Examples of these services include the Mental Health Social Network (http://social.realmentalhealth.com/), and the IJS portal (http://www.theijs.com/), a global community portal centred around the *International Journal of Surgery*.

The web has potential benefit for community oriented services like health care, which can extend extending beyond the geographical reach of organizations. The knowledge is shared between clinicians, and with their patients and public at large. Discussion groups and communities of practice can be created on social networking sites, relationship can be established. Establishing connections with other relevant players is an important factor in patient support, especially in chronic conditions [8]

Almost all of the top pharmaceutical companies, biotechnology firms and medical device manufacturers have some social network presence; some are partnering with third-party social networks such as PatientsLikeMe and Sermo to communicate and collaborate with external stakeholders.

Table No 1 compiled from various sources provides examples of potential applications for key stakeholder groups. As the regulatory environment becomes more defined and innovative organizations demonstrate measurable commercial value from social networks, increasing numbers of health care stakeholders are expected to recognize – and leverage – this transformational technology's role in information acquisition and access. Indeed, social networks are a trend that could change the face of health care information sharing, providing new power for providers and patients alike.

TABLE 1: FORTUNE 100 COMPANY ENGAGEMENT IN SOCIAL NETWORKS

Fortune 100 companies	**Blogs**	**Facebook**	**Twitter**
Telecommunications	75%	100%	100%
Computers, Office Equipment	67%	100%	67%
General Merchandiser	50%	75%	100%
Motor Vehicles and Parts	67%	67%	67%
Specialty Retailers	50%	50%	100%
Food and Drug Stores	17%	33%	50%
Insurance: Property and Casualty	0%	25%	50%
Aerospace and Defense	33%	17%	50%
Commercial Banking	25%	0%	38%
Health care: Insurance and Managed Care	0%	0%	50%
Pharmaceuticals	33%	0%	33%
Petroleum Refining	11%	11%	22%

III. OBJECTIVE OF CURRENT STUDY

This study was conducted to determine

- The extent of usage of Social networking sites by Indian physicians
- To find out if there is an association between physician 's demography and specialization with usage pattern
- Awareness about the Capabilities of these sites

IV. HYPOTHESIS

The following hypothesis were framed

H1 Indian Medical Professional are reluctant to use online social networking sites

H2. Indian medical profession agree that online social networking sites can be used for medical education, medical research, advertising their practice.

V. RESEARCH METHODOLOGY

A Questionnaire was designed and responses were collected from several medical professional in Pune, and similar questionnaire was mailed to several medical professions in India using Google forms.

A. *Finding*

The following were finding obtained from the survey

TABLE 2: SOCIAL NETWORK USAGE PATTERN AMONG HEALTH CARE PROFESSIONAL

Internet usage in hrs/week	Percentage Of users	Males	Fe-males	Allo-path	Homeo-path	Ayur-vedic	Age < 30	Age >30
Less than 5	55	30	25	20	5	15	25	30
5-7	10	40	-	5	-	-		5
7-10	20	15	5	5	-	-		
More than 10	15	15	-	15	-	-	15	

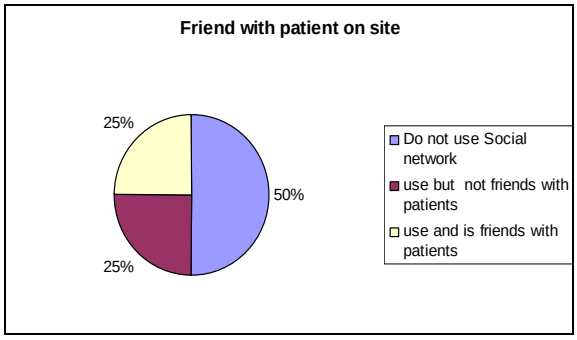


Fig. 1

B. *Participation in Online Social Media*

Out of surveyed professionals 50% said that use online social networking sites, 35 % use medical communities, 15% use medical blogs , 40% use wikis, 25 % 5 % use Discssion forums.

The table no 3 shows the social sites popular with the medical professional, it includes google, facebook,orkut medisoft, medhelp etc

TABLE 3: USAGE OF SOCIAL SITES AMONG MEDICAL PROFESSIONALS

Name of Social Media	Usage in Percentage
Facebook	40
Google	20
Orkut	20
Yahoo	15
Gmail	10
Rediff	10
Medisoft	5
MediHelp	5
Tweeter	5
None of above	35

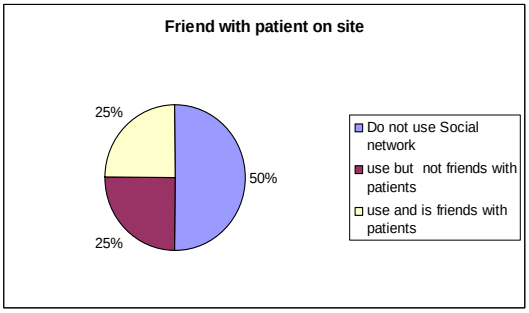


Fig. 1: Percentage of Medical Professional Friends with Patients on Social Network

Awareness about medical specific social sites was also found to be less amongst the surveyed physicians, 30% are aware about medhelp, awareness about pubmed, medline was also found to be 5% each. About 5% respondents said they use health specific group on social networking site like face book.

