ISSN(ONLINE):2045-8741
ISSN(PRINT) :2045-869X

International Journal of Innovative Technology & Creative Engineering

APRIL 2011
No.1 Vol.4

UK: Managing Editor
International Journal of Innovative Technology and Creative Engineering
1a park lane,
Cranford
London
TW59WA
UK
E-Mail: editor@ijitce.co.uk
Phone: +44-773-043-0249

USA: Editor
International Journal of Innovative Technology and Creative Engineering
Dr. Arumugam
Department of Chemistry
University of Georgia
GA-30602, USA.
Phone: 001-706-206-0812
Fax:001-706-542-2626

India: Editor
International Journal of Innovative Technology & Creative Engineering
Dr. Arthanariee. A. M
Finance Tracking Center India
261 Mel quarters
Labor colony,
Guindy,
Chennai -600032.
Mobile: 91-7598208700

IJITCE PUBLICATION

INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY & CREATIVE ENGINEERING

Vol.1 No.4

April 2011

www.ijitce.co.uk

From Editor's Desk

Dear Researcher,

Greetings!

The researchers in this issue reflect a range of perspectives on E-learning, Data Mining, Encryption, Clustering and Security youth librarianship,

BNP economic research reflects it is too early to assess the full consequences of recent global events (the earthquake in Japan, rising commodity prices and debt crisis).

In US February and March job figures suggest that the labor market is strengthening, with 216,000 jobs created in March, after 194,000 in February. This is the strongest performance since last May.

In Japan Economic activity has been severely affected by serious supply constraints. These concern in the first place the electricity supply. Before the disaster, total capacity of the Tokyo Electric Power Company (TEPCO) was 78 GW. In early April, its total capacity amounted to 38.5 GW, which was sufficient to deal with demand. The situation is similar for Tohoku EPCO.

It is possible to generate clean energy from many elements, which are not widely used now for producing energy. Elements like Lithium (Li) and Cobalt (Co) can be used to generate energy. Lithium is highly reactive and has low mass. Because of this property, it is used in rechargeable batteries to store charge. Lithium is found in abundance and is also not dangerous to health. Cobalt is also found in abundance and is used as electrodes in batteries and in highly efficient jet turbines. This can be a good source of clean energy.

To conclude, we need good policies to implement and streamline the usage of these elements in our energy production. These are not only cost effective, but also environment safe.

It has been an absolute pleasure to present you articles that you wish to read. We look forward to many more new technology-related research articles from you and your friends. We are anxiously awaiting the rich and thorough research papers that have been prepared by our authors for the next issue.

Thanks,
Editorial Team
IJITCE

Editorial Members

Review Board Members

Contents

An Adaptive Technique Using Advanced Encryption Standard To Implement Hard Disk Security

Minal Moharir, Dr. A V Suresh,
R V College of Engg., Bangalore-59,India

Abstract: **The main objective of the paper is to study and develop an efficient method for Hard Disk Drive(HDD) Security using Full Disk Encryption (FDE) with Advanced Encryption Standards(AES) for data security specifically for Personal Computers(PCS) and Laptops.**

The focus of this work is to authenticate and protect the content of HDD from illegal use. The paper proposes an adaptive methods for protecting a HDD based on FDE. The proposed method is labeled as DiskTrust. FDE encrypts entire content or a single volume on your disk.

DiskTrust implements Symmetric key cryptography with, Advanced Encryption Standards.

Finally, the applicability of these methodologies for HDD security will be evaluated on a set of data files with different key sizes.

Keywords: Information Security, Integrity, confidentiality, Authentication, Encryption.

1 INTRODUCTION

As of January 2011 the internet connected an estimated 941.7 million computers in more than 450 countries on every continent, even Antarctica (Source: Internet Software Consortium's Internet Domain Survey; www.isc.org/index.pl). The internet is not a single network, but a worldwide collection of loosely connected networks that are accessible by individual computer hosts, in a variety of ways, to anyone with a computer and a network connection. Thus, individuals and organizations can reach any point on the internet without regard to national or geographic boundaries or time of day.

However, along with the convenience and easy access to information come risks. Among them are the risks that valuable information will be lost, stolen, changed, or misused. If information is recorded electronically and is available on networked computers, it is more vulnerable than if the same information is printed on paper and locked in a file cabinet. Intruders do not need to enter an office or home; they may not even be in the same country. They can steal or tamper with information without touching a piece of paper or a photocopier. In this way security of stored information is an important issue. The proposed paper consider the security of Hard Disk Drive which is a fundamental element in computing chain.

The paper organized as follows. Related work, gap & problem is described in Section 2. A view of simulation and experimental design is given in section 3. Simulation results are shown in section 4. Finally the conclusions are drawn section 5.

2 RLATED WORK

The related survey is divided into two parts. The first part is survey about full disk encryption. The second part is survey about advanced encryption standards.

Information security is the process of protecting information. It protects its availability, privacy and integrity. More companies store business and individual information on computer than ever before. Much of the information stored is highly confidential and not for public viewing. Without this information, it would often be very hard for a business to operate. Information security systems need to be implemented to protect this information. There are various ways to implement Information security systems. One of the popular technique is full disk encryption. **F**ull **D**isk **E**ncryption (FDE) is the safest way to protect digital assets, the hard drive is a critical element in the computing chain because it is where sensitive data is stored. Full disk encryption increases the security of information stored on a laptop significantly. It helps to keep business critical data absolutely confidential. Moreover, full disk encryption helps to meet several legislative requirements. Various techniques to implement FDE are discussed as follows: **Tabel1**:

Name	Developed	Released	Licence	OS
TrueCrypt	TrueCrypt Foundation	2009	Free	Linux, Windows
Discryptor	Cosect	2008	Commercial, closed source	Windows, Vista
DriveSentry	DriveSentry	2008	Commercial, closed source	Windows, Vista
R-Crypto	R-Tools Technology Inc	2008	Free, closed source	Windows XP, Vista

The second part of survey covers implementation of Encryption Algorithms. Many encryption algorithms are widely available and used in information security. They can be categorized into Symmetric (private) and Asymmetric (public) keys encryption. In Symmetric keys encryption or secret key encryption, only one key is used to encrypt and decrypt data. The key should be distributed before transmission between entities. Keys play an important role. If weak key is used in algorithm then every one may decrypt the data. Strength of Symmetric key encryption depends on the size of key used. For the same algorithm, encryption using longer key is harder to break than the one done using smaller key. Brief definitions of the most common encryption techniques are given as follows: DES: (Data Encryption Standard), was the firstencryption standard to be recommended by NIST (National Institute of Standards and Technology).DES is (64 bits key size with 64 bits block size) . Since that time, many attacks and methods recorded the weaknesses of DES, which made it an insecure block cipher [3],[4].3DES is an enhancement of DES; it is 64 bit block size with 192 bits key size. In this standard the encryption method is similar to the one in the original DES but applied 3 times to increase the encryption level and the average safe time. It is a known fact that 3DES is slower than other block cipher methods [3]. RC2 is a block cipher with a 64-bits block cipher with a variable key size that range from 8 to128 bits. RC2 is vulnerable to a related-key attack using 234 chosen plaintexts [3]. Blowfish is block cipher 64-bit block - can be used as a replacement for the DES algorithm. It takes a variablelength key, ranging from 32 bits to 448 bits; default 128 bits. Blowfish is unpatented, license-free, and is available free for all uses. Blowfish has variants of 14 rounds or less. Blowfish is successor to Twofish [5]. AES is a block cipher .It has variable key length of 128, 192, or 256 bits; default 256. it encrypts data blocks of 128 bits in 10, 12 and 14 round depending on the key size. AES encryption is fast and flexible; it can be implemented on various platforms especially in small devices[6]. Also, AES has been carefully tested for many security applications [3], [7]. RC6 is block cipher derived from RC5. It was designed to meet the requirements of the Advanced Encryption Standard competition. RC6 proper has a block size of 128 bits and supports key sizes of 128, 192 and 256 bits. Some references consider RC6 as Advanced Encryption Standard [8].

2.1 Research Gap

The FDE technology discussed in above survey are encrypting the entire contents of Hard disk Drive. However encryption of the entire HDD is expensive in terms of time and cost. DiskTrust, the technology proposed here, creates a hidden volume on HDD, which is not visible, accessible to the unauthorized user. The data store in this hidden volume is encrypted using robust(Rijndael) encryption algorithm. In this way DiskTrust technology follows CIA properties of secure information along with hidden partition.

2.2 Problem Definition

DiskTrust technology implements security on the hard drive itself, to provide a foundation for trusted computing.

The technical objectives of the paper are:

1. Create Hidden partition
2. Execute issuance protocol to check authentication.
3. Execute encryption/decryption algorithm while reading /writing data on Hard Disk Drive.

3 SIMULATION AND DESIGN:

This section describes some of the important results that were found as part of the implementatton.

3.1 Implementation of Hidden Volume

DiskTrust Security user interfaces are shown below in the screenshots. The user interface is basically a frame work application where user can use the application

3.1.1 Main Application Window

The Main application window holds multiple options such as CreateVolume, Mount and Dismount All.

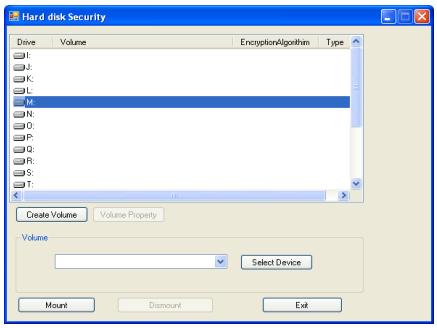

Figure 3.1.1 Screenshot of Main Application Window

3.1.2 Volume Location

Volume Location Window allows the user to select the file for which user want to create volume.

Figure 3.1.2 Screenshot of Volume Location Window

3.2 Volume Password

Volume Password window will allow the user to enter the password and confirm Password.

Password implements user authentication.

Figure 3.2.1 Screenshot **of Volume Password Window**

3.3Encrypt or decrypt data while retrieving from hidden volume:

For our experiment, we use a laptop PentiumV 2.4 GHz CPU, in which performance data is collected. In the experiments, the laptop encrypts a different file size ranges from 321K byte to 7.139Mega Byte. Several performance metrics are collected:

1- encryption time

2- CPU process time

3- CPU clock cycles and battery power.

The encryption time is considered the time that an encryption algorithm takes to produce a cipher text from a plaintext. Encryption time is used to calculate the of an encryption scheme. It indicates the speed of encryption. The CPU process time is the time that a CPU is committed only to the particular process of calculations. It reflects the load of the CPU. The more CPU time is used in the encryption process, the higher is the load of the CPU. The CPU clock cycles are a metric, reflecting the energy consumption of the CPU while operating on encryption operations. Each cycle of CPU will consume a small amount of energy.

4 Simulation Results

The effect of changing key size of AES on power consumption. The performance comparison point is the changing different key sizes for AES algorithm. In case of AES, We consider the three different key sizes possible. In case of AES it can be seen that higher key size leads to clear change in the battery and time consumption. It can

be seen that going from 128 bits key to 192 bits causes increase in power and time consumption about 8% and to 256 bit key causes an increase of 16% [9]. The simulation results with different key sizes are as shown in Table2.

AES Key Size	AES 128	AES 128	AES 256
Time in Milliseconds	287	310	330

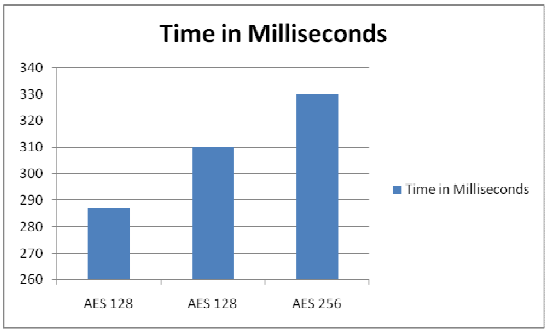

Figure3.3.1 Time with different key size

6 Conclusion

Full disk encryption (FDE) appears to offer an ideal solution to the losses of data on laptops, CDs and thumb drives. The DiskTrust technique proposed in the paper instead encrypting the entire contents of disk, it encryupts a data stored on single volume which less expensive in terms of time & cost. Disktrust provides authentication, integrity and confidentiality for stored data. DiskTrust implements Symmetric key cryptography with AES. AES is more promising according to results.

7 Bibliography

[1] William A. Arbaugh, Angelos D. Keromytis, David J.Farber, and Jonathan M. Smith, Automated Recovery in a Secure Bootstrap Process, Network and Distributed SystemSecurity Symposium, Internet Society, March 1998

[2] Siani Pearson, Trusted Computing Platforms: TCPATechnology in Context, Prentice Hall PTR, 2002.

[3] G. Piret, J.J. Quisquater. "A Differential Fault Attack Technique Against SPN Structures, with Application to the AES and Khazad" Workshop on Cryptographic Hardware and Embedded Systems, CHES 2003. LNCS, vol. 2779, Springer-Verlag, pp. 77-88, 2003.

[4] Hiroshi Maruyama and others, Linux with TCPA Integrity Measurement, IBM Research, Tokyo Research Laboratory, January 28, 2003.

[5] Hiroshi Maruyama and others, Trusted Platform on demand (TPod), IBM, February 1, 2004.

[6]. Michael Austin Halcrow, eCryptfs: An Enterprise-class Cryptographic Filesystem for Linux, International Business Machines Inc., 2005.

[7] Daniela A. S. de Oliveira, Jedidiah R. Crandall, and others, ExecRecorder: VM-based full-system replay for attack analysis and system recovery, Proceedings of the 1st workshopon Architectural and system support for improving software dependability, ACM, 2006.

[8] Peng Shaunghe, Han Zhen, Enhancing PC Security with a U-Key, IEEE Security and Privacy, Volume 4, Issue 5, September 2006.

[9] López-Ongil et al. "Autonomous Fault Emulation: A New FPGA-based Acceleration System for Hardness Evaluation" IEEE T. on Nuclear Science, Vol. 54, Issue1, Part 2, pp. 252-261, Feb. 2007.

[11] W. Diffie and S. Landau, Privacy on the Line: The Politics of Wiretapping and Encryption, updated and expanded edition, MIT Press, 2007, pp. 280–285.

[12] H. Rezaei Ghaleh, PC Secure Bootstrapping, M.Sc. Thesis, Islamic Azad University of Qazvin, March 2008.

[13] BitLocker Drive Encryption Technical Overview, Microsoft TechNet, 2008.

[14] Alexei Czeskis David J. St. Hilaire Karl Koscher Steven D. Gribble, 2008, Defeating Encrypted and Deniable File Systems:TrueCrypt v5.1a and the Case of the Tattling OS and Applications

[15] David Challener, Kent Yoder, Ryan Catherman, DavidSafford, Leendert Van Doorn, A Practical Guide to TrustedComputing, IBM Press, 1 edition, January 6, 2008.

Modification Of Gtd From Flat File Format To Olap For Data Mining

Karanjit Singh and Dr. Shuchita Bhasin
HQ Base Workshop Group EME, Indian Army, Computer Science and IT Dept, Kurukshetra University

Abstract - This document is part of original research work by the authors in a bid to explore new fields for applying Data Mining Techniques. The sample data is part of a large data set from University of Maryland (UMD) and outlines how more meaningful patterns can be discovered by preprocessing the data in the form of OLAP cubes

Keywords: GTD, OLAP, Data Mining, Terror Databases

I. INTRODUCTION

Application of Data Mining Tools for Terror Data Mining is a lesser talked about field[1]. Lot of research efforts are going into capturing the data from incident reports in the past and structuring the data for analysis. Unfortunately there are not many sources on the net. One such database available [2] in a single tabular form, is an Open Source Terrorism Incident events Database called Global Terrorism Database (GTD). This covers terrorism incidents around the world from 1970 through 2008 (with continuing annual updates). It includes systematic data on US, as well as transnational and international terrorist incidents that have occurred during this time period and as on now includes more than 87,000 cases. For each GTD incident, information is available on the date and location of the incident, the weapons used and nature of the target, the number of casualties, and--when identifiable--the group or individual responsible. However the format in which it is available lends itself only to limited analysis unless suitable tools for analysis are used. This paper analyses the available data fields and suggests a format for OLAP and subsequent data mining. The data base has been obtained from the National Consortium for the Study of Terrorism and Responses to Terrorism (START) initiative at University of Maryland, from their online interface at http::://www.start.umd.edu/gtd/ in an effort to increase understanding of terrorist violence so that it can be more readily studied and defeated.

II. CHARACTERISTICS OF AVAILABLE DATA

The main characteristics of the GTD [4] are:-

- Information on over 87,000 terrorist attacks
- Currently the most comprehensive unclassified data base on terrorist events in the world
- Information on more than 38,000 bombings, 13,000 assassinations, and 4,000 kidnappings since 1970
- Includes information on at least 45 variables for each case, with more recent incidents including information on more than 120 variables
- Supervised by an advisory panel of 12 terrorism research experts
- Over 3,500,000 news articles and 25,000 news sources reviewed to collect incident data from 1998 to 2008 alone
- Available to Government representatives and interested researchers directly through their Online interface.

III. AVAILABLE DATABASE VARIABLES

A. GTD ID (eventid) (Numeric)

The incidents follow a 12-digit Event ID system.
The first 8 numbers – Recording date " yyyymmdd".
Next 2 numbers – always Zero Zero "00".
Last two numbers – case number for the given day (01,02 etc.) This will be 00 unless there is more than one case on the same date.
For example, an incident in the GTD occurring on 25 July 1993 would be numbered as "199307250001". An additional GTD case recorded for the same day would be "199307250002". The next GTD case recorded for that day would be "199307250003", etc.

To determine whether an incident is single, incidents occurring in both the same geographic and temporal point will be regarded as a single incident, but if either the *time* of occurrence of incidents or their *locations* are *discontinuous*, the events will be regarded as separate incidents.

Examples:

- *Four truck bombs explode nearly simultaneously in different parts of a major city.* This represents four incidents.

- *A bomb goes off, and while police are working on the scene the next day, they are attacked by terrorists with automatic weapons.* These are two separate incidents, as they were not continuous, given the time lag between the two events.

- *A group of militants shoot and kill five guards at a perimeter checkpoint of a petroleum refinery and then proceeds to set explosives and destroy the refinery.* This is one incident since it occurred in a single location (the petroleum refinery) and was one continuous event.

- *A group of hijackers diverts a plane to Senegal and, while at an airport in Senegal, shoots two Senegalese policemen.* This is one incident, since the hijacking was still in progress at the time of the shooting and hence the two events occurred at the same time in the same place.

If the information available for such complex events does not specify the time lag between or the exact locations of multiple terrorist activities, the event is a single incident.

IV. INCIDENT DATE

A. *ear (iyear) Numeric*

This field contains the year in which the incident occurred. In the case of incident(s) occurring over an extended period, the field will record the year when the incident was initiated. When the year of the incident is unknown, this will be recorded as “0”.

B. *Month (imonth) Numeric*

This field contains the number of the month in which the incident occurred. In the case of incident(s) occurring over an extended period, the field will record the month when the incident was initiated. When the exact month of the incident is unknown, this will be recorded as “0”. For the cube this could form part of the Time dimension.

C. *Day (iday) Numeric*

This field contains the numeric day of the month on which the incident occurred. In the case of incident(s) occurring over an extended period, the field will record the day when the incident was initiated.
When the exact day of the incident is unknown, the field is recorded as “0”.

D. *Approximate Date (approxdate) Text*

Whenever the exact date of the incident is not known or remains unclear, this field is used to record the approximate date of the incident.

- If the day of the incident is not known, then the value for “Day” is “0”.
- For example, if an incident occurred in June 1978 and the exact day is not known, then the value for the “Day” field is “0” and the value for the “Approximate Date” field is “June 1978”.
- If the month is not known, then the value for the “Month” field is “0”.
- For example, if an incident occurred in the first half of 1978, and the values for the day and the month are not known, then the value for the “Day” and “Month” fields will both be “0” and the value for the “Approximate Date” field is “first half of 1978”.

E. *Extended Incident? (extended) Categorical*

1 = "Yes" The duration of an incident extended more than 24 hours.
0 = "No" The duration of an incident extended less than 24 hours.

F. *Date of Extended Incident Resolution (resolution)Date*

This field only applies if “Extended Incident?” is “Yes” and records the date in which the incident was resolved (hostages released by perpetrators; hostages killed; successful rescue, etc.)

It may be seen that variables categorised in this sub section Paras A to F can help form the Time dimension with the desired granularity.

V. INCIDENT LOCATION

A. *Country (country; country_txt) Categorical Variable*

This field identifies the country or location where the incident occurred. This includes non-independent states, dependencies, and territories, such as Northern Ireland and Corsica. If an incident occurs in an autonomous or geographically non-contiguous area, it is listed separately from the “home” country. However, separatist regions, such as Kashmir, Chechnya, South Ossetia, Transnistria, or Republic of Cabinda, are coded as part of the “home” country. West Bank and Gaza Strip have been coded separately from Israel. If an incident took place in a city located in the West Bank or Gaza Strip, it has been coded accordingly.
When an incident occurred in international waters or airspace, the country of departure is listed as the country of the incident. If the departure point is not identified, the incident is coded as “International.”
In cases where hostages were taken, the country where the incident began is recorded as the incident location, and a separate field captures the country where the incident was resolved or ended.