Only half of the physician surveyed use social networking sites and one fourth are friend with their patients (refer figure no 1)

C. *Potential Benefits Usage of Social Networking Sites to Medical Community*

Opinions were sough about potential usage of social networking for benefits of medical community like advertising their practice, medical research, medical education etc. the summary of the response is given Table no 5

TABLE 5: OPINION ABOUT POTENTIAL USAGE OF SOCIAL NETWORKING SITES BY MEDICAL COMMUNITY

Potential use	Yes	No	Not sure
Can be used for medical practice	60%	25%	15%
Can be used for medical education	80%	15%	5%
Can be used for research on rare diseases	70%	10%	20%
Can be use for advertising practice and research	55%	10%	35%

TABLE 6: OPINION REGARDING POTENTIAL BENEFIT OF SOCIAL NETWORK TO THE MEDICAL COMMUNITY

Effects	Percentages Agrees
Good Patient Relationship management (PRM)	55
Effective communication	70
Better chance of Recovery	30
Popularity and Goodwill	30

VI. CONCLUSIONS

- Indian Medical Professional are reluctant to usage online social media because of they are not familiar with the technology
- Some Physician feels online Prescription may give patients false sense of security
- Some say to should used for connecting with friends and advertisement only
- Most of the respondent agree that it can be used for medical education

REFERENCES

[1] Boyd, Danah and Nicole B. Ellison (2007) "Social Networking Sites: Definition, History, and Scholarship", *Journal of Computer-Mediated Communication 13*.

[2] Christian Fuchs (2009) "Social Networking Sites and the Surveillance Society", *A Critical Case Study of the Usage of studiVZ, Facebook, and MySpace by Students in Salzburg in the Context of Electronic Surveillance.*

[3] Bertalan Mesko, University of Debrecen, Debrecen, *Hungary Medical Education and Building an Online Reputation in the World of Web 2.0.*

[4] Rod Ward, University of the West of England, Bristol, *UK The Potential and Challenges of Web 2.0 in the Education of Healthcare Professionals.*

[5] Marcelino Cabrera Giraldez, Institute for Prospective Technological Studies (*Joint Research Centre, European Commission*); José Antonio Valverde , IPTS; Dolores Ibarreta, IPTS.

[6] Socialized Medicine, *How Personal Health Records and Social Networks Are Changing Healthcare Darin Stewart. EContent. Wilton,* Sep 2009, Vol. 32, Iss. 7, p. 30.

[7] Social Networking Poses New Challenges Steven I kern, *Medical Economics*, Oradell, Jan 8, 2010, Vol. 87, Iss. 1, p. 29.

[8] "The Emerging Web 2.0 Social Software", *An Enabling Suite of Sociable Technologies in Health and Health care Education*, Maged N. Kamel Boulos1, Steve Wheeler2 Article first Published Online: 28 FEB 2007 Health Information & Libraries Journal Volume 24, Issue 1, pp. 2–23, March 2007.

[9] "Medicine 2.0 Proceeding of Conference on Medical Internet", Research Toronto Canada September 2008 Published by *Journal of Medical internet Research*.

[10] www.wikipedia.com/ wiki/Social_networking_site

Data Mining usage in Health Informatics: A Case Study

Prof. Sheetal Uplenchwar and Prof. Rajesh More
Allana Institute of Management Sciences, Pune
e-mail: sheetaluplenchwar@gmail.com, more.rajeshmore@gmail.com

***Abstract*—Medical decision process involves classification and diagnosing of diseases. Decisions must be made effectively and reliably. This study will aid the health informatists in decision making process. Health informatics is also called health care informatics, healthcare informatics, medical informatics or biomedical informatics. It is a discipline at the intersection of information science, computer science, and health care.**

Decision trees help reliable and effective decision making with high classification accuracy and a simple representation of gathered knowledge. In this study we have focused on selected diseases with their major symptoms as data support. DTREG tool is used for generating Decision Tree.

Keywords: Data Mining, DTREG, Node, Decision Tree, Target Variable, Predictor Variable, Disease, Symptom

I. INTRODUCTION

The aim of data mining is to find the answer to a strategic query from the data extracted. A decision is usually constructed as a combination of experiences and practices in the specified area. The goal of decision tree is to to provide optimum or shortest path for decision. This paper deals with decision tree, specially constructed for quick identification of possible disease (target) based on symptoms (predictors).

A. DTREG–An Open Source Tool

DTREG is a software tool which accepts a dataset containing number of attributes (columns / dimensions) and a associated classification . One of the variables is the "target variable" whose value is to be modelled and predicted as a function of the "predictor variables". DTREG analyses the data and generates a tree structured model showing how best to predict the values of the target variable based on values of the predictor variables. DTREG can generate two types of trees depending on whether the target variable is continuous (the relative magnitude of the values is significant) or categorical (The actual magnitude of the value is not significant).

B. Decision tree Generation

To generate decision tree using DTREG tool, we have to provide a csv file as an input to DTREG. In most cases, creating data table in a spread sheet and then converting it to a csv file could be the approach. Name of the "Disease" is considered as a "Target Variable" which is predicted with the help of a set of "symptoms" which are the attributes. In our data "twenty symptoms" have been selected to form the set (Table 1), and are treated as a "Predictor Variables".

TABLE 1: DISEASE AND THEIR SYMPTOMS

Sr.	Disease	Symptoms
1	Arthritis	Swelling, joint inflammation Stiffness, tenderness, weight loss, fever, weakness
2	Backache	joint inflammation, stiffness tenderness, weight loss, fever weakness
3	Hypotension	Pain, weakness, malfunctioning in brain, shortness of breath, fast breathing, cold
4	Chikungunya	Swelling, joint inflammation weight loss, stiffness, tenderness, fever, weakness, faint, bodyache, headache, maculopapular rash
5	Chickenpox	Fever, weakness, headache, maculopapular rash, red bumps
6	Diarrhea	Pain, fever, Omiting
7	Malaria	Pain, fever, cold, Omitting, headache
8	Jaundice	Pain, fever, headache, Coloration of skin and eyes
9	Anemia	weakness, faint, shortness of breath, headache, Fatigue or tiredness

Since relative magnitude of values is not significant, we generate decision tree using classification approach. To predict the value of the target variable using a classification tree approach, values of the predictor variables move through the tree until it reach a terminal node or leaf node.

C. Data Collection

Text mining on internet was used to collect data as given in Table 1. The information was converted into a csv file suitable for use on DTREG.

D. Decision Tree for Symptom Based Diagnostics

The decision tree generated is given in Fig 1.

To navigate through this Decision Tree, we have to start from root node (node 1). In the given Decision Tree N stands for total number of Target variable. Decision Tree contains 7 levels. At Level 1 we won't be able to predict Target variable (Disease) hence the

percentage of misclassification is high i.e. 88.89%. At Level 2 it checks for presence of swelling if it is present then it visit node 2 otherwise node 3.

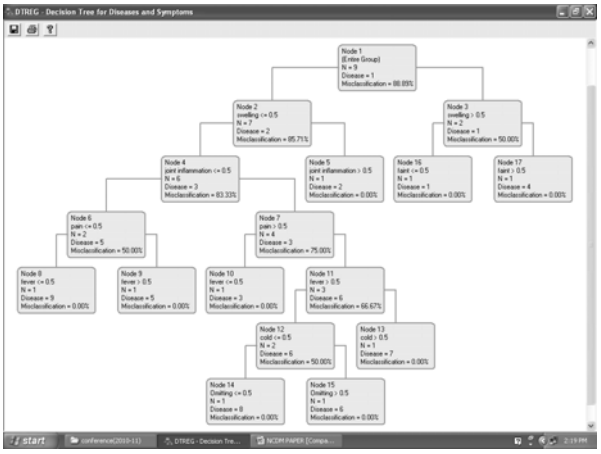


Fig. 1: The Decision Tree

This process is continued until terminal or predictor is obtained, at this stage percentage of leaf node in zero.

E. Importance of Variables

Importance of variable is shown in Fig 2. The variable which include in decision making process has greater importance. In our study variable fever has the most importance.

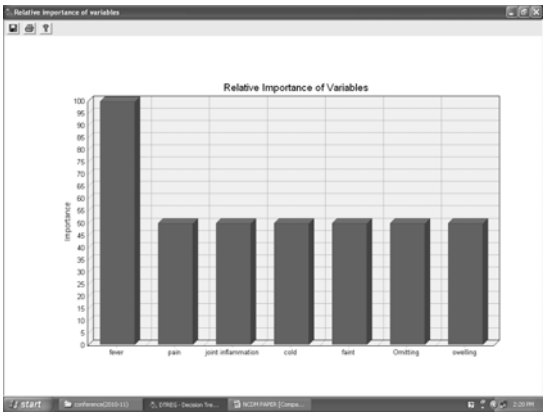


Fig. 2: Importance of Predictor Variables

II. Conclusion

Decesion trees can effectively used to help diagnosi process in medical informatics. DTREG is a powerful open source data mining tool that can generate classification and regression decision trees that model the basic data. DTREG is robust and can be installed easily on any Windows system. Even complex analyses can be set up in minutes. A decision making procedure based on constructing logical arguments for and against a number of available choices, assessing their relative merits and weighing their relative strengths is possible to design. We build decision trees in order to capture underlying relationships in a dataset which help us in classification and prediction as well as in data visualization.

References

[1] Michael Berry & Gordon Linoff, ,(2000.) Mastering Data Mining, John Wiley & Sons,
[2] K. Cios, W. Pedrycz, R. Swiniarski, L. Kurgan, Data Mining,(2007) A Knowledge Discovery Approach, Springer, New York
[3] Margaret Dunham, (2003) Data Mining Introductory and Advanced Topics, Prentice Hall India.
[4] Parr Rud O. (2001) Data mining cookbook: modelling data for marketing, risk, and customer relationship management. USA: Wiley Inc
[5] http://www.autonlab.org/tutorials/
[6] http://www.ncbi.nlm.nih.gov/
[7] http://www.medicinenet.com
[8] http://www.phi.umd.edu/what/

Identification of Mizaj (Temprament) Based on Tibbi Fundamentals using Classification as Tool

Prof. Murtaza M. Junaid Farooque[1#], Dr. Sayyed Abidurrahman[2*] and Prof. Farhana Sarkhawas[2#]
[1]*MCE Society's Allana Institute of Management Sciences, Pune, India*
[2]*MERC Z.V.M Unani Medical College, Pune, India*
e-mail: [#]binhasham@gmail.com, [*]drabidforyou@gmail.com, [#]farhana.ap@gmail.com

***Abstract*—Unani Medicine or Tibb system is based on classification of humans personalities into four different mizaj based on dominance body fluids. In healing drugs as well as diseases are also classified according to the four humors. The four humors correspond to four bodily fluids ie blood(Dam), phlegm(Balgham), black bile(sauda) and yellow bile.(safra) A typical diagnosis of a patient would take the balance of these humors into consideration, before giving any treatment.**

The data of 68 subjects on various parameters was collected and was used to generate an optimum decision tree using Weka using for identifying the mizaj of the subject. The Algorithm used was J48 Algorithm.

Keywords: Unani Medicine, Tibb, mizaj, body fluid, homors, Decision Tree, Machine Learning, Weka, J48

I. INTRODUCTION

Unani Medicine or Tibb is one of the oldest forms of medical therapy practices in India, and several other countries. This system classifies humans personalities into four different mizaj (Temperaments) based on the dominance of body fluids. In healing drugs and diseases are also classified according to the four humors. The four humors correspond to four bodily fluids i.e. blood, phlegm, black bile and yellow bile. A typical diagnosis of a patient would take the balance of these humors into consideration, before giving any treatment.

Hippocrates (460 B.C.) in his book "Human Nature" set forth the doctrine of body fluids i.e. humours or Akhlat, (singular khilt), that human body contains four major kind of humour i.e.

- Dam (blood)
- Balgam (Phlegm)
- Safra (yellow bile)
- Souda (black bile)

A right proportion, according to quality and quantity, and mixing of which (homeostasis) constitutes health and un-right proportion and irregular distribution, according to their quantity and quality constitute disease."