In the case where the country in which an incident occurred cannot be identified, it is coded as "Unknown
The political circumstances of many countries have changed over time. In a number of cases, countries that represented the location of terrorist attacks no longer exist; examples include West Germany, the USSR and Yugoslavia. In these cases the country name for the year the event occurred is recorded. As an example, a 1989 attack in Bonn would be recorded as taking place in West Germany (FRG). An identical attack in 1991 would be recorded as taking place in Germany. The dates which apply as watersheds are as given below.
Eritrea – independence: 24 May 1993.
Germany – unification: 3 October 1990.
Breakup of Czechoslovakia
Czech Republic – independence: 1 January 1993;
Slovakia – independence: 1 January 1993;
Breakup of USSR
Russian Federation – Independence: 24 August 1991;
Armenia – Independence: 21 September 1991;
Azerbaijan – Independence: 30 August 1991;
Belarus – independence: 25 August 1991;
Estonia – independence: 17 September 1991;
Georgia – independence: 9 April 1991;
Kazakhstan – independence: 16 December 1991;
Kyrgyzstan – independence: 31 August 1991;
Latvia – independence: 21 August 1991;
Lithuania – independence: 17 September 1991;
Moldova – independence: 27 August 1991;
Tajikistan – independence: 9 September 1991;
Turkmenistan – independence: 27 October 1991;
Ukraine – independence: 24 August 1991;
Uzbekistan – independence: 1 September 1991;
USSR terminates: 26 December 1991 – 5 January 1992.
Breakup of Yugoslavia:
Bosnia and Herzegovina – independence: 11 April 1992;
Croatia – independence: 25 June 1991;
Kosovo – independence: 17 February 2008:
Macedonia – independence: 8 September 1991;
Yugoslavia turns Serbia-Montenegro: 4 February 2003;
Montenegro – independence: 3 June 2006;
Serbia – independence: 3 June 2006
Slovenia – independence: 1 January 1992.

Country (Location) Codes (Note: These codes are also used for the target nationality fields)

4 = Afghanistan
5 = Albania
6 = Algeria
7 = Andorra
8 = Angola
10 = Antigua and Barbuda
11 = Argentina
12 = Armenia
14 = Australia
15 = Austria
16 = Azerbaijan
17 = Bahamas
18 = Bahrain
19 = Bangladesh
20 = Barbados
21 = Belgium
22 = Belize
23 = Benin
24 = Bermuda
25 = Bhutan
26 = Bolivia
28 = Bosnia-Herzegovina
29 = Botswana
30 = Brazil
31 = Brunei
32 = Bulgaria
33 = Burkina Faso
34 = Burundi
35 = Belarus
36 = Cambodia
37 = Cameroon
38 = Canada
40 = Cayman Islands
41 = Central African Republic
42 = Chad
43 = Chile
44 = China
45 = Colombia
46 = Comoros
47 = Congo (Brazzaville)
49 = Costa Rica
50 = Croatia
51 = Cuba
53 = Cyprus
54 = Czech Republic
55 = Denmark
56 = Djibouti
57 = Dominica
58 = Dominican Republic
59 = Ecuador
60 = Egypt
61 = El Salvador
62 = Equatorial Guinea
63 = Eritrea
64 = Estonia
65 = Ethiopia
66 = Falkland Islands
67 = Fiji
68 = Finland
69 = France
70 = French Guiana
71 = French Polynesia
72 = Gabon
73 = Gambia
74 = Georgia
75 = Germany
76 = Ghana
77 = Gibraltar

78 = Greece
79 = Greenland
80 = Grenada
81 = Guadeloupe
83 = Guatemala
84 = Guinea
85 = Guinea-Bissau
86 = Guyana
87 = Haiti
88 = Honduras
89 = Hong Kong
90 = Hungary
91 = Iceland
92 = India
93 = Indonesia
94 = Iran
95 = Iraq
96 = Ireland
97 = Israel
98 = Italy
99 = Ivory Coast
100 = Jamaica
101 = Japan
102 = Jordan
103 = Kazakhstan
104 = Kenya
106 = Kuwait
107 = Kyrgyzstan
108 = Laos
109 = Latvia
110 = Lebanon
111 = Lesotho
112 = Liberia
113 = Libya
115 = Lithuania
116 = Luxembourg
117 = Macau
118 = Macedonia
119 = Madagascar
120 = Malawi
121 = Malaysia
122 = Maldives
123 = Mali
124 = Malta
125 = Man, Isle of
127 = Martinique
128 = Mauritania
129 = Mauritius
130 = Mexico
132 = Moldova
134 = Mongolia
136 = Morocco
137 = Mozambique
138 = Myanmar
139 = Namibia
141 = Nepal
142 = Netherlands
143 = New Caledonia
144 = New Zealand
145 = Nicaragua
146 = Niger
147 = Nigeria
149 = North Korea
151 = Norway
152 = Oman
153 = Pakistan
155 = West Bank and Gaza Strip
156 = Panama
157 = Papua New Guinea
158 = Paraguay
159 = Peru
160 = Philippines
161 = Poland
162 = Portugal
163 = Puerto Rico
164 = Qatar
166 = Romania
167 = Russia
168 = Rwanda
173 = Saudi Arabia
174 = Senegal
175 = Serbia-Montenegro
176 = Seychelles
177 = Sierra Leone
178 = Singapore
179 = Slovak Republic
180 = Slovenia
181 = Solomon Islands
182 = Somalia
183 = South Africa
184 = South Korea
185 = Spain
186 = Sri Lanka
189 = St. Kitts and Nevis
195 = Sudan
196 = Suriname
197 = Swaziland
198 = Sweden
199 = Switzerland
200 = Syria
201 = Taiwan
202 = Tajikistan
203 = Tanzania
204 = Togo
205 = Thailand
207 = Trinidad and Tobago
208 = Tunisia
209 = Turkey
213 = Uganda
214 = Ukraine
215 = United Arab Emirates
216 = Great Britain
217 = United States
218 = Uruguay
219 = Uzbekistan
220 = Vanuatu

221 = Vatican City
222 = Venezuela
223 = Vietnam
225 = Virgin Islands (U.S.)
226 = Wallis and Futuna
227 = Samoa (Western Samoa)
228 = Yemen
229 = Congo (Kinshasa)
230 = Zambia
231 = Zimbabwe
233 = Northern Ireland
235 = Yugoslavia
236 = Czechoslovakia
238 = Corsica
296 = Kurdish
311 = Roma (Gypsy)
321 = Arab
334 = Asian
338 = African
347 = Timor-Leste
349 = Western Sahara
351 = Commonwealth of Independent States
359 = Soviet Union
362 = West Germany (FRG)
376 = Korea
377 = North Yemen
381 = Jewish
383 = Peru/U.S.
403 = Rhodesia
406 = South Yemen
422 = International
428 = South Vietnam
449 = Hindu
499 = East Germany (GDR)
512 = European
520 = Sinhalese
523 = Tuareg
529 = Middle Eastern
532 = New Hebrides
1003 = Kosovo

B. Region (region; region_txt) Categorical Variable

This field identifies the region in which the incident occurred. The regions are divided into the following 13 categories:

1= North America (Canada, Mexico, United States)
2= Central America & Caribbean (Antigua and Barbuda, Bahamas, Barbados, Belize, Bermuda, Cayman Islands, Costa Rica, Cuba, Dominica, Dominican Republic, El Salvador, Grenada, Guadeloupe, Guatemala, Haiti, Honduras, Jamaica, Martinique, Nicaragua, Panama, Puerto Rico, St. Kitts and Nevis, Trinidad and Tobago, Virgin Islands (U.S.))
3= South America (Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Falkland Islands, French Guiana, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela)
4= East Asia (China, Hong Kong, Japan, Macau, North Korea, South Korea, Taiwan)
5= Southeast Asia (Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, South Vietnam, Thailand, Timor-Leste, Vietnam)
6= South Asia (Afghanistan, Bangladesh, Bhutan, India, Maldives, Mauritius, Nepal, Pakistan, Seychelles, Sri Lanka)
7= Central Asia (Kazakhstan, Kyrgyzstan, Tajikistan, Uzbekistan)
8= Western Europe (Andorra, Austria, Belgium, Corsica, Denmark, East Germany (GDR), Finland, France, Germany, Gibraltar, Great Britain, Greece, Iceland, Ireland, Italy, Luxembourg, Malta, Man, Isle of, Netherlands, Northern Ireland, Norway, Portugal, Spain, Sweden, Switzerland, West Germany (FRG))
9= Eastern Europe (Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Czech Republic, Czechoslovakia, Hungary, Kosovo, Macedonia, Moldova, Poland, Romania, Serbia-Montenegro, Slovak Republic, Slovenia, Yugoslavia)
10= Middle East & North Africa (Algeria, Bahrain, Cyprus, Egypt, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Libya, Morocco, North Yemen, Qatar, Saudi Arabia, South Yemen, Syria, Tunisia, Turkey, United Arab Emirates, West Bank and Gaza Strip, Western Sahara, Yemen)
11= Sub-Saharan Africa (Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Congo (Brazzaville), Congo (Kinshasa), Djibouti, Equatorial Guinea, Eritrea, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Ivory Coast, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Nigeria, Rhodesia, Rwanda, Senegal, Sierra Leone, Somalia, South Africa, Sudan, Swaziland, Tanzania, Togo, Uganda, Zambia, Zimbabwe)
12= Russia & the Newly Independent States (NIS) (Armenia, Azerbaijan, Belarus, Estonia, Georgia, Latvia, Lithuania, Russia, Soviet Union, Ukraine)
13= Australasia & Oceania (Australia, Fiji, French Polynesia, New Caledonia, New Hebrides, New Zealand, Papua New Guinea, Samoa (Western Samoa), Solomon Islands, Vanuatu, Wallis and Futuna)

C. Province / Administrative Region / U.S. State ((provstate) Text Variable)

This variable records the name of the province, administrative region or U.S. State.

D. City (city) Text Variable

This field contains the name of the city in which the incident occurred.

E. Vicinity (vicinity) Categorical Variable

- 1 = "Yes" The incident occurred in the vicinity of the city in question.

- 0 = "No" The incident in the city itself.

F. Location Description (location) Text Variable

This field is used to specify additional information about the location of the incident.

The above region, country, province fields could be used to form a a hierarchical Space Dimension starting at top level regions followed by country and then province. This level could further have the level of City but the data base has not covered this aspect in details as not all cities or provinces are covered by terror incidents.

VI. INCIDENT INFORMATION

A. Incident Summary (summary) Text Variable

A narrative summary of the incident, noting the "when, where, who, what, how, and why." This field is available with incidents occurring after 1997.

B. Criteria Categorical Variables

These variables record, as to which of the inclusion criteria (in addition to the necessary criteria) are met. This allows users to filter out those incidents whose inclusion was based on a criterion which they believe does not constitute terrorism proper.

CRITERION 1: Political, Economic, Religious, or Social Goal (crit1) - The violent act must be aimed at attaining a political, economic, religious, or social goal. This criterion is not satisfied in those cases where the perpetrator(s) acted out of a pure profit motive or from an idiosyncratic personal motive unconnected with broader societal change.

1 = "Yes" The incident meets Criterion 1.
0 = "No" The incident does not meet Criterion 1.

CRITERION 2: Intention to Coerce, Intimidate or Publicize to Larger Audience(s) (crit2) - To satisfy this criterion there must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience (or audiences) than the immediate victims. Such evidence can include (but is not limited to) the following: pre- or post-attack statements by the perpetrator(s), past behavior by the perpetrators, or the particular nature of the target, weapon, or attack type.

1 = "Yes" The incident meets Criterion 2.
0 = "No" The incident does not meet Criterion 2.

CRITERION 3: Outside International Humanitarian Law (crit3) - The action must be outside the context of legitimate warfare activities, i.e. the act must be outside the parameters permitted by international humanitarian law (jus in bello) as reflected in the Additional Protocol to the Geneva Conventions of 12 August 1949 and elsewhere. Specifically, if an attack contravenes any of the following, this criterion is met:

Persons who are not, or are no longer, taking part in hostilities shall be respected, protected and treated humanely. They shall be given appropriate care, without any discrimination.

Captured combatants and other persons whose freedom has been restricted shall be treated humanely. They shall be protected against all acts of violence, in particular against torture. If put on trial, captured combatants shall enjoy the fundamental guarantees of a regular judicial procedure.

The right of parties to an armed conflict to choose methods or means of warfare is not unlimited. No superfluous injury or unnecessary suffering shall be inflicted.

In order to spare the civilian population, armed forces shall at all times distinguish between the civilian population and civilian objects on the one hand, and military objectives on the other. Neither the civilian population as such nor individual civilians or civilian objects shall be the targets of military attacks.

1 = "Yes" The incident meets Criterion 3.
0 = "No" The incident does not meet Criterion 3.

C. Doubt Terrorism Proper? (doubtterr) Categorical Variable

In certain cases there may be some uncertainty whether an incident meets all of the criteria for inclusion. In these ambiguous cases, where there is a strong possibility, but not certainty, that an incident represents an act of terrorism, the incident is included in GTD and is coded as "Yes" for this variable.

- 1 = "Yes" There is doubt as to whether the incident is an act of terrorism.
- 0 = "No" There is essentially no doubt as to whether the incident is an act of terrorism.

This field is presently only available with incidents occurring after 1997. Incidents occurring before 1998 are coded as "-9" for this variable.

D. Alternative Designation (alternative; alternative_txt) Categorical Variable

This variable applies to only those cases coded as "Yes" for "Doubt Terrorism Proper?" (above). This variable identifies the most likely categorization of the incident other than terrorism.

- 1= Insurgency/Guerilla Action
- 2= Purely Criminal Act
- 3= Mass Murder
- 4= Internecine Conflict Action

This field is presently only available with incidents occurring after 1997.

E. Part of Multiple Incident (multiple) Categorical Variable

In those cases where several attacks are connected, but where the various actions do not constitute a single incident (either the time of occurrence of incidents or their locations are discontinuous), then “Yes” is selected to denote that the particular attack was part of a “multiple” incident.

- 1 = "Yes" The attack is part of a multiple incident.
- 0 = "No" The attack is not part of a multiple incident.

F. *Situation of Multi-Party Conflict (conflict) Categorical Variable*

When there are multiple groups in conflict, and some of the groups might be committing terrorist acts, it is often difficult to attribute responsibility or to unequivocally discern various non-state actors. In this case, “Yes” is selected.

- 1 = "Yes" The incident took place in the context of a multi-party conflict.
- 0 = "No" The incident did not take place in the context of a multi-party conflict.

VII. ATTACK INFORMATION

A. *Successful Attack (success) Categorical Variable*

Success of a terrorist strike is defined according to the tangible effects of the attack. For example, in a typical successful bombing, the bomb detonates and destroys property and/or kills individuals, whereas an unsuccessful bombing is one in which the bomb is discovered and defused or detonates early and kills the perpetrators. Success is not judged in terms of the larger goals of the perpetrators. For example, a bomb that exploded in a building would be counted as a success even if it did not, for example, succeed in bringing the building down or inducing government repression.

1 = "Yes" The incident was successful.
0 = "No" The incident was not successful.

B. *Suicide Attack (suicide) Categorical Variable*

This variable is coded “Yes” in those cases where there is evidence that the perpetrator did not intend to escape from the attack alive.

1 = "Yes" The incident was a suicide attack.
0 = "No" The incident was not a suicide attack.

C. *Attack Type (attacktype1; attacktype1_txt) Categorical Variable*

Up to three attack types are recorded for each incident. This field captures the general method of attack and often reflects the broad class of tactics used. It consists of the following nine categories, which are defined below:

1= Assassination An act whose primary objective is to kill one or more specific, prominent individuals. Usually carried out on persons of some note, such as high-ranking military officers, government officials, celebrities, etc. Not to include attacks on non-specific members of a targeted group. The killing of a police officer would be an armed assault unless there is reason to believe the attackers singled out a particularly prominent officer for assassination.

2= Armed Assault An attack whose primary objective is to cause physical harm or death directly to human beings by any means other than an explosive.

3= Bombing/Explosion An attack where the primary effects are caused by an energetically unstable material undergoing rapid decomposition (either deflagration or detonation) and releasing a pressure wave that causes physical damage to the surrounding environment. Can include either high or low explosives but does not include a nuclear explosive device that releases energy from fission and/or fusion, or an incendiary device where decomposition takes place at a much slower rate.

4=Hijacking An act whose primary objective is to take control of a vehicle such as an aircraft, boat, bus, etc. for the purpose of diverting it to an unprogrammed destination, obtain payment of a ransom, force the release of prisoners, or some other political objective. Hijackings are distinct from Hostage Taking because the target is a vehicle, regardless of whether there are people/passengers in the vehicle.

5=Hostage Taking (Barricade Incident) An act whose primary objective is to obtain political or other concessions in return for the release of prisoners (hostages). Such attacks are distinguished from kidnapping since the incident occurs and usually plays out at the target location with little or no intention to hold the hostages for an extended period in a separate clandestine location.

6=Hostage Taking (Kidnapping) As for Barricade Incident above, but distinguished by the intention to move and hold the hostages in a clandestine location. Usually in kidnappings the victims are selected beforehand.

7=Facility / Infrastructure Attack An act, excluding the use of an explosive, whose primary objective is to cause damage to a non-human target, such as a building, monument, train, pipeline, etc. Such attacks consist of actions primarily aimed at damaging property, or at causing a diminution in the functioning of a useful system (mass disruption) yet not causing direct harm to people. Such attacks include arson, cyber attacks, and various forms of sabotage. Can include acts that intend to cause harm to people as a result of the harm done to objects (e.g., blowing up a dam so that the ensuing flood will kill residents downstream). Can include acts which aim to harm an installation, yet also cause harm to people incidentally.

8=Unarmed Assault An attack whose primary objective is to cause physical harm or death directly to human beings by any means other than explosive, firearm, incendiary, or sharp instrument (knife, etc.).

9=Unknown The attack type cannot be determined from the available information.

D. Second Attack Type (attacktype2; attacktype2_txt) Categorical Variable – **Coding is same as above.**

E. Third Attack Type (attacktype3; attacktype3_txt) Categorical Variable- – **Coding is same as above.**

VIII. TARGET INFORMATION

Information on up to three targets is recorded for each incident. The target information fields coded for each of the three targets include target type, target entity, name of entity, specific target, and nationality of the target.

A. A. Target Type (targtype1; targtype1_txt) Categorical Variable

The target type field captures the general type of target. It consists of the following 22 categories, which are defined as under:

1=Business - Businesses are defined as individuals or organizations engaged in commercial or mercantile activity as a means of livelihood. Any attack on a business or private citizens patronizing a business such as a restaurant, gas station, music store, bar, café, etc. This includes attacks carried out against corporate offices or employees of firms like mining companies, or oil corporations. Furthermore, includes attacks conducted on business people or corporate officers. Included in this value as well are hospitals and chambers of commerce and cooperatives. It does not include attacks carried out in public or quasi-public areas such as "business district or commercial area", (these attacks are captured under "Private Citizens and Property", see below.)

2=Government (General) - Any attack on a government building; government member, former members, including members of political parties, their convoys, or events sponsored by political parties; political movements; or a government sponsored institution where the attack is expressly carried out to harm the government. This value includes attacks on judges, public attorneys (e.g., prosecutors), courts and court systems, politicians, royalty, head of state, government employees (unless police or military), election-related attacks, intelligence agencies and spies.

3=Police – This value includes attacks on members of the police force or police installations; this includes police boxes, patrols, Headquarters, academies, cars, checkpoints, etc. This includes attacks against jails or prison facilities, or jail or prison staff or guards. Also includes attacks against private security guards and security forces.

4= Military - Includes attacks against army units, patrols, barracks, and convoys, jeeps, etc. Also includes attacks on recruiting sites, and soldiers engaged in internal policing functions such as at checkpoints and in anti-narcotics activities. It excludes attacks against militia and guerrillas, these types of attacks are coded as "Terrorist" see below.

5=Abortion Related - Attacks on abortion clinics, employees, patrons, or security personnel stationed at clinics.

6=Airports & Airlines – An attack that was carried out either against an airplane or against an airport. Attacks against airline employees while on board are also included in this value. It includes attacks conducted against airport business offices and executives. Attacks where airplanes were used to carry out the attack (such as three of the four 9/11 attacks) are not included.

7=Government (Diplomatic) - Attacks carried out against foreign missions, including embassies, consulates, etc. This value includes cultural centers that have diplomatic functions, and attacks against diplomatic staff and their families and property.

8=Educational Institution - Attacks against schools, teachers, or guards protecting school sites. Includes attacks against university professors, teaching staff and school buses. Moreover, includes attacks against religious schools in this value. As noted below in the "Private Citizens and Property" value, the database has several attacks against students. If attacks involving students are not expressly against a school, university or other educational institution or are carried out in an educational setting, they are coded as private citizens and property. This excludes attacks against military schools (attacks on military schools are coded as "Military,").

9=Food or Water Supply - Attacks on food or water supplies or reserves are included in this value.

10=Journalists & Media - Includes, attacks on reporters, news assistants, photographers, publishers, as well as attacks on media headquarters and offices. Attacks on transmission facilities such as antennae or transmission towers are included in this value (while attacks on broadcast infrastructure are coded as "Telecommunications,").

11=Maritime (Includes Ports and Maritime Facilities) - Implies civilian maritime. Includes attacks against fishing ships, oil tankers, ferries, yachts, etc. (Attacks on fishermen are coded as "Private Citizens and Property," see below).

- ***12=NGO*** - Includes attacks on offices and employees of non-governmental organizations (NGOs). NGOs here are defined as primarily large multinational non-governmental organizations. These include the Red Cross and Doctors without Borders. Peacekeepers also belong to this value. This does not include labor unions, social clubs, student groups, and other non-NGO (such cases are coded as "Other".).
- ***13=Other*** - This value includes acts of terrorism committed against targets which do not fit into other categories.
- ***14=Private Citizens & Property*** -This value includes attacks on individuals, the public in general or attacks in public areas including markets, commercial streets, busy intersections and pedestrian malls. This also includes ambiguous cases where the target was a named

individual, or where the target/victim of an attack could be identified by name, age, occupation, gender or nationality. This value also includes ceremonial events, such as weddings and funerals. The database contains a number of attacks against students. If these attacks are not expressly against a school, university or other educational institution or are not carried out in an educational setting, these attacks are coded using this value. Also, includes incidents involving political supporters as private citizens and property, provided that these supporters are not part of a government-sponsored event. Finally, this value includes police informers. This does not include attacks causing civilian casualties in businesses such as restaurants, cafes or movie theaters (these categories are coded as "Business" see above).