In unani medicine tht law of treatment is based on the temperaments, as long as homeostasis of internal environment is maintained the body remains healthy when this homeostasis is disturb diseases is developed. Thus the humoral theory deals with all aspects of diseases i.e. etiology, pathology, prevention and treatment.[1]

If the disease is of particular temperament, the drug prescribed is usually of the opposite temperament to neutralize the effect of the disease.

According domination of four kind of akhlat the human species can be broadly classified into four types of personalities.

1. Damawi (sanguine or plethoric type)
2. Safrawi (chloretic or bilious type)
3. Balgami (phlegmatic or pituitic type)
4. Saudawi (Melancholic type)

Domination of certain khilt exerts its influence on the mizaj (temperament) of a person. However No Khilt (fluid) is present in isolation they are always present in combination.

II. IDENTIFICATION OF THE TEMPRAMENT

Normal (mizaj)l or abnormal (su'al-mizaj) is not determined by chemical analysis of different akhlat of the body however the Tibbi Physicians have devised mean to find out mizaj. The sign and symptoms by which mizaj can be diagnosed are classified into following 10 divisions

1. Malmus (Tactile sensation)
2. Lahm wa shahm (Muscle and Fat)
3. Ash'ar (hairs)
4. Laun (body color)
5. Hay'at al-aza (Stature)
6. kayfiyat al-infial (quality of passiveness of organ)
7. Naum wa yaqzah (sleep and wakefulness)
8. Af'al-aza (bodily functions)
9. Fadhlat al badan(excreta of the body)
10. infilat nafsaniyah (psychic reactions)

Hakim sayed Ishtiyaq Ahmed in book "Principles of Human physiology in Tibb (An Introduction to Al-Umur Al-Tabiyah) has given a chart which can be useful in identification of Mizaj, this are 55 different parameter belonging to above 10 classes. Number and Nature of variables is reported in Table no 1.

Based above facts the developing A Rule Based Expert System for identification of Mizaj (classification of persons according to their mizaj) with such permutations and combination will be very complex. Moreover the limitation of the resources like time, processor capabilities, memory, etc has to taken be into consideration. Simple program based on if-then else or case switch, may result in a very slow response time. Hence we have to go for data mining and machine learning approach.

Identification of the Mizaj of the subject is the problem of classification in data –mining. Classification infers the defining characteristics of a certain group (in our case Mizaj). These methods involve mapping a set of data into these predefined groups.

Classification routines in data mining also use a variety of algorithms and the particular algorithm used can affect the way records are classified. The common algorithms for classification can be Decision tree, Nearest Neighbor, Naïve Bayes classifiers, Artificial Neural network etc [2]

III. Methodology

We have developed a GUI using java Swings which takes the values of the variables as input and stores the in a spread sheet, the input constitutes of information

- As provided by the observer (doctor) by his observation of the subjects
- Revealed by subject based on his perception about himself.

The numbers of parameter were reduced from 55 to 40, with consultation of expert. The descriptive Parameters like anger fear etc were quantified as high, medium and low. The data about 68 Patients was collected and stored. The collected data was given to practicing unani expert to identify the Mizaj (temperament) of the subject.

This data collected was converted into CSV format and was used to generate decision tree in WEKA using J48 algorithm. The tree can be used in the next phase to design automated software tool for identifying mizaj.

IV. Results

At first stage the tree generated showed a tree with large number of leaves, most of the leaves about 6 or more under two parameter i.e. occupation of the subject and nature of the dreams. This confirms the belief that occupation of the subject is an dominating parameter. According to unani principles it also plays an important role in identification of the mizaj. Exposure to coldness, hotness, dryness or moisture at workplace may have influence on the mizaj of the subject. Similarly the parameter like 'nature of dreams' was also taken into consideration. People gave answers in different terms for eg, No, No dreams, no, none, etc. Hence to generate optimum tree we applied data reduction. The occupation of the subject was reduced to three categories unemployed (students, housewife, retired etc), self employed (businessmen, doctors, professional practice etc) and employed (government or private service etc). Similarly no or none or No dreams means same, happy or soothing dreams have same meaning. After data reduction data set with modified values were input to WEKA in order to generate a decision tree. The tree now generated was of size 17 (nodes) and 10 leaves.

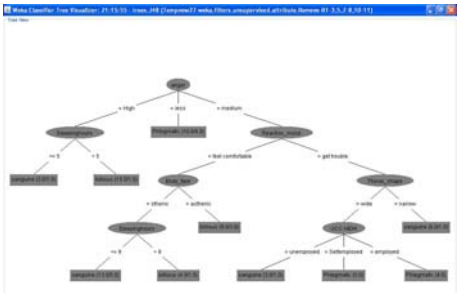


Fig. 2: Decision Tree for Identifying Mizaj as Generated by WEKA

V. Conclusion

It was found, on interpretation of the tree, that, out of 40 parameters, only 6 parameter are dominant in identification of mizaj namely anger, sleeping hours, reaction with moisture, body type, shape of thorax and occupation. The decision tree was able to identify only three types of mizaj, the forth being melancholic type could not be identified as there were only two subject of melancholic type in the data-set. There are chances of bias in the result as the data was provided by the subjects based on their own perception. Physical examination by the physician was not conducted.

TABLE 1: NUMBER AND NATURE OF VARIABLES

No of Possible Values / Data Collection method	2	3	4	6	**Numeric Range**	**Descriptive**	**Total**
By Observation	6	6	3	1	0	6	**22**
By using special Instrument	0	0	0	0	6	0	**6**
By Calculation	0	0	0	0	3	3	**3**
Can be revealed by subject only	2	1	0	0	3	18	**24**
Total	**8**	**7**	**3**	**0**	**12**	**24**	**55**

References

[1] Hakim sayed Istiaq Ahmed,(1980) Introduction to Al-umur Al Tabiyah (Principles of human physiology in Tibb)

[2] Ian H. Witten, Eibe Frank, Data Mining Practical Machine learning Tools & Techniques

[3] Warren P. ( ) the Roots of Scientific Medicine by - The Humoral Theory of Diseases.

[4] Ibne Naifis (1954), Kuliyat-e-Nafisi Idara kitab us shifa

[5] Prof. Jamil Ahmed, Hakim Ashraf Qadeer (1998)Unani The science of greeco-Arabic medicine, Lustre Press Pvt. Ltd.

Data safety and Confidentiality Consideration in Medical Research: An Ethical Approach

Prof. S.D. Bagade[1] and Prof. Mehdi Ali Jafri[2]
MBA Department[1], MCA Department[2] AIMS, Azam Campus, Camp, Pune-411001, India
e-mail: tns321@rediffmail.com, majafri123@rediffmail.com

***Abstract*—In the Medical Research, testing of drug, analysis and its effects is fundamental task. When data is collected for analysis, it is stored in a locked secure database; unless data collection process is not completed it can't be accessed. Ethically the participating subject, to whom drug is given, should know the purpose with right to withdraw participation and the subject who collects the sample should not use the data. The sample collector may be able to recognize what could be expected result after few trials. The participating subjects are human being and hence their disturbing facts should not be affected. This paper points out and suggests making such arrangement of several sample collectors to avoid prediction of results before data mining and analysis with rational satisfaction.**

Keywords: Data warehouse, Data Mining, Database Lock, Ethics, DBA, Clinical Data, Protocol, Access Control.

Before a medicine/drug is introduced into the market; it undergoes several tests and procedures. One of the processes is to find the subjects for test. When the drug is tested on the many subjects, the evaluated data is collected continuously for the overall micro & macro analysis of its effects. For Medical Research; data collected from subject; includes personal. Data is collected in the database and there may be lock on the database before the site closes-out.

Before data is stored in the database and locked, the sample collector should be observed and tested for understanding whether the parameters related to the ethics are scrupulously followed or otherwise.

Introduction: Data Mining is the methodology used to analyze data so as to find trends and patterns for effective result oriented work. Data mining becomes imperative & indispensable in medical research because of continuous collection of data of a subject (Patient) which helps to understand meticulously and error free data/information and critically evaluates the effect of a medicine administered.

This is possible only when data warehouse for enterprise is strengthened.

A. Ethical Principles as Given

Before the subjects will make a decision to provide personal data for analysis; they should know: how it will be used, who will use data, what will be used and what steps will be taken to maintain confidentiality. Appropriate rights should be given / delegated to the participating subjects to continue or to voluntarily withdraw participation.

B. Data Stored and Retrieval Process

Challenges in data security: Before data is subjected to be analyzed, samples are collected at different sites of heterogeneous patients; at distinct locations who are treated by different medical practitioner and kept in the locked database. Sample collector at respective site is primary person to collect the data. Ethically sample collector should not use the data but before data collection-completion, sample collector may understand the effects on his site and may infer the overall analytical result. To avoid such kind of possibility; sample collector should be considered & evaluated from the view point of reliability, integrity & legally.

C. Likely Problems of Ethics

Both the patient and authorized medical practitioners are human beings. Their perception and perseverance are different. Further pattern of behavior; subject to situational factor will be different and hence motive also change.

Value oriented approach and holistic approach are inter related and inter connected "I am to grow along with others and not at the cost of others".__ this is the foundation of all values. We have to protect, nourish, work, share, & study together, grow together. One is to grow for one's inner growth and for the good of the world. It is the core or focus.

The elements of management system in quality environment are focus, methodology (i.e. method & procedure), processes, policies and practices. All these are crucial elements.

The learners after critical analysis feel that invention of life saving or disease curing drug is welcome. However two questions; it is felt are unreplied. One; along with allopathic drug therapy; whether treatment from paramedical streams [e.g. Ayurveda, Unani, Naturopathy etc] can be administered simultaneously. Secondly; for every organization for sustainability, growth, development: invention is a must. When a new invention takes place; endeavor is in the direction of getting patent registration and license. Undoubtedly this entire process is economically extremely vibrant and costing of the product is having an impact & influence on determining the price of a product. Whether " aam aadmi" is at centre for this purpose of deciding reasonable price is a big question.

Administration of the drug is for a noticeable period. Excitement of doctor to selflessly treat the patient and patient's patience to response to prescription/proscription needs crystal clear definition and how it is being monitored needs to be focused.

The data generation and information gathered from the patient's diagnosis, prognosis, response; situation at every phase of the cycle be transmitted, retrieved, examined continuously: by the research centre. The learners suggest that once a data is generated, retrieved and transmitted to research head quarter; expeditious action to lock the same is necessary. This is important from the view point of ensuring preservation of data/information gathered from process – stage results. This will arrest the practices of amending, modifying, adding, deleting the requisite, relevant information. The result yielded will be ensuring error free & true conclusion.

One of the stipulations is that: medical practitioner may publish after taking consent of the manufacturer. Drugs are medicated for expeditious recovery and quickest possible cure. Hence premature opinions, subjective judgments should be avoided. Any expression is to be done cautiously and substantiated with right factual data/information. The communication should be in such a way that the transmitters from–meaning–feeling be in resonance with the receiver of message.

The learners are inclined to state that: products are developed to meet customer needs [quickest healing with no side effects]. Brands are positioned to offer distinctive value [But the market price levied should be reasonable and affordable to "aam aadmi"] communications are used to create expectations of value [There must be transparency and accountability]. Delivery is used to reinforce the value proposition relationship which is built to offer lifetime customer value.