- ***15= Religious Figures/ Institutions*** - This value includes attacks on religious leaders, (Imams, priests, bishops, etc.), religious institutions (mosques, churches), religious places or objects (shrines, relics, etc.). This value also includes attacks on organizations that are affiliated with religious entities that are not NGOs, businesses or schools. Attacks on religious pilgrims are considered "Private Citizens and Property;" attacks on missionaries are considered religious figures.
- ***16=Telecommunication*** - This includes attacks on facilities and infrastructure for the transmission of information. More specifically this value includes things like cell phone towers, telephone booths, television transmitters, radio, and microwave towers.
- ***17=Terrorists*** - Terrorists or members of identified terrorist groups are included in this value. Membership is broadly defined and includes informants for terrorist groups, but excludes former terrorists. This value also includes cases involving the targeting of militias and guerillas.
- ***18=Tourists*** - This value includes the targeting of tour buses, tourists, or "tours." Tourists are persons who travel primarily for the purposes of leisure or amusement. Government tourist offices are included in this value. The attack must clearly target tourists, not just an assault on a business or transportation system used by tourists.
- ***19=Transportation (Other than Aviation)*** - Attacks on public transportation systems are included in this value. This can include efforts to assault public buses, minibuses, trains, metro/subways, highways (if the highway itself is the target of the attack), bridges, roads, etc. The database contains a number of attacks on generic terms such as "cars" or "vehicles." These attacks are assumed to be against "Private Citizens and Property" unless shown to be against public transportation systems. In this regard, buses are assumed to be public transportation unless otherwise noted.
- ***20=Unknown*** - The target type cannot be determined from the available information.
- ***21=Utilities*** - This value pertains to facilities for the transmission or generation of energy. For example, power lines, oil pipelines, electrical transformers, high tension lines, gas and electric substations, are all included in this value. This value also includes lampposts or street lights. Attacks on officers, employees or facilities of utility companies excluding the type of faculties above are coded as business.
- ***22=Violent Political Parties*** -This value pertains to entities that are both political parties (and thus, coded as "government" in this coding scheme) and terrorists. It is operationally defined as groups that engage in electoral politics and appear as "Perpetrators" in the database.

B. Target Entity (entity1; entity1_txt) Categorical Variable

The entity field refers to the type of organization or interest group represented by the specific target attacked, and provides an alternate categorization to "Target Type" above.

1 = Diplomat
2 = Police/Military
3 = Other
4 = Unknown
5 = Government
6 = Political Party
7 = Media
8 = Business
9 = Transportation
10= Utilities
11 = Foreign Business
12 = Domestic Business
13 = Transportation
14 = Utilities
15 = Media
16 = Diplomat
17 = Government
18 = International
19 = Other
20 = Police/Military
21 = Political Party
22 = Unknown
23 = Religious Figures/Institutions
24 = Indiscriminate Civilians/Non-Combatants
25 = Religious Figures/Institutions
26 = Indiscriminate Civilians/Non-Combatants

C. Name of Entity (corp1) Text Variable

This is the name of the corporate entity or government agency that was targeted. If no specific entity was

targeted, this field is left blank. If the element targeted is unspecified, "Unknown" is listed.

D. Specific Target (target1) Text Variable

This is the specific person, building, installation, etc., that was targeted and is a part of the entity named above. (For example, if the U.S. Embassy in Country X was attacked the "Name of Entity" would be "U.S. Department of State" and the "Specific Target" would be "U.S. Embassy in Country X.") However, if the target includes multiple victims (e.g., in a kidnapping or assassination), only the first victim's name is recorded in this field, with remaining names recorded in the "Additional Notes" field.

E. Nationality of Target (natlty1; natlty1_txt) Categorical Variable

This is the nationality of the target that was attacked, and is not necessarily the same as the country in which the incident occurred, although in most cases it is. For hijacking incidents, the nationality of the plane is recorded and not that of the passengers. Numeric nationality codes are same as the country codes .

F. Second Target Type (targtype2; targtype2_txt) Categorical Variable – Same as targtype1 above.

G. Second Target Entity (entity2; entity2_txt) Categorical Variable – Same as entity1 above.

H. Name of Second Entity (corp2) Text Variable

Same as "Name of Entity" field.

I. Second Specific Target (target2) Text Variable

Conventions follow "Specific Target" field.

IX. NATIONALITY OF SECOND TARGET (NATLTY2; NATLTY2_TXT) CATEGORICAL VARIABLE

Conventions follow "Nationality of Target" field. For numeric nationality codes, as per the country codes in section V above.

A. Third Target Type (targtype3; targtype3_txt) Categorical Variable

Conventions follow "Target Entity" field.

B. Name of Third Entity (corp3) Text Variable

Conventions follow "Name of Entity" field.

C. Third Specific Target (target3) Text Variable

Conventions follow "Specific Target" field.

D. Nationality of Third Target (natlty3; natlty3_txt) Categorical Variable

Conventions follow "Nationality of Target" field. For numeric nationality codes, please see the country codes in section III-A.

X. PERPETRATOR INFORMATION

Information on up to three perpetrators is recorded for each incident. This includes the perpetrator group name and the perpetrator group sub-name, in addition to the specific motive of the attack and a record of whether or not the attribution of responsibility is unconfirmed.

A. Perpetrator Group Name (gname) Text Variable

This field contains the name of the group that carried out the attack. In order to ensure consistency in the usage of group names for the database, the GTD database uses a standardized list of group names that have been established by project staff to serve as a reference for all subsequent entries.

B. Perpetrator Sub-Group Namen(gsubname) Text Variable

This field contains any additional qualifiers or details about the name of the group that carried out the attack. This includes but is not limited to the name of the specific faction when available.

C. Second Perpetrator Group Name (gname2) Text Variable

This field is used to record the name of the second perpetrator when responsibility for the attack is attributed to more than one perpetrator. Conventions follow "Perpetrator Group" field.

D. Second Perpetrator Sub-Group Name (gsubname2) Text Variable

This field is used to record additional qualifiers or details about the second perpetrator group name when responsibility for the attack is attributed to more than one perpetrator. Conventions follow "Perpetrator Sub-Group Name" field.

E. Third Perpetrator Group Name (gname3) Text Variable

This field is used to record the name of the third perpetrator when responsibility for the attack is attributed to more than two perpetrators. Conventions follow "Perpetrator Group" field.

F. Third Perpetrator Sub-Group Name (gsubname3) Text Variable

This field is used to record additional qualifiers of details about the third perpetrator group name when responsibility for the attack is attributed to more than two perpetrators. Conventions follow "Perpetrator Sub-Group Name" field.

G. Specific Motive (motive) Text Variable

When reports explicitly mention a specific motive for the attack, this motive is recorded in the "Specific Motive" field.

H. Perpetrator Group(s) Suspected/Unconfirmed? (guncertain) Categorical Variable

- "Yes" is used in circumstances where a government official is reported to be expressing a suspicion, or educated guess or other unconfirmed / speculative position regarding the identity of the terrorist group mounting the attack. Cases where credible, non-government analysts identify probable perpetrators receive a "No" in this field.
- Cases where a terrorist group claims responsibility for the attack are recorded as "No" unless the source specifically notes that authorities doubt the veracity of the claim.
- Cases where a government official expresses a definite position on the perpetrator based on intelligence or other information are recorded as "No".
- 1 = "Yes" The perpetrator attribution(s) for the incident are unconfirmed.
- 0 = "No" The perpetrator attribution(s) for the incident are not unconfirmed.

XI. PERPETRATOR STATISTICS

A. Number of Perpetrators (nperps)Numeric Variable

This field indicates the total number of terrorists participating in the incident. (In the instance of multiple perpetrator groups participating in one case, the total number of perpetrators, across groups, is recorded). There are often discrepancies in information on this value.
Where several independent credible sources1 report different numbers of attackers, the value of this variable reflects the number given by the majority of sources, unless there is reason to do otherwise. Where there is no majority figure among independent sources, the database records the lowest preffered perpetrator figure, unless there is clear reason to do otherwise. In cases where the number of perpetrators is stated vaguely, for example "…at least 11 attackers", then the lowest possible number is recorded, in this example, "11." "-99" or "Unknown" appears when the number of perpetrators is not reported.

B. Number of Perpetrators Captured (nperpcap) Numeric Variable

This field records the number of perpetrators taken into custody.

- "-99" or "Unknown" appears when there is evidence of captured, but the number is not reported.
- Divergent reports on the number of perpetrators captured are dealt with in same manner used for the Number of Perpetrators variable described above.

XII. PERPETRATOR CLAIM OF RESPONSIBILITY

A. A. Claim of Responsibility?(claimed) Categorical Variable

This field is used to indicate whether a group or person(s) claimed responsibility for the attack. If marked "Yes", it indicates that a person or a group did in fact claim responsibility. When there are multiple perpetrator groups involved, this field refers to the First Perpetrator Group (separate fields for the Second and Third groups follow below).

- 1 = "Yes" A group or person claimed responsibility for the attack.
- 0 = "No" No claim of responsibility was made.
- -9 = "Unknown" It is unknown whether or not a claim of responsibility was made.

B. Mode for Claim of Responsibility (claimmode; claimmode_txt) Categorical Variable

This records one of 10 modes used by claimants to claim responsibility and might be useful to verify authenticity, track trends in behavior, etc. If greater detail exists (for instance, a particularly novel or strange mode is used) this information is captured in the "Additional Notes" field.
Mode Values:
1 = Letter
2 = Call (post-incident)
3 = Call (pre-incident)
4 = E-mail
5 = Note left at scene
6 = Video
7 = Posted to website, blog, etc.
8 = Personal claim
9 = Other
10 = Unknown

C. Claim Confirmed? (claimconf) Categorical Variable

"Yes" or "No", indicate whether or not the claim is confirmed. "Unknown" appears if this information is not available.

D. Second Group Claim of Responsibility? (claim2) Categorical Variable

1 = "Yes" A group or person claimed responsibility for the attack.
0 = "No" No claim of responsibility was made.
-9 = "Unknown" It is unknown whether or not a claim of responsibility was made.
Conventions follow "Claim of Responsibility" field.

E. Mode for Second Group Claim of Responsibility

(claimmode2; claimmode2_txt) Categorical Variable

Conventions follow "Mode for Claim of Responsibility" field.

F. Second Group Claim of Responsibility Confirmed? (claimconf2) Categorical Variable

Conventions follow "Claim of Responsibility Confirmed?" field.

G. Third Group Claim of Responsibility? (claim3) Categorical Variable

Conventions follow "Claim of Responsibility" field.

H. Mode for Third Group Claim of Responsibility (claimmode3; claimmode3_txt) Categorical Variable

Conventions follow "Mode for Claim of Responsibility" field.

I. Third Group Claim of Responsibility Confirmed? (claimconf3) Categorical Variable

Conventions follow "Claim of Responsibility Confirmed?" field.

J. Competing Claims of Responsibility? (compclaim) Categorical Variable

This field is used to indicate whether more than one group claimed separate responsibility for the attack. If marked "Yes", it indicates that the groups entered in conjunction with the case each claimed responsibility for the attack (i.e., they did not work together, but each independently tried to claim credit for the attack).

- 1 = "Yes" There are competing claims of responsibility for the attack.
- 0 = "No" There are not competing claims of responsibility for the attack.
- -9 = "Unknown" It is unknown whether or not the claim of responsibility is confirmed.

XIII. WEAPON INFORMATION

Information on up to four types and sub-types of the weapons used in an attack are recorded for each case, in addition to any information on specific weapon details reported.

A. Weapon Type (weaptype1; weaptype1_txt) Categorical Variable

This field records the general type of weapon used in the incident. It consists of the following 13 categories:

1 = Biological
2 = Chemical
3 = Radiological
4 = Nuclear
5 = Firearms
6 = Explosives/Bombs/Dynamite
7 = Fake Weapons
8 = Incendiary
9 = Melee
10 = Vehicle (not to include vehicle-borne explosives, i.e., car or truck bombs)
11 = Sabotage Equipment
12 = Other
13 = Unknown

B. Weapon Sub-type (weapsubtype1; weapsubtype1_txt) Categorical Variable

This field records a more specific value for most of the Weapon Types identified immediately above.

Values for Weapon Type and corresponding Sub-type

- Biological [no corresponding weapon sub-types]
- Chemical
  1 = Poisoning
- Radiological [no corresponding weapon sub-types]
- Nuclear [no corresponding weapon sub-types]
- Firearms
  2 = Automatic Weapon
  3 = Handgun
  4 = Rifle/Shotgun (non-automatic)
  5 = Unknown Gun Type
  6 = Other Gun Type
- Explosives/Bombs/Dynamite
  7 = Grenade
  8 = Land Mine
  9 = Letter Bomb
  10 = Pressure Trigger
  11 = Projectile (rockets, mortars, RPGs, etc.)
  12 = Remote Trigger
  13 = Suicide (carried bodily by human being)
  14 = Time Fuse
  15 = Vehicle
  16 = Unknown Explosive Type
  17 = Other Explosive Type
- Fake Weapons [no corresponding weapon sub-types]
- Incendiary
- Melee
  18 = Arson/Fire
  19 = Flame Thrower
  20 = Gasoline or Alcohol
  21 = Blunt Object
  22 = Hands, Feet, Fists
  23 = Knife
  24 = Rope or Other Strangling Device
  25 = Sharp Object Other Than Knife
  26 = Suffocation
- Vehicle (not to include vehicle-borne explosives, i.e., car or truck bombs) [no corresponding weapon sub-types]
- Sabotage Equipment [no corresponding weapon sub-types]
- Other [no corresponding weapon sub-types]
- Unknown [no corresponding weapon sub-types]

C. Second Weapon Type (weaptype2; weaptype2_txt) Categorical Variable

Conventions follow "Weapon Type" field.

D. Second Weapon Sub-Type (weapsubtype2; weapsubtype2_txt) Categorical Variable

Conventions follow "Weapon Sub-Type" field.

E. Third Weapon Type (weaptype3; weaptype3_txt) Categorical Variable

Conventions follow "Weapon Type" field.

F. Third Weapon Sub-Type (weapsubtype3; weapsubtype3_txt) Categorical Variable

Conventions follow "Weapon Sub-Type" field.

G. Fourth Weapon Type (weaptype4; weaptype4_txt) Categorical Variable

Conventions follow "Weapon Type" field.

H. Fourth Weapon Sub-Type (weapsubtype4; weapsubtype4_txt) Categorical Variable

Conventions follow "Weapon Sub-Type" field.

I. Weapon Details (weapdetail) Text Variable

This field notes any pertinent information on the type of weapon(s) used in the incident. Such notes could include the novel use or means of concealing a weapon, specific weapon models, interesting details of the weapons' origins, etc.

XIV. CASUALTY INFORMATION

If several cases are linked together, the open-source reports sometimes list the number of casualties cumulatively. In such cases the preservation of statistical accuracy is preserved by the GTD by evenly distributing casualties across the linked incidents.

A. Total Number of Fatalities (nkill) Numeric Variable

- This field stores the number of total confirmed fatalities for the incident. The number includes all victims and attackers who died as a direct result of the incident.
- Where there is evidence of fatalities, but the number is not reported, "-99"or "Unknown" is the value given to this field.
- Where several independent sources report different numbers of casualties, the database will usually reflect the number given by the most recent source, unless there is reason to do otherwise. Where there are several "most recent" sources published around the same time, then the majority figure will be used. Where there is no majority figure among independent sources, the database will record the lowest proffered fatality figure, unless there is clear reason to do otherwise.

B. Number of U.S. Fatalities (nkillus) Numeric Variable

Limited to only U.S. fatalities, this field follows the conventions of "Total Number of Fatalities" above.

C. Number of Perpetrator Fatalities (nkillter)Numeric Variable

Limited to only perpetrator fatalities, this field follows the conventions of "Total Number of Fatalities" field.

D. Total Number of Injured (nwound) Numeric Variable

This field records the number of confirmed non-fatal injuries. Conventions follow the "Total Number of Fatalities" field.

E. Number of U.S. Injured (nwoundus) Numeric Variable

Conventions follow the "Number of U.S. Fatalities" field.

F. Number of Perpetrators Injured (nwoundte) Numeric Variable

Conventions follow the "Number of Perpetrator Fatalities" field.

XV. CONSEQUENCES

A. Property Damage? (property) Categorical Variable

"Yes" appears if there is evidence of property damage during the incident.

- 1 = "Yes" The incident resulted in property damage.
- 0 = "No" The incident did not result in property damage.
- -9 = "Unknown" It is unknown whether or not the incident resulted in property damage

B. Extent of Property Damage (propextent; propextent_txt) Categorical Variable

If "Property Damage?" is "Yes" then one of four categories describe the extent of the property damage:
1 = Catastrophic (likely > $1 billion)
2 = Major (likely > $1 million but < $1 billion)
3 = Minor (likely < $1 million)
4 = Unknown

C. Value of Property Damage (in U.S. $) (propvalue) Numeric Variable

If "Property Damage?" is "Yes" then the exact U.S. dollar amount (at the time of the incident) of total damages is listed. If no dollar figure is reported, the field is blank. That is, a blank field here does not indicate that there was no property damage but, rather, that no precise

estimate of the value was available. The value of damages only includes direct economic effects of the incident (i.e. cost of buildings, etc.) and not indirect economic costs (longer term effects on the company, industry, tourism, etc.). Protocols for recording inconsistent numbers, etc., listed above are followed (see, for example, "Number of Perpetrators").

D. Property Damage Comments (propcomment) Text Variable

If "Property Damage?" is "Yes" then non-monetary or imprecise measures of damage may be described in this field.

XVI. HOSTAGE / KIDNAPPING ADDITIONAL INFORMATION

A. Hostages or Kidnapping Victims? (ishostkid) Categorical Variable

This field records whether or not the victims were taken hostage or kidnapped.

- 1 = "Yes" The victims were taken hostage or kidnapped.
- 0 = "No" The victims were not taken hostage or kidnapped.
- -9 = "Unknown" It is unknown whether or not the victims were taken hostage or kidnapped.

B. Total Number of Hostages/ Kidnapping Victims (nhostkid) Numeric Variable

This field records the total number of hostages or kidnapping victims. As with the number of perpetrators, where several independent sources report different numbers of hostages, the GTD reflects the number given by the majority of sources, unless there is reason to do otherwise. Where there is no majority figure among independent sources, the database will record the lowest proffered hostage figure, unless there is clear reason to do otherwise. In cases where the number of hostages or kidnapping victims is stated vaguely, for example, "…at least 11 hostages", then the lowest possible number will be recorded, in this example "11." If the number of hostages is unknown or unidentified, this field records "-99" or "Unknown."

C. Number of U.S. Hostages/ Kidnapping Victims (nhostkidus) Numeric Variable

Conventions follow the "Total Number of Hostages/ Kidnapping Victims" field, but only include U.S. hostage/kidnapping victims.

D. Hours of Kidnapping / Hostage Incident (nhours) Numeric Variable

- If the "Attack Type" is "Hostage Taking (Kidnapping)," "Hostage Taking (Barricade Incident)," or "Hijacking" then the duration of the incident is recorded either in this field or in the next field.
- If the incident lasted for less than 24 hours, this field records the number of hours.
- If the incident lasts for more than 24 hours (i.e., at least one day), then the number of days is recorded in the next field.

E. Days of Kidnapping / Hostage Incident (ndays) Numeric Variable

If the "Attack Type" is "Hostage Taking (Kidnapping)," "Hostage Taking (Barricade Incident)," or "Hijacking" and if the duration of the kidnapping / hostage incident last for more than 24 hours, this field records the duration of the incident in days. If information on hours and days is provided, the figure is rounded to the nearest day.

F. Country That Kidnappers/Hijackers Diverted To (divert) Text Variable

If the "Attack Type" is "Hostage Taking (Kidnapping)" or "Hijacking" then this field lists the country that the hijackers diverted the vehicle to. If the hijackers did not divert the vehicle to another country, this field is blank.

G. Country of Kidnapping/Hijacking Resolution (kidhijcountry) Text Variable

If the "Attack Type" is "Hostage Taking (Kidnapping)" or "Hijacking" then this field lists the country in which the incident was resolved or ended. If the incident was not resolved in another country, this field is blank.

H. Ransom Demanded? (ransom) Categorical Variable

"Yes" is recorded if the incident involved the demand of some form of ransom.

- 1 = "Yes" The incident involved a demand of ransom.
- 0 = "No" The incident did not involve a demand of ransom.
- -9 = "Unknown" It is unknown whether or not the incident involved a demand of ransom.

I. Total Ransom Amount Demanded (ransomamt) Numeric Variable

If a ransom was demanded then the amount of ransom demanded is listed in U.S. dollars. If a ransom was demanded but the monetary figure was unknown then this field is recorded with "-99" or "Unknown."

J. Ransom Amount Demanded from U.S. Sources (ransomamtus) Numeric Variable

If a ransom was demanded from U.S. sources then this figure is listed in U.S. dollars. If a ransom was demanded from U.S. sources but the monetary figure was unknown then this field is recorded with "-99" or "Unknown."

K. Total Ransom Amount Paid (ransompaid) Numeric Variable

If a ransom amount was paid then this figure is listed in U.S. dollars. If a ransom was paid but the monetary

figure was unspecified then this field is recorded with "-99" or "Unknown."

L. Ransom Amount Paid By U.S. Sources (ransompaidus) Numeric Variable

If a ransom amount was paid by U.S. sources then this figure is listed in U.S. dollars. If a ransom was paid by U.S. sources but the monetary figure was unspecified then this field is recorded with "-99" or "Unknown."