The learners further feel that market research [participation of medical practitioner & patient] is used to understand customer needs. Hence the respondent base should be widened. The company; manufacturer when the customer wants to buy. Hence ethical practices to market newly invented drugs needs to be persued. Distribution enhances shopping experience. Hence accessibility, affordability, proximity and timely delivery has got its own importance. The truth is; we are all interconnected, inter-related, inter dependent. Hence prevent all types of pollutions.

I. Conclusion

Contemplate the criticality of subject in vogue; it is opined evaluation of process should be followed meticulously. Further post result perpetual ; examination is also indispensable; as being practiced in European countries & USA Transparency and action oriented for investigation should be "self less" and logical.

References

[1] Ian H. Witten and Eibe Frank, *DATA MINING Practical Machine Learning Tools and Techniques:* Morgan Kaufmann Publishers, An Imprint of Elsevier. 2005.

[2] *Declaration of Helsinki regarding Ethical Principles for Medical Research Involving Human Subjects.*

[3] Hillol Kargupta, Anupam Joshi, Krishna Moorthy Sivakumar and Yelena Yesha, *Data Mining Next generation challenges and future directions*: Prentice-Hall of India Pvt. Ltd. 2005.

[4] David Hand, Heikki Manila and Padhraic Smyth, *Principles of Data Mining*: Prentice-Hall of India Pvt. Ltd. 2005.

Proceedings of National Conference on Data Mining (NCDM-2011)

Clustering: an Efficient Technique for XML Data Management

Darshana Desai
(Lecturer)Department of MCA, Indira Institute of Management, Tathawade, Pune
e-mail: darshana.desai@indiraiimpmca.edu.in

***Abstract*—The eXtensible Markup Language (XML) has become an international standard for representation and exchange of information on the Web, because of its simple, self-describing capability and flexible organization of data over the heterogeneous environment. The increasing availability of XML data has made researcher aware of number of issues regarding management of XML data. As a result, discovering knowledge to infer semantic organization of XML data has become a major challenge for research in XML data management. This paper mentions different techniques used to measures similarity between two XML documents both for its semantics & data. It also describes how similarities that are computed are exploited by an agglomerative clustering algorithm to group xml data having similarity schemas & data. This paper focuses various clustering techniques for xml data and how hierarchical clustering can be used for efficient xml data management.**

Keywords: Clustering, Schema, Hierarchical Clustering, Agglomerative Clustering,

I. INTRODUCTION

The World Wide Web is a huge resource of data and this data is created, stored and transferred incrementally. As the current need of the distributed corporate world, is to have some data representation and transformation tool which can work on heterogeneous environment. XML (eXtensible Markup Language) documents are one of the best tools for representing and transferring data because of their flexibility and self description and increasing use of the information resources. Since XML tags describe the structural and semantic concepts of information in texts, these documents are known to be semi structured. XML is one of the widely used semi structured, self-describing language.

XML document is the collection of elements and each element is a pair of matching well- formed start and end-tags and all the data that appears between them. XML tags are user defined tags which represents data meaning enclosing its actual data. The elements of XML are organized in a nested structure with one root such that XML can be modeled as an ordered labeled tree [1]. Each node in this tree corresponds to a tag in the document and is labeled with the tag's name. Each edge in this tree represents inclusion of the tag corresponding to the child node under the tag corresponding to the parent node in the XML. The purpose of data clustering is to derive some relevant information from the various data for further data processing and managing data. The clustering of XML documents is to group similar documents to facilitate searching because similar documents can be searched and processed within a specific category.

The clustering of XML documents can be done based on data, structure or both. The appropriate clustering of XML documents is also effective for systematic document management and the efficient storage of XML documents. This paper describes various clustering methods to cluster XML documents efficiently. The path which composes of the important node for each level of a XML tree can represent both the structure and the contents of a XML document. We then propose a method to apply the well known hierarchical clustering techniques to the representative paths to cluster XML documents. We will also explain how the hierarchical clustering process is terminated when the clustering that best fits the data has been achieved according to some criterion.

II. RELATED WORK

The need for organizing and clustering XML data has become challenging, due to the increase of heterogeneity of XML sources. Recently, several clustering techniques which consider the structure and/or the contents of XML documents are studied. Ref. [2] applies a *K*-means clustering technique to XML documents represented in a vector-space model. In this representation, each document is represented by an *N*-dimensional vector, with *N*being the number of document features such as text features, tag features, and a combination of both in the collection. They only consider the contents of XML. In [3,4] a new bitmap indexing based technique to cluster XML documents is described. A BitCube is presented in as a 3-dimensional bitmap index of triplets (document, XML-element path, word). BitCube indexes can be manipulated to partition documents into clusters by exploiting bit-wise distance and popularity measures. However, this method needs manual operations to create a bitmap index. Ref. [5] devises features for XML data, focusing on content information extracted from textual elements and structure information derived from tag paths. They

introduce the notion of tree tuple in the definition of an XML representation model that allows for mapping XML document trees into transactional data, i.e., variable length sequences of objects with categorical attributes. A partitioned clustering approach has been developed and applied to the XML transactional domain. On the other hand, [6] transforms the structure of the XML document into a discrete function. The discrete function, then, is transformed into frequency domain by FFT. The result of FFT is a pair of complex numbers consisting of x and y values and considered to be a pair of n-dimensional vectors. The pairs of n-dimensional vectors are compared using a weighted Euclidean distance metric in an incremental and unsupervised fashion. This approach considers solely the structure of elements

III. CLUSTERING

Clustering of XML documents determines the groups of XML documents having similarity between its schema and the structure of the documents. There have been many techniques developed for finding similarity between XML documents or XML schemas. These techniques are used mainly in data or schema integration or query approximation. As well, these techniques facilitate the clustering process. They does by considering the XML semantic information (linguistic and context elements) as well as the hierarchical structure. The process usually starts by representing the XML document or schema into a tree presentation.

A. Clustering Framework

The increasing availability of heterogeneous XML data has raised a number of issues concerning how to represent and manage semi-structured data. As a result, discovering knowledge to infer semantic organization of XML data has become a major challenge in XML data management. A possible solution is to group similar XML data based on their content and structures. Grouping similar XML data according to structure or content or both among heterogeneous set of data is the process of XML data clustering. Clustering is a useful technique for grouping data objects such that objects within a single group/cluster have similar features, while objects in different groups are dissimilar [11]. The main steps involved in the data clustering activity, as shown in Fig. 1, are: (1) data representation: data objects are represented using a common data model;(2) definition of data proximity measures suitable to the data domain and data representation: data features and the proximity function to measure the similarity between pairs of data objects are determined; and(3) clustering or grouping: the similar data objects are grouped together based on the proximity function using clustering algorithms [11, 17].

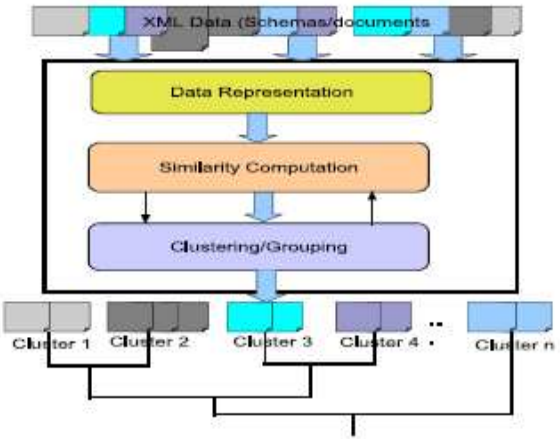


Fig. 1: Generic XML Data Clustering Framework

Sequential pattern mining algorithms [7] have been used by many researchers [8][9][10] to measure structural similarity. These algorithms represent a tree by a set of paths or sequences. A path is represented by a unique sequence of element nodes following the containment links from root to leaf nodes. The sequential pattern algorithm computes the maximal similar paths between XML documents. The combination of semantic and structural similarity is represented as a similarity matrix. K-means or hierarchical agglomerative clustering algorithms [11] generate clusters of XML documents. In order to perform clustering of XML documents similarity approach can be classified as categories based on their structural or semantic similarity and schema.

B. Clustering Based on Structure-Level Similarity Approach

The structure level similarity approaches detect and measure three different sets of data; (1) structural and content similarities between documents [12][13][14] (2) the structural similarity between documents and schemas[15] and (3) the structural and content similarity between schemas[8][16] apply one of the methods calculating similarity or distance between the XML document structures.

C. Clustering Based on Schema Level Similarity Approaches

The *Schema level similarity approaches* also known as Element level similarity to determine the semantic correspondence between elements of two schemas. These methods use the document schema to cluster XML documents. Relevant schema information is used to efficiently determine the similarity of corresponding elements in XML documents. The document schema provides a definitive description of the document, while document instances represent examples of content. The document definition outlined in a schema holds true for all document instances of that

schema, hence schema clustering results hold true for all document instances and can be reused for other instances. *Instancebased matchers*use either metadata or statistical data collected from data instances to annotate the schema or directly correlated schema elements [21].

IV. HIERARCHICAL CLUSTERING

Hierarchical clustering solutions are in the form of trees called *dendograms*, which provide a view of the data at different levels of abstraction. The consistency of clustering solutions at different levels of granularity allows flat partitions of different granularity to be extracted during data analysis, making them ideal for interactive exploration and visualization [11].Two primary methods to obtain hierarchical clustering solutions: *agglomerative algorithms* and *partitioned algorithms*.

In agglomerative algorithms, objects are initially assigned to it owns cluster and then the pairs of clusters are repeatedly merged until the whole tree is formed. However, partitioned algorithms can also be used via a sequence of repeated bisections. The partitioned algorithms are well suited for clustering large datasets due to their relatively low computational requirements. However, the agglomerative algorithms outperform partitioned algorithms. We propose a method to apply the well known hierarchical clustering algorithm when a representative path is used as the feature of a XML document. The hierarchical clustering algorithms produce a hierarchy of nested clustering. A clustering $\Re$containing *k*clusters is said to be nested in the clustering $\Re$, which contains *r*(<*k*) clusters, if each cluster in $\Re$is proper subset of . The pseudo code of general agglomerative clustering algorithm is described in Fig. 2 when the total number of patterns is *n*.

- Begin with n clusters, each consisting of one pattern.
- Repeat step a total n-I times.
- Find the most similar clusters C and C_j and merge $C_{I\ and}$ C_j into one cluster.
- Fig. 2. Agglomerative clustering algorithm

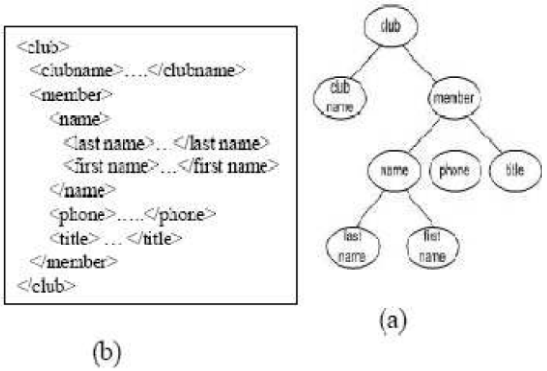


Fig. 2: Club XML document and its corresponding tree

One of the important issues for clustering process is how the similarity measure between patterns is quantified. We consider not only the node's name but also the node's position in the path to measure the similarity between XML documents. The similarity between the names of the two compared nodes can be obtained by the set of synonyms or thesauruses.