M. Ransom Notes (ransomnote) Text Variable

If a ransom was demanded this field may be used to record any specific comments relating to the ransom not captured in other fields.

N. Kidnapping/Hostage Outcome (hostkidoutcome; hostkidoutcome_txt) Categorical Variable

If the "Attack Type" is "Hostage Taking (Kidnapping)" then this field applies. The seven values for this field are:

1 = Attempted Rescue
2 = Hostage(s) released by perpetrators
3 = Hostage(s) escaped (not during rescue attempt)
4 = Hostage(s) killed (not during rescue attempt)
5 = Successful Rescue
6 = Combination
7 = Unknown

If the hostages suffered a variety of the above fates, "Combination" is selected. Further details about the fate of hostages may be recorded in the "Additional Notes" field.

O. Number Released/Escaped/Rescued (nreleased) Numeric Variable

If the "Attack Type" is "Hostage Taking (Kidnapping)" then this field will apply. This field records the number of hostages who survived the incident. All previous protocols for recording numbers apply, including using "-99" for "Unknown."

As with the total number of kidnapping victims, where several independent sources report different numbers of hostages, the database will reflect the number given by the majority of sources, unless there is reason to do otherwise. Where there is no majority figure among independent sources, the database will record the lowest proffered hostage released/escaped/rescued figure, unless there is clear reason to do otherwise. In cases where the number of hostages released/escaped/rescued is stated vaguely, for example "…at least 11 hostages were released", then the lowest possible number will be recorded, in this example "11". If the number of hostages released/escaped/rescued is unknown or unidentified, this is recorded as "Unknown".

XVII. ADDITIONAL INFORMATION

A. (addnotes) Text Variable

This field is used to capture the following information:

- Additional information that could not be captured in any of the above fields, such as details about hostage conditions or additional countries hijacked vehicles were diverted to.
- Supplemental important information not specific to the particular attack, such as multiple attacks in the same area or by the same perpetrator.
- Uncertainties about the data (such as differing reports of casualty numbers or perpetrators responsible).
- Unusual factors, such as a shift in tactics, the reappearance of an organization, the emergence of a new organization, an attack carried out on a historical date, or an escalation of a violent campaign.
- The fate (legal, health, or otherwise) of either victims or perpetrators where this is mentioned in GTD source documents.

B. In addition, the instructions for several fields listed above have specific indications for placing additional information in this "Additional Notes" field, as needed:

- Specific Target - If the Target is multiple victims (e.g., in a kidnapping or assassination), only the first name is recorded in the "Specific Target" field, with remaining names recorded in the "Additional Notes" field.
- Perpetrator Individual(s)' Name(s) - Names of individuals identified as planners, bomb-makers, etc., who are indirectly involved in an attack, may recorded in the "Additional Notes" field.
- Mode for Claim of Responsibility - If greater detail is needed than provided for the "Mode for Claim of Responsibility" field (for instance, a particularly novel or strange mode is used) this information may be captured in the "Additional Notes" field.
- Kidnapping/Hostage Outcome - If greater detail is available than the Kidnapping/Hostage Outcome field allows, then further details about the fate of hostages/kidnapped may be recorded in the "Additional Notes" field.

XVIII. SOURCE INFORMATION

A. First Source Citation (scite1) Text Variable

This field cites the first source used to compile information on the specific incident.

B. Second Source Citation (scite2) Text Variable

This field cites the first source used to compile information on the specific incident.

C. Third Source Citation (scite3) Text Variable

This field cites the first source used to compile information on the specific incident.

XIX. DATA MINING TECHNIQUES CONSIDERED

A. Traditional data mining techniques such as association analysis, classification and prediction, cluster analysis, and outlier analysis identify patterns in structured data [5] and hence these techniques are applicable in this scenario as well. Various forms of data mining especially, crime / terror data mining raises privacy concerns [6]. Nevertheless, researchers are developing various automated data mining techniques for both local law enforcement and national security applications.

B. Clustering techniques can group the above data items into classes with similar characteristics to maximize or minimize intraclass similarity—for example, to identify perpetrators who claimed to have carried out the incident in similar ways or distinguish among groups belonging to different terrorist outfits. These techniques do not have a set of pre-defined classes for assigning items. However since the data under question is a Global Data a great deal of presummarisation is involved and hence it would be more apt to have multi dimensional cubes generated and explore/ evolve data mining techniques suited for OLAP. So we do not propose direct application of this method.

C. To predict terrorist activity trends, classification can reduce the time required to identify the perpetrators. However, the technique requires a predefined classification scheme[13]. We can evolve a scheme but then classification also requires reasonably complete training and testing data because a high degree of missing data would limit prediction accuracy. The GTD data is sparse in this regard and hence this approach prima facie does not appear promising and hence we do not propose to use this

D. With association rule mining we can discover frequently occurring item sets in the GTD database and present the patterns as rules. We can apply this technique to the incidents and perpetrators to help detect potential future incidents of similar nature [12].

E. Similar to association rule mining, we shall also try sequential pattern mining to find frequently occurring sequences of incidents that occurred at different times. This approach can identify attack patterns among time-stamped data. Showing hidden patterns benefits terror incidence analysis, but to obtain meaningful results GTD which is a feature rich data has to be summarised and highly structured for which we propose to constuct relevant OLAP cubes for analysis and data mining.

F. On these OLAP cubes we shall also be trying out deviation detection / outlier analysis by appyling our own uses specific measures in the form of outlier score functions to study incident data that differs markedly from the rest of the data.

XX. ACKNOWLEDGMENT

A. The authors acknowledge the efforts being made by researchers and Journalists all over the world who are compiling the Global Terror Data Base. The Text, Introduction to Data Mining with Case Studies By G.K. Gupta [1] presents a rich collection of Data Mining Case Studies which has motivated us to explore application of suitable data mining methods to GTD.

XXI. REFERENCES

[1] *Introduction to Data Mining with Case Studies By G.K. Gupta*

[2] *http://www.start.umd.edu/gtd/contact/*

[3] *W. Chang et al., "An International Perspective on Fighting Cybercrime," Proc. 1st NSF/NIJ Symp. Intelligence and Security Informatics, LNCS 2665, Springer-Verlag, 2003.*

[4] *http://www.start.umd.edu/gtd/downloads/Codebook.pdf*

[5] *J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.*

[6] *H. Kargupta, K. Liu, and J. Ryan, "Privacy-Sensitive Distributed Data Mining from Multi-Party Data," Proc. 1st NSF/NIJ Symp. Intelligence and Security Informatics, LNCS 2665, Springer-Verlag, 2003.*

[7] *M.Chau, J.J. Xu, and H. Chen, "Extracting Meaningful Entities from Police Narrative Reports, Proc. Nat'l Conf. Digital Government Research, Digital Government Research Center, 2002.*

[8] *A. Gray, P. Sallis, and S. MacDonell, "Software Forensics: Extending Authorship Analysis Techniques to Computer Programs," Proc. 3rd Biannual Conf.*

[9] *Int'l Assoc. Forensic Linguistics, Int'l Assoc. Foren sic Linguistics, 1997.*

[10] *R.V. Hauck et al., "Using Coplink to Analyze Criminal-Justice Data," Computer, Mar. 2002.*

[11] *T. Senator et al., "The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions," AI Magazine, vol.16, no. 4, 1995.*

[12] *.W. Lee, S.J. Stolfo, and W. Mok, "A Data Mining Framework for Building Intrusion Detection Models," Proc. 1999 IEEE Symp. Security and Privacy, IEEE CS Press, 1999.*

[13] *O. de Vel et al., "Mining E-Mail Content for Author Identification Forensics," SIGMOD Record, vol. 30, no. 4, 2001.*

[14] *.G. Wang, H. Chen, and H. Atabakhsh, "Automatically Detecting Deceptive Criminal Identities," Comm. ACM, Mar. 2004.*

Unsteady Hydromagnetic Flow Of Viscoelastic Fluid Down An Open Inclined Channel

S.Sreekanth1,, R.Saravana2, S.Venkataramana3, R.Hemadri Reddy4
1Department of Mathematics, Sreenivasa Institute of Technology and Management Studies, Chittoor 517127, A.P. India.
2 & 3Department of Mathematics, Sri Venkateswara University, Tirupati, A.P. India.
4School of Advanced Sciences, VIT University, Vellore – 632 014, T.N. India.

Abstract-In this paper, we study the unsteady hydromagnetic flow of a Walter's fluid (Model B') down an open inclined channel of width 2a and depth d under gravity, the walls of the channel being normal to the surface of the bottom under the influence of a uniform transverse magnetic field. A uniform tangential stress is applied at the free surface in the direction of flow. We have evaluated the velocity distribution by using Laplace transform and finite Fourier Sine transform technique. The velocity distribution has been obtained taking different form of time dependent pressure gradient g(t), viz., i) constant ii) exponential decreasing function of time and iii) Cosine function of time. The effects of magnetic parameter M, Reynolds number R and the viscoelastic parameter K are discussed on the velocity distribution in three different cases.

Key words: Walter's B' fluid, open inclined channel, Laplace transform and finite Fourier Sine transform technique.

1. INTRODUCTION

A flowing liquid is said to have a free surface when the upper part of the bounding surface of the liquid is in contact with the overlying atmosphere, rather than with a solid, as would be the case of the flow were in a pipe completely full of the liquid. A flow with a free surface proceeding in a natural or artificial channel or conduit is called an open-channels flow. Open-channels may be divided into two types namely (i) Natural channels and (ii) Artificial channels. Natural channels range inform from the bounder-strewn bed of a mountain torrent to the relatively uniform channel of a large river. Artificial channels are man-made and are constructed in many forms. A pipe in which water flows with a free surface and flume of rectangular cross section constructed of sheet iron are two kinds of artificial channels. Other types are canals excavated in earth or blasted in rock, which either are left unlined or lined with smooth concrete or another suitable material. Unlined or lined tunnels bored through rock may also contain free surface flows.

The flow of a liquid in an open inclined channel with a free surface has a wide application in the designs of drainage, irrigation canals, flood discharge channels and coating to paper rolls etc. Hence the flow of a liquid in an open inclined channel with a free surface under gravity has long been studied experimentally and several interesting empirical results have been reported by many investigators [3, 6, 7, 10, 11, 14]. The steady laminar flow of a viscous fluid flowing down an open inclined channel has been discussed by Satyaprakash [13], Gupta *et al* [4] have studied the flow of a viscous fluid through a porous medium down an open inclined channel. Venkataramana and Bathaiah [18] have studied the flow of a hydromagnetic viscous fluid down an open inclined channel with naturally permeable bed under the influence of a uniform transverse magnetic field. Unsteady laminar flow of an incompressible viscous fluid between porous, parallel flat plates has been investigated by Singh [12], taking (i) both plates are at rest and (ii) Generalized plane Coutte flow. The free surface was exposed to atmospheric pressure and bottom was taken as impermeable. Bakhmeteff [1], Henderson [5] and Chow [2] have discussed many types of open channel flows. Recently, many authors [8, 9 and 14] have studied the flow of Walter's B' fluid.

The subject of Rheology is of great technological importance in many branches of industry. The problem arises of designing apparatus to transport or to process substances which cannot

be governed by the classical stress-strain velocity relations. Example of such substances and process are many, the extrusion of plastics, in the manufacture of rayon, nylon or other textile fibres, viscoelastic effects transported or forced through spinnerts and in the manufacture of lubricating grease and rubber.

Non-Newtonian fluids have wide importance in the present day technology and industries; the Walter's fluid is one of such fluid. The model of Walter's B fluid is chosen for our study as it involves non-Newtonian parameter. The Cauchy stress tensor *T* in such a fluid is related to the motion in the following manner

$$T = -PI + 2\eta_0 e - 2K_0 \frac{\delta e}{\delta t} \qquad (1.1)$$

In this equation *P* is the pressure, *I* is the Identity tensor and the rate of strain tensor e is defined by

$$2e = \nabla v + (\nabla v)^T \qquad (1.2)$$

where *v* is the velocity vector, ∇ is the gradient operator and $\frac{\delta}{\delta t}$ denotes the convicted differentiation of a tensor quantity in relation to the material in motion. The convicted differentiation of the rate of strain tensor is given by

$$\frac{\delta e}{\delta t} = \frac{\partial e}{\partial t} + v \cdot \nabla e - e \cdot \nabla v - (\nabla v)^T \cdot e \qquad (1.3)$$

Here η_0 and K_0 are, respectively, the limiting viscosity at small rate of shear and the short memory coefficient which are defined through

$$\eta_0 = \int_0^\infty N(\tau)\, d\tau \qquad (1.4)$$

and

$$K_0 = \int_0^\infty \tau N(\tau)\, d\tau \qquad (1.5)$$

$N(\tau)$ being the relaxation spectrum as introduced by Walter's [19, 20]. This idealized model is a valid approximation of Walter's fluid (model B′) taking very short memory into account, so that terms involving

$$\int_0^\infty \tau^n N(\tau)\, d\tau, \qquad n \geq 2 \qquad (1.6)$$

have been neglected.

In addition to equation (1.1), the equation of motion and continuity are

$$\nabla \cdot e = 0 \qquad (1.7)$$

$$\rho(v \cdot \nabla v) = \nabla \cdot T \qquad (1.8)$$

In this paper, we study the unsteady hydromagnetic flow of a Walter's fluid (model B) down an open inclined channel under gravity of width 2a and depth d, the walls of the channel being normal to the surface of the bottom, under the influence of uniform transverse magnetic field. A uniform tangential stress is applied at the free surface in the direction of flow. We have evaluated the velocity distribution by using Laplace Transform and Finite Fourier Sine Transform techniques. Here it is assumed that (i) the fluid flows in the steady state for $t \leq 0$, (ii) Unsteady state occurs at $t > 0$ and (iii) the unsteady motion is influenced by time dependent pressure gradient. The velocity distribution has been obtained in some particular cases i.e. when (i) $g(t) = c^*$ (ii) $g(t) = c^* e^{-bt}$ and (iii) $g(t) = c^* \cos bt$, where b and c^* are constants. The effects of magnetic parameter M, Reynolds number R and viscoelastic parameter K are investigated on the velocity distribution in three different cases.

2. FORMULATION AND SOLUTION OF THE PROBLEM

We consider the unsteady Hydromagnetic flow of a Walter's fluid (model B') down an open inclined channel of width 2a and depth d under gravity, the walls of the channel being normal to the surface of the bottom under the influence of uniform transverse magnetic field. A uniform tangential stress S is applied at the free surface. The bottom of the channel is taken at angle $\beta\left(0<\beta\le\pi/2\right)$ with the horizontal. The x-axis is taken along central line in the direction of the flow at the free surface, y-axis along the depth of the channel and z-axis along width of the channel. A uniform magnetic field of intensity H_o is introduced in y-direction. Therefore the velocity and the magnetic field are given by $\overline{q}=\left(u,0,0\right)$ and $\overline{H}=\left(0,H_0,0\right)$. The fluid being slightly conducting, the magnetic Reynolds number is much less than unity, so that the induced magnetic field can be neglected in comparison with the applied magnetic field (Sparrow and Cess [15]). In the absence of any input electric field the equations of continuity and motion of the unsteady hydromagnetic Walter's fluid (model B') flowing down an open inclined channel at t>0 are given by

$$\frac{\partial u}{\partial x}=0 \tag{2.1}$$

$$\rho\frac{\partial u}{\partial t}=-\frac{\partial p}{\partial X}+\rho g\sin\beta+\mu\left(\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2}\right)-K_0\left(\frac{\partial^3 u}{\partial t\partial y^2}+\frac{\partial^3 u}{\partial t\partial z^2}\right)-\sigma{\mu_e}^2{H_0}^2u \tag{2.2}$$

$$0=-\frac{\partial p}{\partial y}+\rho g\cos\beta \tag{2.3}$$

$$0=-\frac{\partial p}{\partial z} \tag{2.4}$$

Where ρ = density of the fluid

g = acceleration due to gravity

p = pressure

μ = coefficient of viscosity

σ = electrical conductivity of the fluid

μ_e = magnetic permeability

K_0 = viscoelastic parameter

The boundary conditions are

$$\left.\begin{array}{l} t \le 0;\ u = u_0 \\ t > 0;\ z = \pm a, u = o \\ y = 0,\ \mu \dfrac{\partial u}{\partial y} = S \\ y = d\ \ u = 0 \end{array}\right\} \qquad (2.5)$$

where u_0 is the initial velocity.

we introduce the following non-dimensional quantities

$$\left.\begin{array}{l} u^* = u/U, \qquad x^* = x/d, \qquad y^* = y/d, \qquad z^* = z/d, \qquad t^* = t\upsilon/d^2 \\ p^* = p/\rho U^2, \quad K^* = K_0/\rho d^2, S^* = S/\rho U^2 \end{array}\right\} \qquad (2.6)$$

where the depth of the channel d is the characteristic length and the mean flow velocity U is the characteristic velocity.

In view of the equation (2.6), the equations (2.1) to (2.3) reduce to (dropping the superscript *)

$$\frac{\partial u}{\partial t} = -R\frac{\partial P}{\partial X} + \frac{R}{F}\sin\beta + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} - K\left(\frac{\partial^3 u}{\partial t \partial y^2} + \frac{\partial^3 u}{\partial t \partial z^2}\right) - Mu \qquad (2.7)$$

$$\frac{\partial u}{\partial x} = 0 \qquad (2.8)$$

Where $M = \sigma\mu_e^{\ 2} H_o^{\ 2} d^2/\rho\upsilon$ Magnetic parameter

$R = Ud/\upsilon$ Reynolds number

$F = U^2/gd$ Froude number

$K = K_0/\rho d^2$ viscoelastic parameter

The non-dimensional boundary conditions are

$$\left.\begin{array}{l} t \le 0;\ u = u_0 \\ t > 0;\ z = l\left(= \pm a/d\right), u = o \\ y = 0\ \partial u/\partial y = SR \\ y = 1\ u = 0 \end{array}\right\} \qquad (2.9)$$

3. METHOD OF SOLUTION

Assuming

$$-R\frac{\partial p}{\partial X}+\frac{R}{F}\sin\beta \qquad \begin{matrix} = g(t) \;\; at \;\; t>0 \\ = P \qquad at \;\; t\le 0 \end{matrix} \tag{3.1}$$

substituting z = (2lf/π)-1 in equation (2.7) reduces to

$$\frac{\partial u}{\partial t}=g(t)+\frac{\partial^2 u}{\partial y^2}+\frac{\pi^2}{4l^2}\frac{\partial^2 u}{\partial f^2}-K\left(\frac{\partial^2 u}{\partial t\partial y^2}+\frac{\pi^2}{4l^2}\frac{\partial^3 u}{\partial t\partial f^2}\right)-Mu \tag{3.2}$$

and the boundary conditions are reduced to

$$\left.\begin{array}{l} t\le 0;\; u=u_0 \\ t>0;\; f=0,\pi;\; u=o \\ y=0,\; \dfrac{\partial u}{\partial y}=SR \\ y=1,\; u=0 \end{array}\right\} \tag{3.3}$$

Now, since u_0 is the initial velocity i.e. at $t\le 0$, therefore taking $g(t)=P$ in equation (3.2)

$$u_0=\frac{2}{\pi}\sum_{n=1}^{\infty}\left(\frac{1-\cos n\pi}{n}\right)\left[\frac{P}{C^2}\left(1-\frac{\cosh Cy}{\cosh C}\right)-\frac{SR}{C}\frac{\sinh C(1-y)}{\cosh C}\right]\sin nf \tag{3.4}$$

Where

$$C^2=Q^2+M, \qquad Q=\frac{n\pi}{2l}$$

Now to solve equation (3.2) we take Laplace transform of equation (3.2) with respect to t (Sneddon [16])

$$\bar{u}(y,f,s)=\int_0^{\infty}u(y,f,s)e^{-st}dt, \qquad s>0 \tag{3.5}$$

we get

$$\frac{\partial^2\bar{u}}{\partial y^2}+\frac{\pi^2\partial^2\bar{u}}{4l^2\partial f^2}-\frac{(M+s)}{1-Ks}\bar{u}=\frac{1}{1-Ks}\left(\frac{pKM}{C^2}\frac{\cosh Cy}{\cosh C}+\frac{PKQ^2}{C^2}+\frac{SRMK}{C}\frac{\sinh C(1-y)}{\cosh C}-\bar{g}(s)-u_0\right)$$

(3.6)

Where $\bar{g}(s) = \int_0^\infty g(t) e^{-st} dt$

On taking the finite Fourier sine transform the equation (3.6) with respect to f (Sneddon [16])

$$\bar{u}^*(y, N, s) = \int_0^\pi \bar{u}(y, f, s) \sin Nf df \tag{3.7}$$

We get

$$\frac{\partial^2 \bar{u}^*}{\partial y^2} - H^2 \bar{u}^* = \left(\frac{1-\cos n\pi}{N(1-Ks)} \right) \left(\frac{PKQ^2}{C^2} - \bar{g}(s) - \frac{P}{C^2} + \frac{P(1+MK)}{C^2} \frac{\cosh Cy}{\cosh C} + \frac{SR(1+MK)}{C} \frac{\sinh Cy}{\cosh C} \right)$$

(3.8)

Where $H^2 = Q^2 + \frac{s+M}{1-Ks}$

Now, the boundary conditions are reduced to

$$\left.\begin{array}{l} y = 0, \quad \dfrac{\partial \bar{u}^*}{\partial y} = \dfrac{SR(1-\cos N\pi)}{SN} \\ y = 1, \quad \bar{u}^* = 0 \end{array}\right\} \tag{3.9}$$