For example, suppose that the names of the compared nodes are different such that an 'actor' in one node and a 'star' in the other node. Then, the system assigns a value from 0 to 1 to the name weight according to the conformance level of the two names after the synonyms of '*star*' are extracted from a synonym database such as the WordNet [11]. However, even if the names of the compared nodes are the same, the weight can be different according to the positions of the nodes in the paths. This is because as the path to the node becomes short, more weight is assigned when the similar XML documents are searched [12]. Let the level weight of a node *xi*, ($1 \le i \le n$) in a representative path is *Lev x*, then the level weights satisfying the following conditions.

$$Lev_{x_1} \ge Lev_{x_2} \ge \ldots \ge Lev_{x_n} \text{ and } \sum_{i=1}^{n} Lev_{x_i} = 1 \quad (1)$$

Let the representative paths *X*and *Y*of the twoXML documents are *x1* / *x2* /K/ *xn* and y*1* / *y2* /K/ *ym*, Then, the similarity between two XML documents *Sim*(*X*,*Y*) (0≤*Sim*(*X*,*Y*)≤1) isdefined in the following way:

$$Sim(X,Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} Name(x_i, y_i) \times \min(Lev_{x_i}, Lev_{y_j}) \quad (2)$$

The reason of choosing the minimum value of the two level weights is that we want to reduce the influence of a weight as the difference between the levels of the compared nodes is large. Different agglomerative clustering algorithms are obtained by using different methods to determine the imilarity of clusters. The single-linkage algorithm is obtained by defining the distance between two clusters to be the smallest distance between two patterns such that one pattern is in each cluster. Therefore, if *Ci*and *Cj*are clusters, the distance between them is defined as:

$$d_{SL}(C_i, C_j) = \max_{X \in C_i, Y \in C_j} Sim(X,Y) \quad (3)$$

On the other hand, the complete-linkage algorithm is obtained by defining the distance between two clusters to be the largest distance between a pattern in one cluster and a pattern in the other cluster. Therefore, if *Ci*and *Cj*are clusters, the distance between them is defined as:

$$d_{CL}(C_i, C_j) = \min_{X \in C_i, Y \in C_j} Sim(X,Y) \quad (4)$$

Conclusion &Future Direction:

This paper proposes different approaches for xml data clustering like schema based or structure based with inter structure and intra structure. It also explains

hierarchical clustering algorithm in which agglomerative clustering algorithm is explained with path representative method. This work can be further implemented in details by implementing schema based clustering with different method like bit vector and also with representative path with actual data sets measuring performance with different method can be compared with the help of dendograms, to achieve best method for efficient data management for xml data.

REFERENCES

[1] Behrens, R., (2000), "A Grammar Based Model for XML Schema Integration, Proc. ofthe17thBritish National Conf. on Databases", pp.172–190.

[2] Doucet, A., and Ahonen-Myka, H., (2002), "Navie Clustering of a Large XML Document Collection", *Proc. 1st Annual Workshop of the Initiative for the Evaluation of XML retrival* (INEX), Germany, Dec., pp.81–88.

[3] Yoon, J., Raghavan, V., and Chakilam, V., (2001), "Bit Cube: Clustering and Statistical Analysis for XML Documents", *Proc. ofthe13th Int. Conf. onScientificand StatisticalDatabaseM anagement*, Fairfax, Virginia, July.

[4] Yoon, J., Raghavan, V., Chakilam, V., and Kerschberg, L., (2001), "Bit Cube: A 3-D Bitmap Indexing for XML Documents, Journal of Intelligent Information Systems", Vol. 17, November, pp.241–254.

[5] Tagarelli, A., and Greco, S., (2006), "Toward Semantic XML Clustering, *6th SIAM International Conferenceon Data Mining (SDM'06)*, Bethesda, Maryland" USA, April, pp. 188–199.

[6] Lee, H., (2007), "An Unsupervised Clustering Technique of XML Documents based on Function Transform and FFT, *Journal of Korea Information Processing Society*".

[7] Agrawal, R., & Srikant, R., (1996), "Mining Sequential Patterns: Generalizations and Performance Improvements. Paper presented at the the 5th International Conference on Extending Database Technology (EDBT'96), France".

[8] Nayak, R., and Iryadi, W., (2007), "XML schema clustering with semantic and hierarchical similarity measures. *Knowledge-Based Systems", 20*(4), pp. 336–349.

[9] Lee, J., W., and Park, S. S. (2004). "Finding Maximal Similar Paths Between XML Documents Using Sequential Patterns. Paper presented at the ADVIS, Izmir, Turkey".

[10] Leung, H.-p., Chung, F.-l., & Chan, S. C.-f.(2005), "On the use of hierarchical information in sequential miningbased XML document similarity computation. Knowledgeand InformationSystems",7(4), pp. 476–498.

[11] Jain, A. Murty, K., M. N., and Flynn, P. J. (1999), "Data Clustering: A Review. *ACM Computing Surveys" (CSUR),31*(3), pp. 264–323.

[12] Dalamagas, T., Cheng, T., Winkel, K., and Sellis, T.K., (2004). "Clustering XML documents by Structure. Paperpresented at the SETN".

[13] Flesca, S., Manco, G., Masciari, E., Pontieri, L., et al. (2005), Fast Detection of XML Structural Similarities. IEEE Transaction on Knowledge and Data Engineering, 7(2), pp. 160–175.

[14] Huang, Z. (1997), *A fast clustering algorithm to cluster very large categorical data setsin data mining.*Paper presented at the SIGMOD workshop on Research Issues on Data Mining and Knowledge Discovery *Surveys*, 31(3):264–323, 1999.

[15] Y. B. Idrissi and J. Vachon. Evaluation ofhierarchical clustering algorithms for documentdatasets. In the11th InternationalConferenceonInformation andknowledgeM anagement, pages515–524, 2002.

[16] A. Jain, M. Murty, and P. Flynn. Dataclustering: A review. *ACM Computing*

[17] P. Berkhin. Grouping Multidimensional Data:Recent Advances in Clustering, chapterSurvey of Clustering Data Mining Techniques ,pages 25–71. Springer BerlinHeidelberg, 2006.

Author Index