Integrating equation (3.8) under the boundary conditions (3.9), we get

$$\bar{u}^* = \left(\frac{1-\cos N\pi}{N} \right) \left\{ \begin{array}{l} \dfrac{PKQ^2}{C^2 H^2 (1-Ks)} \left(\dfrac{\cosh Hy}{\cosh H} - 1 \right) + \dfrac{g(\bar{s})}{H^2(1-Ks)} \left(\dfrac{\cosh Hy}{\cosh H} - 1 \right) - \dfrac{P}{H^2 C^2 (1-Ks)} \left(\dfrac{\cosh Hy}{\cosh H} - 1 \right) \\ \quad + \dfrac{P}{SC^2} \left(\dfrac{CoshHy}{\cosh H} - \dfrac{\cosh Cy}{\cosh C} - \dfrac{SR}{SC} \dfrac{\sinh C(1-y)}{\cosh C} \right) \end{array} \right\}$$

(3.10)

Now, inverting the finite Fourier sine transform as given by (Sneddon [16])

$$\bar{u}(y, f, s) = \frac{2}{\pi} \sum_{N=1}^{\infty} \bar{u}^*(y, N, s) \sin Nf$$

In equation (3.10) we get

$$\bar{u}(y,f,s)=\frac{2}{\mathrm{p}}\sum_{N=1}^{\infty}\left(\frac{1-\cos N\mathrm{p}}{N}\right)\left[\frac{PKQ^2}{C^2H^2(1-Ks)}\left(\frac{\cosh Hy}{\cosh H}-1\right)-\frac{g(s)}{H^2(1-Ks)}\left(\frac{\cosh Hy}{\cosh H}-1\right)\right.$$
$$\left.-\frac{P}{H^2C^2(1-Ks)}\left(\frac{\cos Hy}{\cosh H}-1\right)+\frac{P}{SC^2}\left(\frac{\cosh Hy}{\cosh H}-\frac{\cosh Cy}{\cosh C}\right)-\frac{SR}{SC}\frac{\sinh C(1-y)}{\cosh C}\right]\sin Nf \quad (3.11)$$

On inverting Laplace transform as defined by (Sneddon[16])

$$u(y,f,t)=\frac{1}{2\mathrm{p}i}\int_{r-i\infty}^{r+i\infty}\bar{u}(y,f,s)e^{st}dt$$

In equation (3.11), we obtain

$$u=\frac{2}{\pi}\sum_{N=1}^{\infty}\left(\frac{1-\cos N\pi}{N}\right)\left[\sum_{r=0}^{\infty}\frac{4P(-1)^r e^{-A_r t}\cos a_r y}{\pi(2r+1)\left(a_r^{\,2}+c^2\right)}+\int_0^t h(u)\,g(t-u)\,du-\frac{SR}{C}\frac{\sinh C(1-y)}{\cosh C}\right]\sin Nf \quad (3.12)$$

Where $h(u)=\sum_{r=0}^{\infty}\frac{4(-1)^r\cos aye^{-A_r u}}{\pi(2r+1)\left\{1-K\left(a_r^{\,2}+Q^2\right)\right\}}$, $a_r=\frac{\pi}{2}(2r+1)$

and $A_r=\frac{\left(a_r^{\,2}+c^2\right)}{1-k\left(a_r^{\,2}+Q^2\right)}$

Particular cases

Case 1. When $g(t)=C^*$

Using in equation (3.12), we get

$$u=\frac{2}{\pi}\sum_{N=1}^{\infty}\left(\frac{1-\cos N\pi}{N}\right)\left[\sum_{r=0}^{\infty}\frac{4(-1)^r e^{-A_r t}\cos a_r y}{\pi(2r+1)\left(a_r^{\,2}+c^2\right)}(P-C^*)+\frac{C^*}{C^2}\left(1-\frac{\cosh Cy}{\cosh C}\right)\right.$$
$$\left.-\frac{SR}{C}\frac{\sinh C(1-y)}{\cosh C}\right]\sin Nf \quad (3.13)$$

If we take the limit $M\to 0, K\to 0,\ g(t)=P=C^*$ in equation (3.13) then we get the velocity distribution in the case of nonmagnetic and Newtonian fluid. In this case the velocity distribution is

$$u=-\frac{2}{\pi}\sum_{N=1}^{\infty}\left(\frac{1-\cos N\pi}{N}\right)\left[\frac{C}{Q^2}\left(1-\frac{\cosh Qy}{\cosh Q}\right)-\frac{SR}{Q}\frac{\sinh Q(1-y)}{\cosh Q}\right]\sin Nf \qquad (3.14)$$

This is in agreement with Sathyaprakash [13]

Case II

When $g(t)=C^{*}e^{-bt}$, $b>0$, using in equation (3.12), we get

$$u=\frac{2}{\pi}\sum_{N=1}^{\infty}\left(\frac{1-\cos N\pi}{N}\right)\left[\sum_{r=0}^{\infty}\frac{4P(-1)^r\cos a_r y e^{-A_r t}}{\pi(2r+1)\left(a_r{}^2+C^2\right)}\right.$$

$$\left.+\sum_{r=0}^{\infty}\frac{4C^{*}(-1)^r\cos a_r y e^{-bt}-e^{-A_r t}}{\pi(2r+1)\left\{1-K\left(a_r{}^2+Q^2\right)\right\}(A_r-b)}-\frac{SR}{C}\frac{\sinh C(1-y)}{\cosh C}\right]\sin Nf \qquad (3.15)$$

Case III

When $g(t)=C^{*}\cos bt$, using in equation (3.12), we get

$$u=\frac{2}{\pi}\sum_{N=1}^{\infty}\left(\frac{1-\cos N\pi}{N}\right)\left[\sum_{r=0}^{\infty}\frac{4P(-1)^r\cos a_r y e^{-A_r t}}{\pi(2r+1)\left(a_r{}^2+C^2\right)}\right.$$

$$\left.+\sum_{r=0}^{\infty}\frac{4C^{*}(-1)^r\cos a_r y\left(A_r\cos bt+b\sin bt-A_r e^{-A_r t}\right)}{\pi(2r+1)\left\{1-K\left(a_r{}^2+Q^2\right)\right\}\left(A_r{}^2+b^2\right)}-\frac{SR}{C}\frac{\sinh C(1-y)}{\cosh C}\right]\sin Nf$$

(3.16)

4. CONCLUSIONS

Three distinct time dependent pressure gradient namely (i) $g(t)=C^{*}$ (ii) $g(t)=C^{*}e^{-bt}$ and (iii) $g(t)=C^{*}\cos bt$ have been chosen to discuss the velocity profiles. Figures (1) to (3), (4) to (6), (7) to (9) are

drawn to investigate the effects of magnetic parameter M, time t and viscoelastic parameter K on the velocity distribution u in three different cases (i), (ii), (iii) respectively. In all the three cases we noticed that the velocity distribution increases with the increase in M or t where as it decreases with increase in K. Further we observe that in all the three cases the velocity profiles in nonmagnetic case are greater than with the magnetic case. Table (1) is drawn to bring out the effects of Reynolds number R on velocity distribution in three different cases. We noticed that in all the three cases the velocity distribution decreases with the increase in R. Further it is observed that in all three cases the velocity profiles in nonmagnetic case are less than with the magnetic case.

GRAPHS

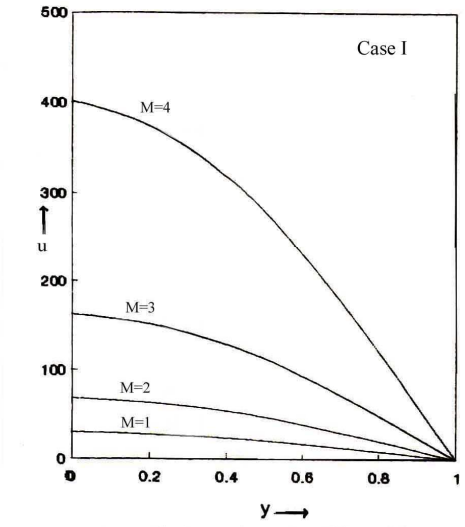

Fig.1 u against y for different M

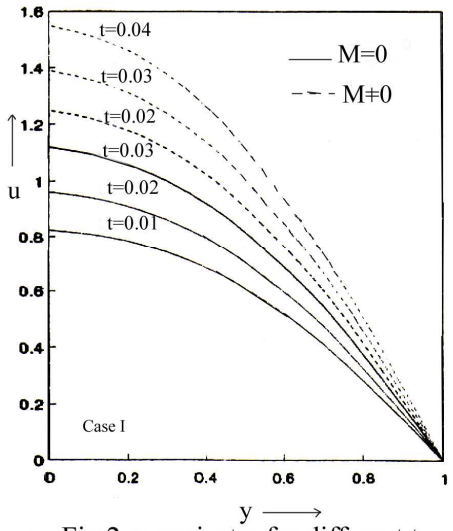

Fig.2 u against y for different t

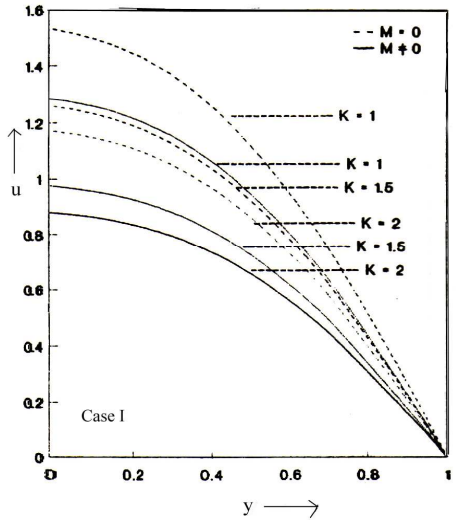

Fig.3 u against y for different K

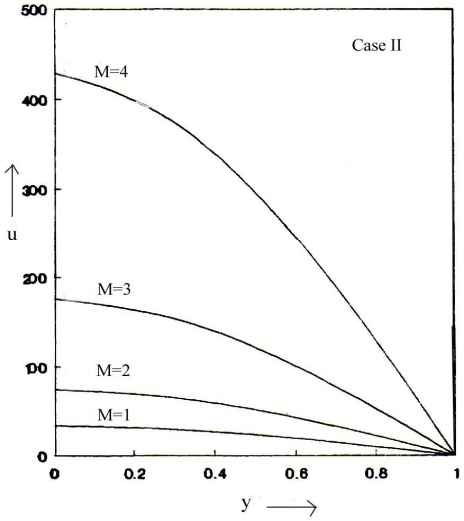

Fig.4 u against y for different M

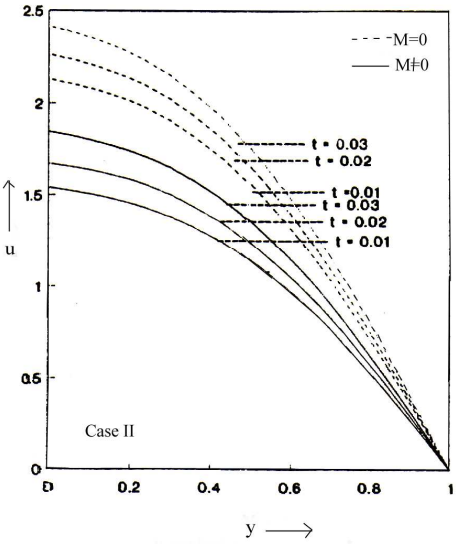

Fig.5 u against y for different t

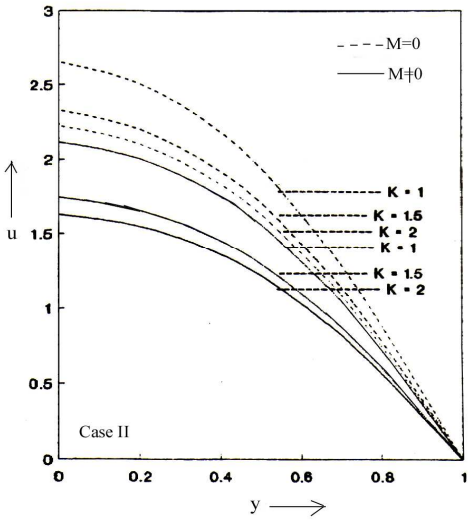

Fig.6 u against y for different K

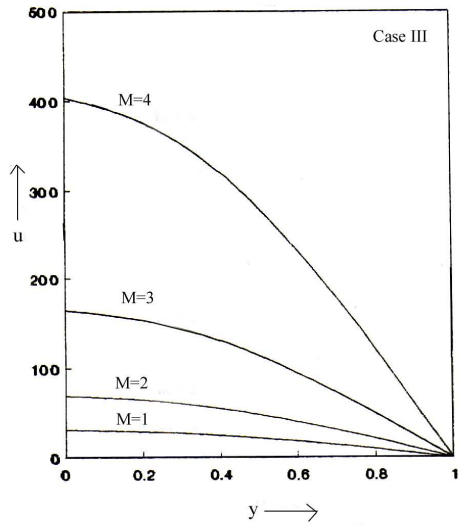

Fig.7 u against y for different M

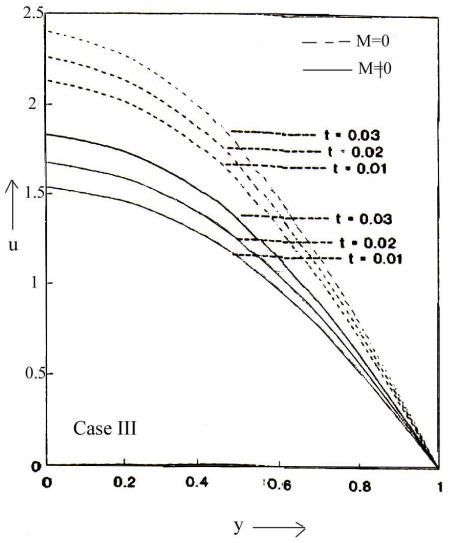

Fig.8 u against y for different t

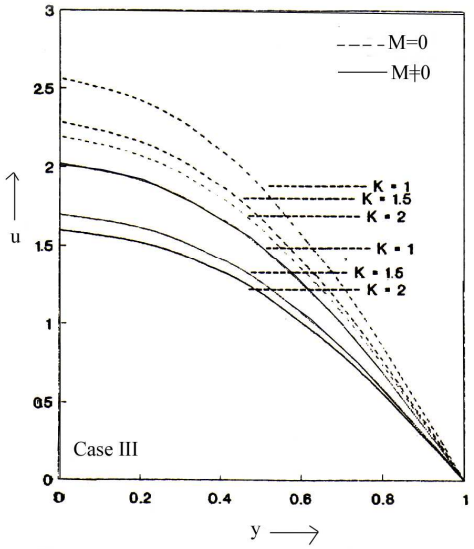

Fig.9 u against y for different K

Table I

Case I U against y for different R

R	M	y=0	0.2	0.4	0.6	0.8	1
2	1	30.1757	28.72465	24.47127	17.8251	9.406355	0
2	0	14.51567	13.83189	11.8035	8.622231	4.568540	0
4	1	30.16235	28.72465	24.47127	17.8251	9.406355	0
4	0	14.50283	13.83189	11.8035	80622231	4.56854	0
6	1	30.14952	28.72465	24.47127	17.8251	9.406355	0
6	0	14.48990	13.83189	11.8035	8.622231	4.56854	0
8	1	30.13668	28.72465	24.47127	17.8251	9.406355	0
8	0	14.47716	13.83189	11.8035	8.622231	4.56854	0
10	1	30.12385	28.72465	24.47127	17.8251	9.40635	0
10	0	14.46432	13.83189	11.8035	8.622231	4.56854	0

Table II

Case II U against y for different R

R	M	y=0	0.2	0.4	0.6	0.8	1
2	1	34.14086	32.50528	27.71082	20.20846	10.68239	0
2	0	17.33922	16.52671	14.12038	10.33622	5.493201	0
4	1	34.12803	32.50528	27.71082	20.20846	10.68239	0
4	0	17.32638	16.52671	14.12038	10.33622	5.493201	0
6	1	34.11519	32.50528	27.71082	20.20846	10.68239	0
6	0	17.3135	16.52671	14.12038	10.33622	5.493201	0
8	1	34.10235	32.50528	27.71082	20.20846	10.68239	0
8	0	17.30071	16.52671	14.12038	10.33622	5.493201	0
10	1	34.08952	32.50528	27.71082	20.20846	10.68239	0
10	0	17.282788	16.52671	14.12038	10.33622	5.493201	0

Table III

Case III U against y for different R

R	M	y=0	0.2	0.4	0.6	0.8	1
2	1	31.2881	29.79183	25.40008	18.52646	9.795776	0
2	0	15.78622	15.04873	12.86061	9.417838	5.008056	0
4	1	31.27597	29.79183	25.40008	18.52646	9.795776	0
4	0	15.77338	15.04873	12.86061	9.417838	5.008056	0
6	1	31.26314	29.79183	25.40008	18.52646	9.795776	0
6	0	15.76055	15.04873	12.86061	9.417838	5.008056	0
8	1	31.2503	29.79183	25.40008	18.52646	9.795776	0
8	0	15.74771	15.04873	12.86061	9.417838	5.008056	0
10	1	31.23747	29.79183	25.40008	18.52646	9.795776	0
10	0	15.73487	15.04873	12.86061	9.417838	5.008056	0

References

[1] Bekhmeteff, B.A., Hydralics of open channels, McGraw-Hill Book Company, New York, 1932.

[2] Chow, V.T., open channel Hydraulics, McGraw-Hill, Inc., New-York, (1959).

[3] Franzini, J.B., and Chisholm, P.S., Wat. Sewage wks 110, (1963), 342.

[4] Gupta, P.C., Chaudhary, J.S., and Sharma, R.G., Indian J. Theor. Phys., 81 (1983), 65.

[5] Henderson, F.M., open channel flow, The Mac. Millian company New York, 1966.

[6] Johnson, J.W., Trans. Amc. Geophys. Un 25 (1944), 906.

[7] Johnson, J.W., and O Brieu, M.P., Engng. News. Rec. 113(1934), 214.

[8] Joneidi, A.A., Domairry, G., Babaelahi, M., Meccanica Vol.45, (2010), 857–868.

[9]

[10] Mahantesh M. Nandeppanavar., M. Subhas Abel., and Jagadish Tawade, Journal of Biorheology Vol. 24. No.1. (2009), 22-28.

[11] Powell, R.W., Trans. ASCE, 3 (1946), 531.

[12] Powell, R.W., Trans. Amc, Geophys. Un. 31 (1950), 57.

[13] Raveer Singh, Ph. D Thesis, Rajasthan University, Jaipur, 1992.

[14] Satyaprakash, Indian J. Pure Appl. Maths. 2 (1971), 103.

[15] Sharma R.C, Kumar P, Sharma S. Int. J. Appl. Mech. Eng.Vol.7. No.2. (2002), 433–44.

[16] Sparrow, E.M. and Cess, R.D., Trans. ASME., J. Appl. Mech., Vol. 29, No.1. (1962), 18.

[17] Sneddon, I.N., The use of integral transforms, Tata McGraw-Hill Book Co. Ltd., New Delhi, (1974).

[18] Vanoni, V.A., Civ. Engng., 11 (1941), 356.

[19] Venkataramana, S., and Bathaiah, D., Ganit (J. Bangladesh Math Soc) Vol.2, No.2. (1972), P. 94-100.

[20] Walter's, K., Quart. J. Math. Appl. Mech., 13 (1960), 444.

Web Pages Clustering: A New Approach

Jeevan H E [#1], Prashanth P P [#2], Punith Kumar S N [#3], Vinay Hegde [#4]
[#] *Dept. of Computer Science and Engineering, RV College of Engineering, Bangalore, Karnataka, India*

***Abstract*—The rapid growth of web has resulted in vast volume of information. Information availability at a rapid speed to the user is vital. English language (or any for that matter) has lot of ambiguity in the usage of words. So there is no guarantee that a keyword based search engine will provide the required results. This paper introduces the use of dictionary (standardised) to obtain the context with which a keyword is used and in turn cluster the results based on this context. These ideas can be merged with a metasearch engine to enhance the search efficiency.**

Keywords: Clustering, concept mining, information retrieval, metasearch engine

I. INTRODUCTION

As information availability increases with the growth of the web, the number of users who want to retrieve that information also increases. This has led to the rise of search engines. A search engine typically is based on a keyword as a query, uses this to search its indexed database which has data about different web sites and their content and presents the results to the user. But users still find it fairly difficult to find the exact information required by them, even though it may be present in the web. There are various reasons for this.

One reason for this is that many users search the Internet with keywords that are ambiguous to certain degree.

For example : If one searches for “keyboard” in a search engine expecting sites containing information about the musical instrument, he gets a list that is a mix of links to pages containing information about typing keyboard and musical instrument.

Today we have many sophisticated search engines like Google, Yahoo, Bing etc. But still we are not guaranteed of accurate search results. Apart from the above mentioned reason, it may also be due to the fact that a single search engine may not be able to index the entire web which has grown to such a large extent. Every day thousands of new web sites are created and millions of existing pages get updated. To keep track of every such detail is impossible.

In order to solve this problem, many meta-search engines emerge such as, Excite, WebCrawler and so on, which make further processing of search results gathered from many existing search engines as explained in [1]. For example Excite issues queries to three other search engines, including Google, Yahoo, and Bing. The results from these search engines are combined to find the most relevant pages. The advantage is obvious. People can fast identify the information they need.

In this paper we propose a simple and effective method to cluster web pages and extract concepts from a keyword. We also introduce an improved ranking algorithm for metasearch engines.

II. WEB PAGES CLUSTERING AND CONCEPT MINING

A. Web Pages Clustering

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”.

A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters as defined in [2].

Web pages clustering, in particular, mean removing irrelevant links from the obtained results. The result from multiple search engines is processed to obtain the final search result page. The result which appears in results of more search engines will be listed above the others.

B. Concept Mining

Concept mining is an activity that results in the extraction of concepts from artefacts. Solutions to the task typically involve aspects of artificial intelligence and statistics, such as data mining and text mining. Because artefacts are typically a loosely structured sequence of words and other symbols (rather than concepts), the problem is nontrivial, but it can provide powerful insights into the meaning, provenance and similarity of documents.

The idea is to use the dictionary available in the Internet to determine the different contexts in which the keyword can appear, that is, the same keyword explaining different concepts.

III. USE OF DICTIONARY

Concept mining as mentioned earlier involves Artificial Intelligence. Extracting concepts from short text snippets retrieved from the search results may not be accurate enough. To achieve good amount of accuracy, we may require the entire text to be available. Hence it can be computationally intensive and consume high bandwidth to function at an acceptable speed [3]. For the internet environment, a better solution can be to use a dictionary. A dictionary can be used for the queries that the user gives. Each ambiguous word will lead to multiple meanings obtained from the dictionary. Based on these multiple meanings clusters can be done for each type of result.

This clustering can be done in two ways. One is to process the search results. Compare the context of the results with the meanings retrieved from the dictionary. This is again not straightforward and requires considerable data mining techniques [4]. Hence we propose a simple alternative but an efficient technique. The technique is to submit the meanings retrieved itself as queries to the search engine. This eliminates the need for any data mining algorithm. Each result retrieved already belongs to a particular cluster (the meaning used for searching). So this eliminates the need for a clustering algorithm.

Now consider a query such as “Bank”. The dictionary can provide meanings such as financial institution, sides of a water body and rely upon. The search engine can resolve the ambiguity by forming three clusters of results, one for each meaning. The meaning itself is sent to the search engine as a query. Further, the results can be improved by concatenating the user query and the meaning and making it a single new query. In this case it can be “Bank financial institution”.

```
//Module to retrieve meanings from a dictionary
//Input- user query – string
//Output- list of meanings
Dictionary (String query)
do
        meanings = getFromDictionary (query);
        for each meaning from the dictionary
        do
                AddToList (list, meaning)
        end
        if (list is NULL) // no meaning found
                // query may be a noun
                AddToList (list, query)
return list
end
```

When it comes to implementation of the same, the dictionary can be maintained either online or offline. An online dictionary such as that of the WorldNet is a better choice, since it is updated regularly and is widely accepted standard dictionary. On the other hand an offline, local dictionary is also possible, provided it is sophisticated enough to provide the results with minimum delay and can be updated regularly.

One problem with this is the use of multiword queries. In this case, it may still be possible to get the meaning of each word of the query from the dictionary, but constructing a new query from that will be a problem. Different solutions can be provided for the same. The algorithm may be designed to select only one word for querying, based on the number of meanings retrieved for each word in the multiword query. The word with maximum number of different meanings can be used. Another solution is to perform a quick concept mining from the multi word query and obtain a single word query. For Example, a query such as “Where is Bangalore”, can be reduced to just “Bangalore”

Another problem with the use of dictionary is for the queries that involve proper nouns. The dictionary is not expected to provide results for these. Even proper nouns can be ambiguous to some extent. For example consider “Sachin”. This could refer to cricket player Sachin Tendulkar or any other individual with the same name (music director Sachin Dev Burman), resolving such ambiguities is non-trivial and may require more input from the user itself. One approach to remove such ambiguities is to use the history of searches by the same user [5]. This can inherently point to a certain context. In this case if the user had earlier searched things about sports, then the probability is more that the query “Sachin” meant “Sachin Tendulkar”. This requires data mining and statistical analysis of previous data available.

IV. METASEARCH ENGINE

A metasearch engine is a search tool that sends user requests to several other search engines and/or databases and aggregates the results into a single list and provides it to the user in way similar to any other search engine. The concept of metasearch engine arises from the fact that the web is too large for one search engine to index it completely and more comprehensive results can be obtained by combining the results of various search engines [6]. The obvious advantage of this technique is that the search space is more i.e. more web pages are covered. Since a metasearch engine has to deal with different search engines, it requires a parsing stage to convert the results from all the search engines into a uniform manner. The implementation can typically involve XML and HTML parsing.

The usage of a metasearch engine must be done in an intelligent manner to extract the maximum benefit out of it. The ranking of results is very crucial to provide the user with the required information in minimum time. A straightforward algorithm that can be adopted to provide a well refined search result is given below. The underlying assumption is that a few

results will be same, from all the search engines. Here we consider the count of each result link from all the search engines used. Then rank it, based on the decreasing order of the count.

```
//Module to search and unify the results
//Performs ranking based on the count
//Input: user query – string
//Output: list of browsable search results

MetaSearchEngine (query)
do
  Submit the search query to the search engines
      for each search engine
      do
            for each result_link from the given
                  search engine
            do
            if (Final_Results has result_link)
            // increment count
            SetCount (result_link, getCount
            (result_link) +1)
            else
            //add it to result list and set count to 1
            AddToFinalResults(result_link)
            SetCount(result_link, 1)
            end
      end
  Sort the Final_Results in the decreasing order of
the
  count of the result
  Display the search results in this order

end
```

The use of this approach provides a far more efficient ranking than simply performing a union of all the results. Moreover it's a simple approach and easily implementable. This ranking can also be done on client side (using client side scripting). Hence it provides a flexible approach for implementation. Experimental implementation of the same technique has been done, with a good amount of success.

V. CONCLUSION

The paper proposed a new basis for web pages clustering and concept extraction from a keyword based on results of multiple search engines on the Internet. It will help user to get relevant information needed upon querying. We also did an experimental implementation of the same ideas, which performed to meet our expectations of speed and efficiency.

It can be said that providing context sensitive results increases the efficiency of the user, so that he can easily find the document he is searching for in the web.

Current keyword based search engines rank the web pages based on frequency of the keywords, inbound link count etc. Hence these results require user to go through all the returned links for finding the right one. With the use of a metasearch engine the relevance of results is also high, since it uses multiple search engines like Google, Yahoo and Bing. The links that appear in most of search engines' results are given higher priority.

Further enhancements include support for queries from languages other than English, enabling caching mechanism for recently queried keywords and moving forward to implement the above idea for image searching as well as video searching.

ACKNOWLEDGMENT

We would like to thank. Dr. T. M. Rangaswamy, Professor, IEM department, R.V College of Engineering for providing support and guidance for the study and research regarding the subject.

REFERENCES

[1] *Fang Li, Martin Mehlitz, Li Feng and Huange Sheng,"Web Pages Clustering and Concept Mining- An approach towards intelligent information retrieval", 2006.*

[2] *Oren Zamir and Oren Etzioni, Department of Computer Science and Engineering, University of Washington," Web Document Clustering: A Feasibility Demonstration".*

[3] *David A Grossman and Ophir Frieder," Information Retrieval – Algorithms and Heuristics", 2004.*

[4] *Jiawei Han and Micheline Kamber, "Data mining: concepts and techniques", 2006.*

[5] *Y Taher H. Haveliwala, Aristides Gionis and Piotr Indyk,"Scalable Techniques for Clustering the Web".*

[6] *Mike Perkowitz and Oren Etzioni, Department of Computer Science and Engineering, Box 352350, University of Washington, Seattle,"Towards adaptive Web sites: Conceptual framework and case study".*

A Performance Study of Data Mining Techniques: Multiple Linear Regression vs. Factor Analysis

Abhishek Taneja, R.K.Chauhan
Assistant Professor, Dept. of Computer Sc. & Applications, DIMT, Kurukshetra
Professor, Dept. of Computer Sc. & Applications,Kurukshetra University, Kurukshetra

Abstract:The growing volume of data usually creates an interesting challenge for the need of data analysis tools that discover regularities in these data. Data mining has emerged as disciplines that contribute tools for data analysis, discovery of hidden knowledge, and autonomous decision making in many application domains. The purpose of this study is to compare the performance of two data mining techniques viz., factor analysis and multiple linear regression for different sample sizes on three unique sets of data. The performance of the two data mining techniques is compared on following parameters like mean square error (MSE), R-square, R-Square adjusted, condition number, root mean square error(RMSE), number of variables included in the prediction model, modified coefficient of efficiency, F-value, and test of normality. These parameters have been computed using various data mining tools like SPSS, XLstat, Stata, and MS-Excel. It is seen that for all the given dataset, factor analysis outperform multiple linear regression. But the absolute value of prediction accuracy varied between the three datasets indicating that the data distribution and data characteristics play a major role in choosing the correct prediction technique.

Keywords: Data mining, Multiple Linear Regression, Factor Analysis, Principal Component Regression, Maximum Liklihood Regression, Generalized Least Square Regression

1. Data Introduction

A basic assumption concerned with general linear regression model is that there is no correlation (or no multi-collinearity) between the explanatory variables. When this assumption is not satisfied, the least squares estimators have large variances and become unstable and may have a wrong sign. Therefore, we resort to biased regression methods, which stabilize the parameter estimates [17]. The data sets we have chosen for this study have a combination of the following characteristics: few predictor variables, many predictor variables, highly collinear variables, very redundant variables and presence of outliers.

The three data sets used in this paper viz., marketing, bank and parkinsons telemonitoring data set are taken from [8],[9], and [10] respectively.

From the foregoing, it can be observed that each of these three sets has unique properties. The marketing dataset consists of 14 demographic attributes. The dataset is a good mixture of categorical and continuous variables with a lot of missing data. This is characteristic for data mining applications.

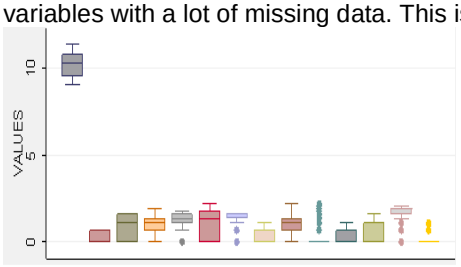

Fig 1 Box Plot of Marketing Dataset

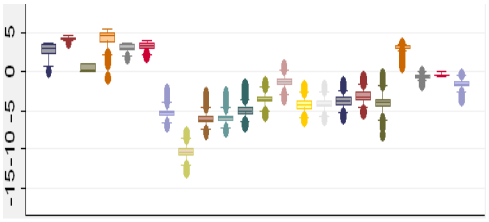

Fig 2: Box Plot of Parkinson Dataset

The bank dataset is synthetically generated from a simulation of how bank-customers choose their banks. Tasks are based on predicting the fraction of bank customers who leave the bank because of full queues.

Each bank has several queues, that open and close according to demand. The tellers have various affectivities, and customers may change queue, if their patience expires.

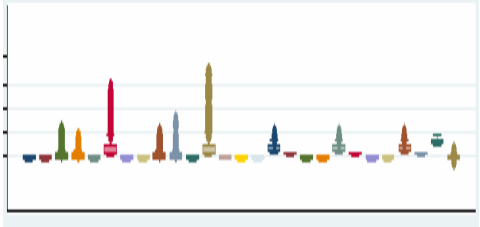

Fig 3: Box Plot of Bank Dataset

In the rej prototasks, the object is to predict the rate of rejections, i.e., the fraction of customers that are turned away from the bank because all the open tellers have full queues. This dataset consists of 32 continuous attributes and having 4500 records.
The parkinsons telemonitoring data set is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring. The recordings were automatically captured in the patient's homes. Columns in the table contain subject number, subject age, subject gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measures. Each row corresponds to one of 5,875 voice recording from these individuals. The main aim of the data is to predict the total UPDRS scores ('total_UPDRS') from the 16 voice measures. This is a multivariate dataset with 26 attributes and 5875 instances. All the attributes are either integer or real with lots of missing and outlier values.
The box plot of the three datasets (fig 1 to fig.3) shown above display measure of dispersion between these variables, compares the mean of different variables, and also shows the outliers in three datasets. In this regard, it becomes necessary to scale these three datasets to reduce the measure of dispersion and bring all the variables of all datasets to the same unit of measure.

2. Prediction Techniques

There are many prediction techniques (association rule analysis, neural networks, regression analysis, decision tree, etc.) but in this study only two linear regression techniques have been compared.

2.1 Multiple Linear Regression

Multiple linear regression model maps a group of predictors x to a response variable y [4]. The multiple linear regression is defined by the following relationship, for $i = 1, 2, n$:

$$y_i = a + b_1x_{i1} + b_2x_{i2} + \cdots + b_kx_{ik} + e_i$$

or, equivalently, in more compact matrix terms:

$$Y = Xb + E$$

where, for all the n considered observations, $\boldsymbol{Y}$ is a column vector with n rows containing the values of the response variable; $\mathbf{X}$ is a matrix with n rows and $k + 1$ columns containing for each column the values of the explanatory variables for the n observations, plus a column (to refer to the intercept) containing n values equal to 1; $\boldsymbol{b}$ is a vector with $k + 1$ rows containing all the model parameters to be estimated on the basis of the data: the intercept and the k slope coefficients relative to each explanatory variable. Finally $\boldsymbol{E}$ is a column vector of length n containing the error terms. In the bivariate case the regression model was represented by a line, now it corresponds to a $(k + 1)$-dimensional plane, called the regression plane. This plane is defined by the equation

$$\hat{y}_i = a + b_1x_{i1} + b_2x_{i2} + \cdots + b_kx_{ik} + \mu_i$$

Where $\hat{y}_i$ is dependent variable. $X_i^{'s}$ are independent variables, and μ_i is stochastic error term. We have compared three basic methods under this multiple linear regression technique. They are full method (which uses the least square approach), forward method, and stepwise approach (which used discriminant approach or all possible subsets) [5].

2.2 Factor Analysis

Factor analysis attempts to represent a set of observed variables $X_1, X_2 \ldots X_n$ in terms of a number of 'common' factors plus a factor which is unique to each variable. The common factors (sometimes called latent variables) are hypothetical variables which explain why a number of variables are

correlated with each other- it is because they have one or more factors *in common* [7].

Factor analysis is basically a one-sample procedure [6]. We assume a random sample $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_n$ from a homogeneous population with mean vector μ and covariance matrix $\sum$. The factor analysis model expresses each variable as a linear combination of underlying *common factors* f_1, f_2, . . . , f_m, with an accompanying error term to account for that part of the variable that is unique (not in common with the variables). For y_1, y_2, y_p in any observation vector $\mathbf{y}$, the model is as follows:

$$y_1 - \mu_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1m} f_m + \varepsilon_1$$

$$y_2 - \mu_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2m} f_m + \varepsilon_2$$

...

$$y_p - \mu_p = \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pm} f_m + \varepsilon_p.$$

Ideally, m should be substantially smaller than p; otherwise we have not achieved a parsimonious description of the variables as functions of a few underlying factors. We might regard the *f's* in equations above as random variables that engender the *y*'s. The coefficients λ_{ij} are called *loadings* and serve as weights, showing how each y_i individually depends on the *f* 's. With appropriate assumptions, λ_{ij} indicates the importance of the *j*th factor f_j to the *i*th variable y_i and can be used in interpretation of f_j. We describe or interpret f_2, for example, by examining its coefficients, λ_{12}, λ_{22}, λ_{p2}. The larger loadings relate f_2 to the corresponding *y*'s. From these *y*'s, we infer a meaning or description of f_2. After estimating the λ_{ij} 's, it is hoped they will partition the variables into groups corresponding to factors. There is superficial resemblance to the multiple linear regression, but there are fundamental differences. For example, firstly f's in above equations are unobserved, secondly equations above represents one observational vector, whereas multiple linear regression depicts all n observations.

There are a number of different varieties of factor analysis: the comparison here is limited to principal component analysis, generalized least square and maximum likelihood estimation.

3. Related Work

There are many data mining techniques (decision tree, neural networks, regression, clustering etc.) but in this paper we have compared two linear techniques viz., multiple linear regression, and factor analysis. In this domain there have been many researchers and authors who compared various data mining techniques from varied aspects.

In year 2004 Munoz et. al did a comparison of three data mining methods: linear statistical methods, neural network method, and non-linear multivariate methods [11]. In 2008, Saikat and Jun Yan compared PCA and PLS on simulated data [12]. Munoz et.al compared logistic regression, principal component regression, and classification and regression tree with multivariate adaptive regression spines [16]. In 1999, Manel et.al compared discriminate analysis, neural networks, and logistic regression for predicting species distribution [13]. In year 2005, Orsalya et. al compared ridge regression, pair wise correlation method, forward selection, best subset selection, on quantitative structure retention relationship study based on multiple linear regression on predicting the retention indices for aliphatic alcohols[14]. In year 2002 Huang et. al compared least square regression, ridge and partial least square in the context of the varying calibration data size using only squared prediction errors as the only model comparison criteria [15].

4. Preparation and Methodology

Both the techniques under study are linear in nature and the choice of technique is vital for getting significant results. When a nonlinear data are fitted to a linear technique, the results obtained are biased and when linear data are fitted to a non-linear technique, the results have increased variance. As the techniques undertaken for this study are both linear, so to get significant results we need to apply the same on linear data sets. Both the techniques are linear regression techniques, we mean that they are linear in parameters [1] [2]; the β's (that is, the parameters are raised to the first power only. It may or may not be linear in explanatory variables, the X's. To make our data sets linear it is preprocessed by taking natural log of all the instances of the data sets or normalized using z-score [3] normalization. After

scaling and standardizing the three datasets, it is found that skewness is reduced that is shown by histogram diagram of all three datasets. For proving linearity of these data sets box-plot, histogram and JB Test (Jarque Bera Test) with p-value (exact significance level or probability value of committing type-I error) have been used.

After scaling and standardizing the data sets are divided into two parts, taking 70% observations as the "training set" and the remaining 30% observations as the "test validation set"[3]. For each data set training set is used to build the model and various methods of that technique are employed. For example in Multiple Linear Regression (MLR), three methods are associated in this study: the full model, forward model and stepwise model. The model is validated using test validation data set and the results are presented using ten goodness of fit criteria. Both the techniques are intra and inter compared for their performance on the underlying three unique datasets.

5. Interpretation and Findings

Refer to table 1and table 2 given below.

5.1 Interpreting Marketing Dataset

In marketing dataset, the value of R^2 and Adj.R^2, of full model was found with good explanatory power i.e., 0.47, which is higher than both stepwise and forward model.

On the behalf of this explanatory power value we can say that among all methods of multiple linear regression, full model was found best method for data mining purpose, since 47% change in variation in dependent variable was explained by independent variables. But 0.47 value of explanatory power is not significant up-to the mark which requires another regression model than multiple regression model for reporting data set, since 0.53 means 53% of the total variation was found unexplained. So, within multiple regression techniques full model was found best but not up-to the mark. Value of R^2 suggest for using another regression model.

	Methods	MSE	MAE	CN	No. of variables	R Square	Adj. R Square	RMSE	F Value (dF, No. of Observations)	Modified Coefficient of efficiency	Test of normality
MLR (MARKETING DATASET)	FULL MODEL	0.333	0.33	6.87e+6	13	0.4765	0.4751	.57728	336.50 (13, 4805)	-0.009	0.6325
	STEPWISE MODEL	0.603	4.94	5.10e+5	13	0.436	0.435667	0.77	1042.32 (11,4805)	0.047	0.6162
	FORWARD MODEL	.584	0.897	3.53e+3	13	.459	.458	0.76	410.48 (13,4805)	0.077	0.6826
MLR (PARKINSON DATASET)	FULL MODEL	1.256	18.55	8.284e+8	19	0.9073	0.9068	.13359	2106.68 (19, 4092)	56.10	0.7251
	STEPWISE MODEL	0.021	9.936	8.67e+8	19	.910	.909	0.144	2288.954 (18,4092)	0.090	0.7343
	FORWARD MODEL	.171	10.04	0.607e+.6	19	0.196	0.193	0.42	62.351 (16,4092)	0.139	0.7651
MLR (BANK DATASET)	FULL MODEL	3.818	2.534	0.33212	32	0.0348	0.0248	1.9542	3.51 (32, 3116)	6.786	0.7876
	STEPWISE MODEL	4.54	2.865	0.4534	32	0.0563	0.0527	2.1307	4.45 (32, 3116)	7.896	0.765
	FORWARD MODEL	4.86	2.476	0.4653	32	0.0564	0.05383	2.204	3.851 (32,3116)	6.765	0.5876

Table 1

The inclusion of some other independent variables (either relevant or irrelevant) in multiple regression model mostly generate non-decreasing explanatory value or R^2 value. In this case we can use anther good measure of R^2 i.e., Adj. R^2, which accounts for the effect of new explanatory variables in the model, since it incorporate degree of freedom of the model, or denominator of the explained and unexplained variation[18]. The

expression for the adjusted multiple determination is:

Adj. $R^2 = 1-(1-r^2)\quad \frac{n-1}{n-k}$

Adj. $R^2 = 1-\left[\frac{\sum e_i^{\,2}/(n-k)}{\sum y^2/(n-1)}\right]$

If n is large Adj. R^2 and R^2 will not differ much. But with small samples, if the number of regressors X's is large in relation to the sample observations Adj. R^2 will be much smaller than R^2 and can even assume negative values in which case Adj. R^2 should be interpreted as being equal to zero.

For marketing data set, all methods of multiple linear regression Adj. R^2 was found similar to R^2 value which means sample size is sufficiently large as required for data mining purpose [19].

	Methods	MSE	MAE	CN	No. of variables	R Square	Adj. R Square	RMSE	F-Value (dF, No. of Observations)	Modified Coefficient of efficiency	Test of normality
FACTOR ANALYSIS (*MARKETING DATASET*)	PCR	0.756	3.67	12	13 *(with four components)*	0.584	0.56	0.8694	323.65 (13,4819)	5.754	0.6654
	MAXIMUM LIKLIHOOD	0.775	3.98	9.78e+9	13	0.589	0.576	0.8803	367.455 (13,4819)	5.9876	0.6792
	GLS	0.746	3.998	11	13	0.587	0.573	0.8602	386.78 (13,4819)	5.7685	0.6776
FACTOR ANALYSIS (*PARKINSON DATASET*)	PCR	0.456	0.67	7.87e+7	19 *(with six components)*	0.63	0.51	0.6749	543.5 (19,4112)	8.56	0.87
	MAXIMUM LIKLIHOOD	0.582	0.655	7.10e+7	19	0.64	0.54	0.763	513.65 (19,4112)	9.38	1.73
	GLS	0.398	1.677	6.54e+6	19	0.67	0.56	0.63	665.45(11, 4112)	11.09	1.96
FACTOR ANALYSIS (*BANK DATASET*)	PCR	0.643	0.58	8.86e+8	33 *(with six components)*	0.74	0.69	0.80	654.45 (34,3150)	0.0544	0.6758
	MAXIMUM LIKLIHOOD	0.665	0.598	8.75e+8	33	0.728	0.684	0.815	675.65 (34,3150)	0.0546	0.0754
	GLS	0.678	0.612	8.74e+8	33	0.715	0.682	0.823	688.45 (34,3150)	0.0568	0.0543

Table 2

The R^2 in case of marketing dataset for factor analysis was found around 0.58. So, all methods have equal explanatory power under factor analysis. More over, under all methods viz., PCR, Maximum Likelihood, and GLS, explained variation is 58% out of total variation in the dependent variable which signifies that factor analysis extraction is better than multiple linear regression. R^2 can also be estimated through the following notations: $R^2=\frac{ESS}{TSS}$

TSS = Explained Sum Square(ESS)+ Residual Sum Square(RSS)

The Adj. R^2 i.e., adjusted for inclusion of new explanatory variable was also found 0.56 less than R^2. The 58% variation was captured due to regression, it explains the overall goodness of fit of the regression line to marketing dataset due to use of factor analysis.

So, on the behalf of first order statistical test (R^2), we can conclude that factor analysis technique is

better than multiple regression technique due to explanatory power.

Mean Square Error (MSE) criteria is a combination of unbiased-ness and the minimum variance property. An estimator is a minimum MSE estimator if it has smallest MSE, defined as the expected value of the squared differences of the estimator around the true population parameter b. MSE($\hat{b}$) =E($\hat{b}$ -b)^2 . It can be proved that it is equal to

MSE($\hat{b}$)'s
=Var($\hat{b}$)'s+bias^2 ($\hat{b}$)

The MSE criteria for unbiased-ness and minimum variance were found increasing under multiple linear regression models. It signifies that full method MSE is less than all model's MSE, which further means that under full model of multiple linear regression of marketing dataset there is less unbiased-ness and less variance.

The minimum variance also increases the probability of unbiased-ness and gives better explanatory power like R^2 in marketing dataset.

The inter comparison of two techniques multiple linear regression and factor analysis generated that in factor analysis models MSE is significantly different which signifies that under factor analysis all b's are unbiased but with large variance. Due to large variance in factor analysis techniques the probability value of unbiased-ness increases that generates a contradictory result about the explanatory power of the factor analysis methods. But factor analysis methods may have questionable values of MSE, due to this reason new measure of MSE that is RMSE (root mean square error) was used in the study.

RMSE was found considerably similar in methods of both the techniques. Due to less variation in RMSE of both MLR and factor analysis of marketing dataset it can be stated that both techniques have equal weights for consideration.

A common measure used to compare the prediction performance of different models is Mean Absolute Error (MAE).

If Y^p be the predicted dependent variable and Y be the actual dependent variable then the MAE can be computed by

$$\text{MAE}=\frac{1}{n}\frac{\sum|Y-Y^p|}{Y}$$

In marketing dataset MAE was found less under full model, which is less than stepwise and forward model. MAE signifies that full model under MLR techniques give better prediction than other mode

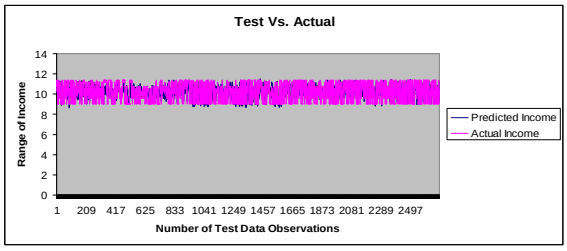

Fig 4: MLR-Full Model (Marketing)

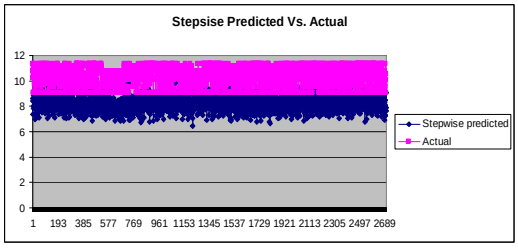

Fig 5: MLR-Stepwise Model (Marketing)

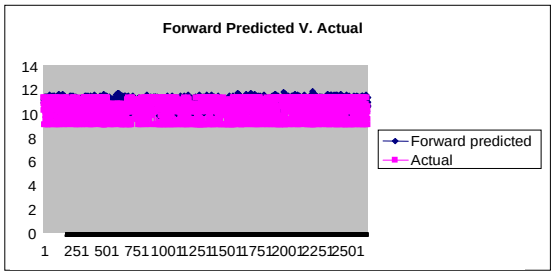

Fig 6: MLR-Forward Model (Marketing)

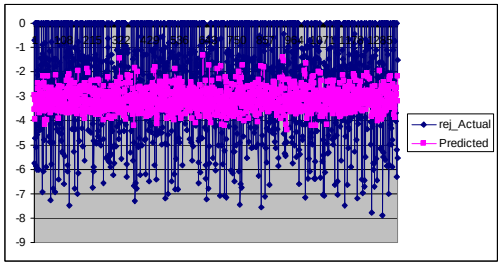

Fig 7: MLR-Full Model (Bank Dataset)

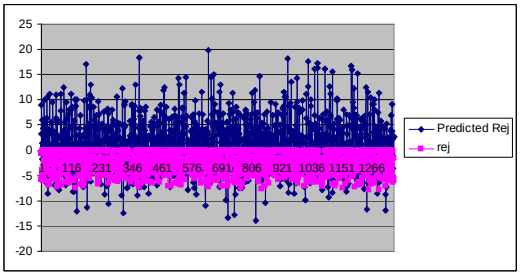

Fig 8: MLR-Forward Model (Bank Dataset)

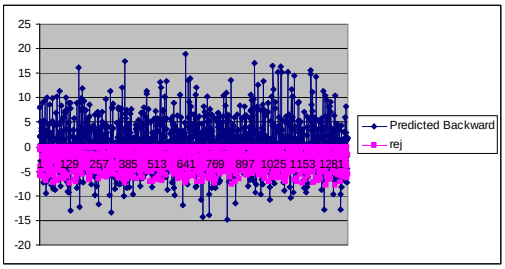

Fig 9: MLR-Stepwise Model (Bank Dataset)

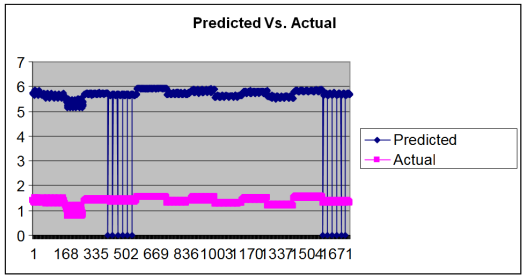

Fig 10: MLR-Full Model (Parkinson Dataset)

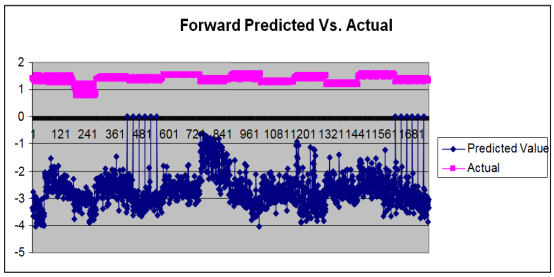

Fig 11: MLR-Forward Model (Parkinson Dataset)

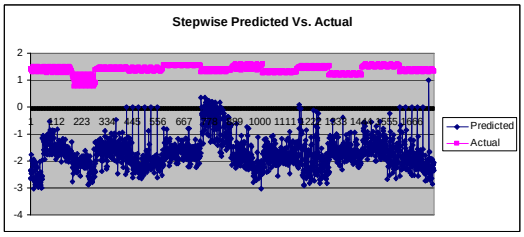

Fig 12: MLR-Stepwise Model (Parkinson Dataset)

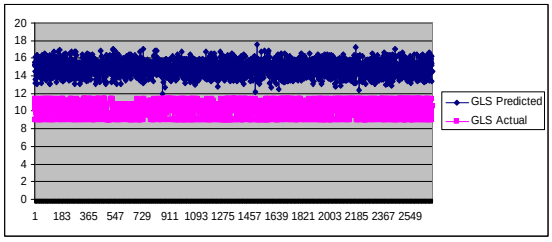

Fig 13: Factor Analysis-GLS Model (Marketing Dataset)

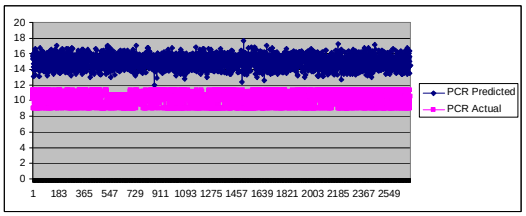

Fig 14: Factor Analysis-PCR Model (Marketing Dataset)

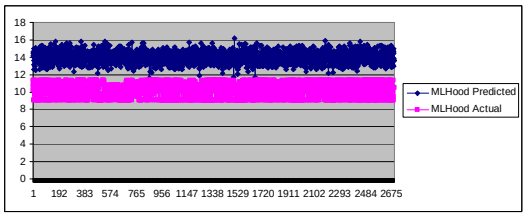

Fig 15: Factor Analysis-Maximum Likelihood Model (Marketing Dataset)

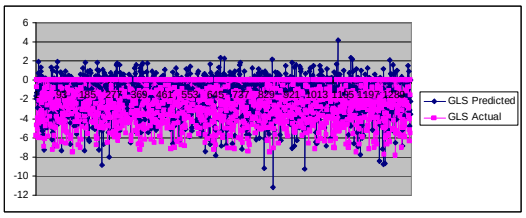

Fig 16: Factor Analysis-GLS Model (Bank Dataset)

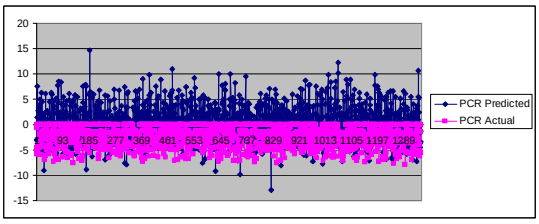

Fig 17: Factor Analysis-PCR Model (Bank Dataset)

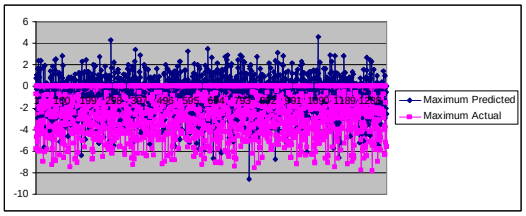

Fig 18: Factor Analysis-Maximum Likelihood Model (Bank Dataset)

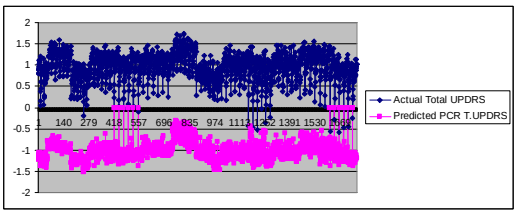

Fig 19: Factor Analysis-PCR Model (Parkinson Dataset)

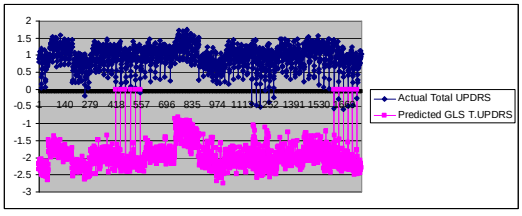

Fig 20: Factor Analysis-GLS Model (Parkinson Dataset)

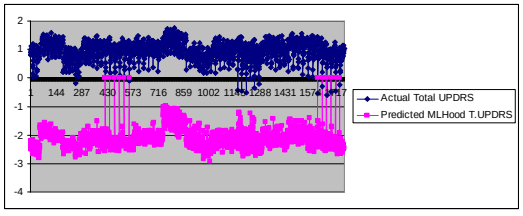

Fig 21: Factor Analysis-Maximum Likelihood Model (Parkinson Dataset)

Under factor analysis marketing dataset MAE in all models was found considerably similar but higher than multiple regression techniques, therefore we can say factor analysis models for such kind of datasets generate poor prediction performance.

The diagnosis index of multi collinearity was found significantly below 100 under MLR methods in marketing dataset, which means there is no scope for high and severe multi collinearity. In case of same dataset condition number was found lower

than factor analysis technique. This means factor analysis is better technique to diagnosis the effect of multi collinearity. But in marketing dataset both factor analysis and MLR techniques were found with less multi collinearity in regressors than severe level of multi collinearity.

The F value in case of marketing dataset was found more than critical value with respect to dF(degree of freedom), in both techniques, which signifies that overall regression model is significantly estimated but stepwise model of MLR technique was found high F corresponding to its dF which means overall significance of the regression model was up-to the mark in case of stepwise method. The prediction plots of two techniques on marketing dataset better represent above discussion visually (see fig. 4-fig. 6 and fig. 13- fig. 15)

5.2 Interpreting Bank Dataset

In case full model of bank dataset explanatory power (R^2) was found considerably low due to residual, whereas in stepwise and forward model MLR generated satisfactory explanatory power. Due to stepwise and forward model 56% variation in dependent variable was explained with respect to independent variables. Another measure of explanatory power was also found satisfactory in case of stepwise and forward model but not in full model.

On the other hand factor analysis models on bank dataset generated higher value of both R^2 and adjusted R^2, which signifies that the explanatory power of factor analysis in case of bank dataset is more than MLR technique. Overall one drastic point was found that in all models of factor analysis and MLR, full model of MLR generated very poor R^2 value, which means this dataset is not having proper specification according to magnitude change.

The MSE criteria for unbiasness and minimum variance for all parameters is found increasing under both factor analysis and MLR techniques, but all models of factor analysis are found with low unbiasness and variance than all models of MLR. It means both the technique parameters are significant, but MLR techniques parameters are significant with high variance.

The RMSE is also satisfactory and upto the mark in case of factor analysis. Therefore, we can say that factor analysis parameters have low variance and unbiasness.

The prediction power of the regression model is also found good fit in all factor analysis models. In case of bank dataset MLR is having more MAE due to test dataset skewness.

Modified coefficient of efficiency was found low in case of factor analysis model in case of bank dataset, since this dataset does not satisfy the center limit theorem due to constant number of variables; but in MLR model modifies coefficient of efficiency was found considerably significant for all models. This may be due to the successful implementation of center limit theorem.

In case bank dataset the diagnosis index of multi-collinearity was found higher in factor analysis than MLR, which signifies that factor analysis is better technique to identify multi-colinearity problem.

The F value in case of bank dataset was found significant under MLR model but F value was found very low rather in case of factor analysis was found 200 times more than the critical value, which means overall significance of all factor analysis model is higher than MLR model. The prediction plots of the two techniques (see fig. 7-fig. 9 and fig. 16- fig. 18) corroborate our discussion.

5.3 Interpreting of Parkinson Dataset

In case of Parkinson dataset forward model of MLR was found very low explanatory power, it is due to hetroscedasticity in stochastic error term of the model, but the full and the stepwise model was found to have 90% explanatory power of the model. In all models of factor analysis R^2 was found to have 60%, which is considerably sufficient for satisfactory explanatory power of the model. Moreover adjusted R^2 was found similar in both techniques i.e., MLR and factor analysis, due to no intrapolation.

In case of MLR models on Parkinson dataset MSE was found low and up-to the mark, which signifies that MLR technique is better technique for the extraction of structural parameters with unbiasness and low variance. On the other hand factor analysis was found having high biasness and high variance for extracting structural parameters of the model.

RMSE was found similar in all models of MLR and factor analysis which signifies the same consideration for unbiasness and variance.

The prediction power (MAE) of two models of factor analyis viz. PCR and maximum likelihood was found significant but GLS model prediction power was found considerably higher than PCR and maximum likelihood methods. On the other hand MLR prediction power was found significantly different in all three models. In case of stepwise and forward models prediction power increased more than full model.

The center limit theorem for getting efficiency of the model was found incompatible, but in case of factor analysis it was found satisfactory to the center limit theorem. Overall inn case of factor analysis modified coefficient of efficiency was found increasing.

In Parkinson dataset multi-colinearity extraction index was found higher under all models of MLR techniques except forward model. In factor analysis on the same dataset, this index was found lower than MLR model. This means MLR is better technique for diagnosing multi-colinearity particularly with full and stepwise methods.

The significance of overall model was found higher in two models of MLR viz. full and stepwise methods but in case of factor analysis, overall significance of regression model was found similar in all methods. The forward method of MLR generated considerably low F value, which means overall significance is poor than another models of both technique. The prediction plots of two techniques on Parkinson dataset is given in figure 10 to figure 12 and figure 19 to figure 21.

6. Conclusion and Future Work

The analysis of linear techniques (MLR and Factor Analysis) suggests that factor analysis is considerably better technique than MLR. The principal component model extracted good performance on all datasets of the study. The good performance is said on the basis of higher explanatory power, higher goodness of fit, and higher prediction power.

In diagnosis of multi-colinearity PCR model of factor analysis was found better model. However, full model of MLR also extracted satisfactory result. All other models of both the techniques were found with high explanatory power but with moderate prediction power.

All models are best fit from the point of view of linearity and unbiased ness due to moderate variance and heteroscedasticity, distribution of residual term. Their prediction power was found considerably moderate fit.

From the point of view of structural parameters and overall significance of regression model again factor analysis was found significantly up-to the mark.

From overall analysis of regression technique we can say that data with high skew ness and large structural observations should be estimated/treated with principal component model of factor analysis. The dataset with high multi-colinearity should also be treated through factors/components according to relevancy. The small dataset on the other hand should be extracted through full model of multiple regression.

The compatibility of a technique on particular dataset also depends on particular dataset's distribution of residual term of the model. In our study marketing or Parkinson dataset are having normal distribution of the residual term, on the other hand bank dataset residual term was found non normally distributed considerably. The violation of this residual assumption is affecting the prediction power for removing heteroscedastic variance of residual term. The method GLS should be adopted to estimate the structural parameters with suitable suggested forms of the regression model.

The techniques in which estimators satisfy BLUE (best, linear, unbiased, and efficient) properties of structural parameters estimates and stochastic random error term are considered better than others.

The skewness of predictors and random term in the linear regression model is creating obstacles to satisfy BLUE properties. Reducing skewness with some advance data mining tool and then comparing performance of said techniques can further enlighten us, which is an area that can be further explored.

References

[1] *Gujarati N. Damodar, Sangeetha, "Basic Econometrics" 4th edition, New York: McGraw Hill, (2007).*

[2] Walpole, R.E, S.L Myers, and K. Ye., Probability and Statistics for Engineers and Scientists, 7th edition. Englewood Cliffs, NJ: Prentice Hall (2002).

[3] Myatt J. Glenn, "Making Sense of Data-A practical guide to exploratory data analysis and data mining" New Jersy: Wiley-Interscience (2007).

[4] Giudici Paolo, "Applied Data Mining-Statistical methods for business and industry" wiley, (2003)

[5] Dash, M., and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis. 1:3 (1997) pp. 131-156.

[6] Rencher C. Alvin, "Methods of Multivariate Analysis" 2nd Edition, Wiley Interscience, (2002).

[7] Kim, Jae-on.; Mueller, Charles W., "Introduction to Factor Analysis-What it is and how to do it.", Sage Publications, Inc., (1978).

[8] http://www-stat.stanford.edu/~tibs/ElemStatLearn/,

[9] http://www.cs.toronto.edu/~delve/data/bank/desc.html

[10] http://archive.ics.uci.edu/ml/datasets.html

[11] Munoz, Jesus, and Angel M. Felicisimo, "Comparison of Statistical Methods

Commonly used in Predictive Modeling," Journal for Vegetations Science

15 (2004), pp. 285-292.

[12] Maitra Saikat and Yan Jun," Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression" Casualty Actuarial Society, 2008 Discussion Paper Program, pp. 79-90

[13] Manel, S., J. M. Dias, and S. J. Ormerod, "Comparing Discriminant Analysis,

Neural Networks and Logistic Regression for Predicting Species

Distribution: A Case Study with a Himalayian River Bird," Ecol. Model.

120 (1999), pp. 337-347.

[14] Farkas, Orsolya, and Heberger Karoly, "Comparison of Ridge Regression, PLS,

Pairwise Correlation, Forward and Best Subset Selection methods for

Prediction of Retention indices for Aliphatic Alcohols," Journal of

Information and Modeling, 45:2 (2005) pp. 339-346.

[15] Huang, J. et al., "A Comparison of Calibration Methods Based on Calibration

Data size and Robustness," Journal of Chemometrics and Intelligent Lab.

Systems, 62:1 (2002) pp. 25-35.

[16] Specht, D. F., "A General Regression Neural Network," IEEE Transactions on

Neural Networks, 2:6 (1991), pp. 568-576.[17] Al-Kassab M, "A Monte Carlo Comparison between Ridge and Principal Components Regression Methods" Applied Mathematical Sciences, Vol. 3, 2009, no. 42, 2085 - 2098

[18] Larose T. Daniel, "Data Mining-Methods and Models" Wiley Interscience, (2002), pp 114.

[19] Han Jiawei and Kamber Micheline "Data Mining Concepts and Techniques" Morgan Kaufmann Publishers, 2006, PP6.

E- Learning: An effective pedagogical tool for learning

Boumedyen [#1], Kaneez [*2], Rafael [#3] Victor [**4]
[#1]*Information System, University of Nizwa,* [#3]*Information System, SPIRAN University*
[#1]*Birkut- ul- Mauz,Nizwa, Sultanate of Oman, ,* [#3]*St. Petersburg, Russia*
[*2] *Management, University of Nizwa*
Birkut- ul- Mauz,Nizwa, Sultanate of Oman
, [**4]Information System, SPIRAN University,, St. Petersburg, Russia

***Abstract*— In the info-tech age E-Methods of learning are becoming the most important vehicle in disseminating knowledge in higher education institutions. This sector is growing and changing at a rapid speed due to developments in technologies. But teaching is an art. Can there be fun learning with raw and dry technology? How can we make the best use of E- Methods, can we make the required information and data available to the students in a flexible manner, at ease all the time? What are the advantages of traditional methods of teaching and learning? Is E-learning a progressive stage incubating all the benefits of the Manual learning or it is only a window dressing on the face of advancement? Can we convert the boring, tedious subjects into interactive, monotony breaking joyous learning? In this paper the researchers have focused on the modernization of E- Pedagogy vis-à-vis the traditional method of learning. They have highlighted the effectiveness of using the E-learning elements and various E- Methods. This work has used the decision tree algorithms particularly Classifiers.trees.J48 The obtained results show that using online examination attribute plays major role in increasing the average grade of the class in higher education. The novelty of this work is that the researchers have focused on the teaching methodology used by the faculty members and the tools available in the universities. We believe that this work will play a constructive role in building higher education system. Our generated rules/output can be used by the decision makers in the improvement of higher education system processes.**

Key words: E- learning, higher education systems, modernization, decision tree algorithm.

VI. INTRODUCTION

I n the info-tech age E-Methods of learning are becoming the most important vehicle in disseminating knowledge in higher education institutions. This sector is growing and changing at a rapid speed due to developments in technologies.

But teaching is an art. Ever since Socrates thought of teaching geometry to the slave boy in Plato's Meno, the nature of learning has been an active topic of investigation [1]. Both undergraduate and

graduate courses are experiencing a migration away from the traditional classroom and toward a greater emphasis for electronic delivery of content [2]. This trend is equally applicable on all departments and schools in the university system but is especially critical in business schools, since the preparation of students for successful business careers will depend on the students' abilities to accurately assess the quality of teaching and rapidly adapt to the changing pedagogy that reflect radical technological advances. The researchers have tried to examine the adoption of pedagogical changes in Higher Education with respect to the introduction and growth of e-learning. Our professed aim is to use e-learning to improve the quality of the teaching-learning experience for faculties and students. Unfortunately, the teaching quality ranks poor in relation to most of changes required in higher education. Out of the many traditional and non traditional pedagogical tools for learning such as oral conversation, case study method, group discussions, classroom games, simulation exercises, distance learning etc E-learning emerges to be the best one. This is high time when higher education institutions think seriously on improving faculty-student learning, and one humble step in this direction is to exploit e-learning. There is always resistance to change, each time we have to ride on a new wave of technological innovation. The novelty of this work is that unlike other researches which focuses on students learning behavior we have focused on the teaching pedagogy used by the faculty members and the instruments available in the universities. Usually the blame goes to the students for their non-performance as being inattentive or careless or indolent, lethargic or idle but we ignore the important vehicle in education that is the Instructor, education imparting pedagogy, electronic media, and instruments of teaching and available resources. We believe that this work will play a constructive role in building higher education system. It will encourage the use of technology in teaching. Our generated

rules/output can be used by the decision makers in the improvement of higher education system processes. In this paper we have studied the instructor's behavior and the class attributes. Further research can be carried out on the basis of this work so as to compare the student's behavior or learning attitude with the present study.

What is M-learning (Manual learning)?

The classroom face to face teaching with the help of text books and scheduling the class timings, following strict timetable, physical presence, hardcopy of class notes, books, assignments, pre-decided meeting places and timings, face to face interaction, communication and question answer sessions, group discussions and physical participation in educational games, usually one teacher and students of similar age form traditional learning [1] [2].

C. *What is E- learning?*

Using the electronic methods as learning tools such as multimedia, internet, computer, software, online textual materials to streaming video to chat rooms and discussion boards which means that students have many more choices in an E-learning environment than they had in a more M-learning, face-to-face environment. Catalog of media-enhanced Power-Point slides, streaming video lectures, some interactive Excel-based practice problems which may be individualized to the students form part of E-learning [3]

D. *Why is e-learning important for HE (Higher Education)?*

[3]Higher education is the link between knowledge gained and practical implementation in industry. E-learning allows students access to learning without the constraints of time and location , [4] Some of the added benefits of E-education include flexibility; ease of participation; absence of labeling due to such things as race, gender, and appearance; training in electronic communication; and exposing students to information technology [5] E-learning in general and online college education specifically are having a profound effect on the future of postsecondary education and is transforming the educational model from an instructor-driven to an interactive and community-driven educational environment in which all students share responsibility for learning outcomes [6].

VII. RELATED WORK

Recently many researchers have worked to enhance and evaluate the higher education tasks. Some researchers have proposed methods and architectures by using data mining in higher education. In [7] they investigated the current trends in improving the higher education systems, to understand from the outside which factors might create devoted students. They used Data mining methods to extract valuable information from existing students so as to handle prospective students in a better manner. They have generated some rules which may be used to understand the behavioral pattern and learning attitude of the prospective enrollments at an early stage and thus the concentration of effort in higher education systems may be implemented. The research by [8] proposed a model to represent how data mining is used in higher educational system to improve the efficiency and effectiveness of the traditional processes. [9] Discusses different Al technologies and compares them with genetic algorithm based induction of decision trees and discusses why the approach has a potential for developing an alert tool. The researchers in [10] proposed a model to represent how data mining is used in Institutes of higher learning to improve the competence and effectiveness of the traditional processes. In this model a guideline was presented for higher educational system to improve their decision-making processes. The Work by [11] is to use Rough Set theory as classification approach to analyze student data. In this research the Rosetta toolkit has been used to evaluate the student data to describe the dependencies between the attributes and the student status.

Many other related Information and works can be found in [12,13,14,15,16,17,18,19,20,21,22,23,24].

VIII. DATA COLLECTION

For implementing this work we have used the data base of specific university. For some reason we are not able to reveal the name of the university. The data base has information about the teaching attributes and overall marks achieved by the class for one semester. The database has information about marks of 139 courses from different domains (subjects) for one semester, provided by specific university. Each of the attributes in this data set has dichotomous values- Yes and No value. If the instructor uses the particular teaching method the value will be equal to 'Yes' and if not the value will be 'No'. For simplification of the processing of the

collected data set we replaced the nominal value to numerical value i.e. Yes=1, and No=0. All the attributes have two instances except the 'mark' attribute which has six instances i.e. D, D+, C, C+, B, B+ (see fig.3.1)

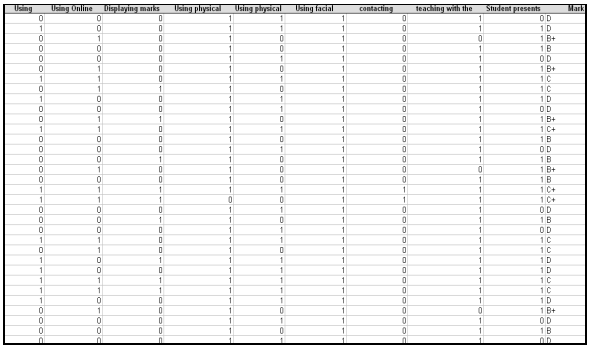

Fig 3.1 Data Set Sample in binary view

IX. EXPERIMENT

Using machine learning to find the conditions that are suitable for improving higher education system using E- learning methods and M- learning (manual learning) methods.

The general instances in the data set are characterized by the values of attributes that measures different aspects of the instances. In this work there are 25 attributes such as Using power point, Using multimedia in the class, Using physical models, Using e-mail for communication with students, Using websites to display instruction material, Narrative discussions in classroom, Using white board-marker for teaching, Using mobile for communication, Using group discussion in the class, Using computers in the lab, Using any kind of software, Using the internet for giving assignments, Using internet for submitting assignments, Using local Eduwave for teaching, Physical Visit to instructor's office, Using online system for registration, Using Online examination, Displaying marks on line, Using physical handouts, Using physical instruments, Using facial expressions, movement of hands, Contacting instructor by mail, Teaching with the help of book, Student presents their work using PowerPoint and Marks. The outcome shall be the effect on student's performance due to the use of E-learning, M- learning or any other teaching method and to highlight which attributes play the main role in improving the student's performance. In its simplest form as shown in fig 3.1, we have renamed the attributes for simplifying the process of the data set as shown in fig 4.1. For Example we have assigned label 'UOE' to describe 'Using online exam' and so on.

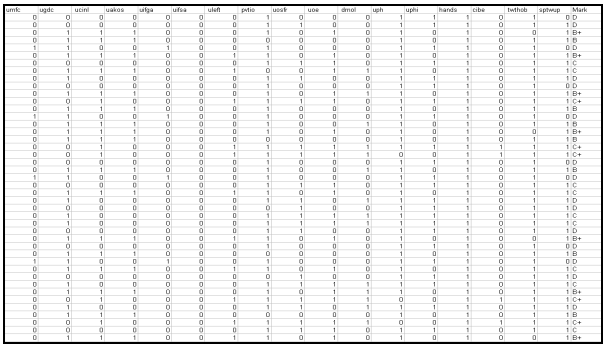

Figure 4.1 Renamed attributes of data set

We prepared description for each attribute as follows- Name, Type, Missing, Distinct and Unique; see the following for further detail.

Sample Selected Attributes:

Using Power Point

Name: Using power point Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.777
StdDev	0.418

Figure 4.2 Value description of using power point attribute

From the above figure 4.2 the Using power Point attribute has no missing values ,Two distinct values and no unique values .

Name: Using mulimedia in the class Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.245
StdDev	0.431

Figure 4.3 Value description of using multimedia in the class attribute

Name: using physical models Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.712
StdDev	0.454

Figure 4.4 Value description of using physical models attribute

Name: Using facial expressions, movement of hands Type: Numeric
Missi... 0 (0%) Distinct: 1 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	1
Mean	1
StdDev	0

Figure 4.5 Value description for using facial expression attribute

Name: Mark Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

No.	Label	Count
1	D	50
2	B+	19
3	B	20
4	C	27
5	C+	19
6	D+	4

Figure 4.5 Value description for using mark attribute

Figure 4.6 displays Histogram and it shows how often each of six values of class, Mark, occurs for each values of different attributes.

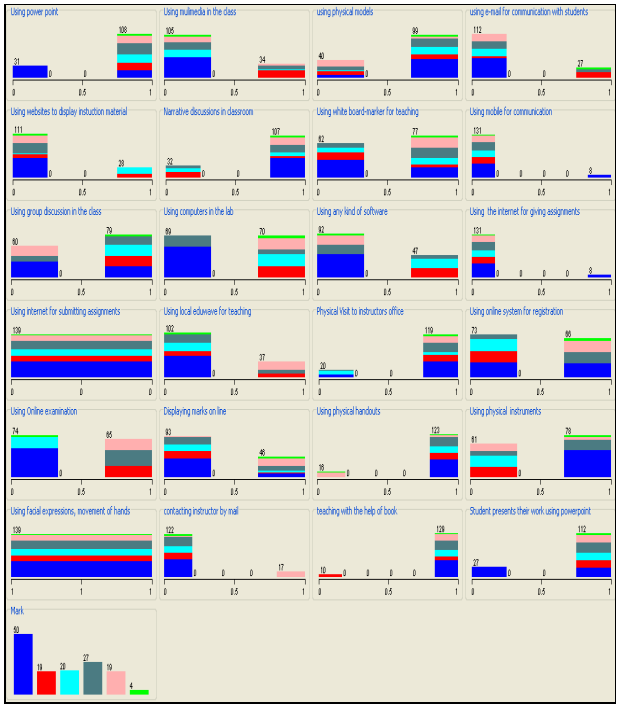

Figure 4.6 Histogram showing how often each of six values of class, Mark, occurs for each values of different attributes.

The histogram shows the distribution of the class as a function of these attributes.

We used the J4.8 algorithm to implement The C4.5 decision tree.

TABLE I

Result-

Run information :
Scheme: .classifiers.trees.J48 -C 0.25 -M 2
Relation: Survey on E- Learning
Instances: 139
Attributes: 25
Using power point
Using multimedia in the class
Using physical models
Using e-mail for communication with students
Using websites to display instruction material
Narrative discussions in classroom
Using white board-marker for teaching
Using mobile for communication
Using group discussion in the class
Using computers in the lab
Using any kind of software
Using the internet for giving assignments
Using internet for submitting assignments
Using local Eduwave for teaching
Physical Visit to instructor's office
Using online system for registration
Using Online examination
Displaying marks on line
Using physical handouts
Using physical instruments
Using facial expressions, movement of hands
Contacting instructor by mail
Teaching with the help of book
Student presents their work using PowerPoint
Mark
Test mode: 10-fold cross-validation
Classifier model (full training set)
J48 pruned tree:
Using Online examination <= 0
Using computers in the lab <= 0: D (50.0)
Using computers in the lab > 0
Using any kind of software <= 0: D+ (4.0)
Using any kind of software > 0: B (20.0)
Using online examination > 0
Contacting instructor by mail <= 0
Using any kind of software <= 0
Using multimedia in the class <= 0: C (19.0)
Using multimedia in the class > 0: C+ (3.0)
Using any kind of software > 0

```
Using physical models <= 0
  Teaching with the help of book <= 0: B+ (6.0)
  Teaching with the help of book > 0: C (11.0/3.0)
  Using physical models > 0: B+ (10.0)
  Contacting instructor by mail > 0: C+ (16.0)

Number of Leaves:          9

Size of the tree:          17
```

Stratified cross-validation :
Figure 4.7 summarize the experiment details

```
Correctly Classified Instances         131               94.2446 %
Incorrectly Classified Instances         8                5.7554 %
Kappa statistic                          0.9258
Mean absolute error                      0.021
Root mean squared error                  0.1183
Relative absolute error                  8.1055 %
Root relative squared error             32.9231 %
Total Number of Instances              139

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Clas
                 1         0          1         1         1          1        D
                 0.842     0.033      0.8       0.842     0.821      0.985    B+
                 0.95      0          1         0.95      0.974      0.975    B
                 0.852     0.027      0.885     0.852     0.868      0.989    C
                 1         0          1         1         1          1        C+
                 1         0.007      0.8       1         0.889      0.996    D+
Weighted Avg.    0.942     0.01       0.944     0.942     0.943      0.992

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 50  0  0  0  0  0 |  a = D
  0 16  0  3  0  0 |  b = B+
  0  0 19  0  0  1 |  c = B
  0  4  0 23  0  0 |  d = C
  0  0  0  0 19  0 |  e = C+
  0  0  0  0  0  4 |  f = D+
```

Figure 4.7 Summary of the result

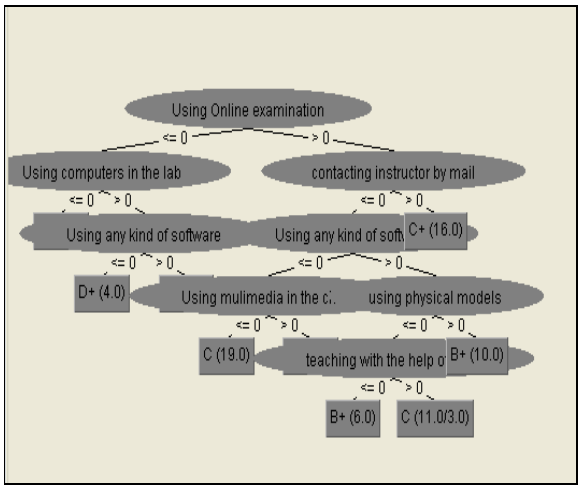

Figure 4.8 Decision Tree for experiment attributes.

To give a better presentation of the decision tree we have used acronyms for each attributes as described if Figure 4.9 .

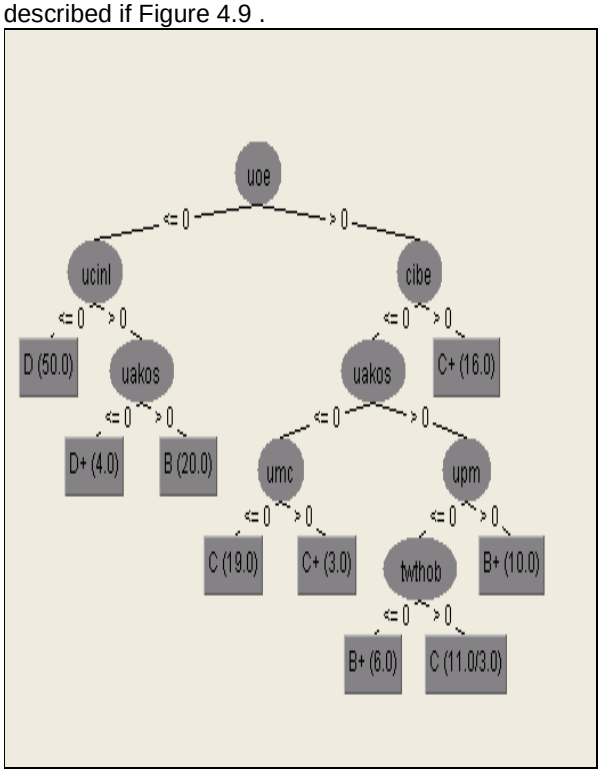

Figure 4.9 Decision tree using acronyms

X. OUTCOME/RESULTS

From the decision tree as shown above, we notice that the top node represents the entire data i.e.

139 instances and 28 attributes. The classification tree algorithm finds out that the best way to explain the dependent variable 'Mark' is by using variable 'UOE' which stands for 'Using online Examination'. Using the categories of the variable 'UOE' two different groups (under condition) were observed.If the instructor does not use 'uoe' (using online examination), next question is 'ucinl'(using computer in lab), we can conclude that:
If the instructor neither uses 'uoe'(Online Examination) nor uses 'ucinl' (Using computers in the lab), the 'mark' of the student will be 'D'.
In the same manner if we study other attributes on each level, depending on the decision tree algorithm to trace each path, we generate the following proposals and rules.

Sample of the rule – >0 = Yes
<=0 =No
If 'uoe'(using online examination)<=0 'ucinl'(Using computers in the lab) <=0 then the mark is D
If 'uoe'(using online examination)<=0 'ucinl'(Using computers in the lab) >0 and 'uakos'(using any kind of software)<=0 then the mark is D+
If 'uoe'(using online examination)<=0 'ucinl'(Using computers in the lab) >0 and 'uakos'(using any kind of software)>0 then the mark is B
If 'uoe' (using online examination) >0 and 'cibe' (contacting instructor by mail)>0 then the mark is C+
If 'uoe' (using online examination) >0 and 'cibe'(contacting instructor by mail)<=0 and 'uakos'(using any kind of software)<=0 and 'umc'(using multimedia in class)<=0 then the result is C
If 'uoe'(using online examination) >0 and 'cibe'(contacting instructor by mail)<=0 and 'uakos'(using any kind of software)<=0 and 'umc'(using multimedia in class)>0 then the result is C+
If 'uoe' (using online examination) >0 and 'cibe'(contacting instructor by mail)<=0 and 'uakos' (using any kind of software)>0 and 'upm'(using physical models)>0 then the result is B+
If 'uoe'(using online examination) >0 and 'cibe'(contacting instructor by mail)<=0 and 'uakos'(using any kind of software) >0 and 'upm'(using physical models)<=0 and 'twthob'(teaching with the help of books)>0 then the result is C
If 'uoe' (using online examination) >0 and 'cibe'(contacting instructor by mail)<=0 and 'uakos'(using any kind of software) >0 and 'upm'(using physical models)<=0 and 'twthob'(teaching with the help of books)<=0 then the result is B+

XI. CONCLUSION

This study extends the research reported by [18] and enhances our understanding of the relationship between various attributes of teaching in a number of ways. First, our decision tree confirmed that among the various pedagogies, E-learning methods play major role in enhancing the performance of the students. Our findings suggest that the E-learning methods require the consideration of a number of interrelated decisions and antecedent conditions. Successful implementation of E- educational delivery takes a commitment from both the students and the faculty and is not as simple as merely using computer in the labs or contacting the instructor by mail. Second, the study scores the fact that classes that integrate use of software and using computers while teaching even though there is no online examination yet their effect on the marks is significant. Third, the resultant mark of the student, if multimedia is used in the classroom even though no any software is being used and the students communicate with the instructor via mail and appear for on line examination, is better than the students for whom multimedia was not used. Fourthly, the marks of the students who were neither taught by the help of books in the classroom nor with the help of physical models instead with the help of any kind of software and online examination and the students communicated with the instructor on mail gave the best result in mark of the students. Overall, the findings show that teaching without the use of books gives better results and thus enhances the performance of the students. The instructor needs to develop a learning community using the E-learning pedagogy . Although this research represents a step forward in the development and evaluation of E-learning, it also raises many additional questions. How can E-learning be combined with the traditional pedagogy? More research is also needed to determine whether undergraduate and graduate business students enrolled in various courses differ in their learning needs. Research is also needed that goes beyond student perceptions to examine more quantifiable learning outcomes. Lastly, because this study was conducted at a particular institution, the generalization of these findings to others is unknown. Future research should target multiple institutions, both national and abroad. In conclusion, not only will faculty members get better E-teaching opportunity, students will also become better learners as they gain more experience with this educational medium. The end result will be improved performance of the students and overall

satisfaction of the faculty and institution. Business students will come to expect highly integrated, effective, and efficient learning experiences. Those schools unwilling to commit significant resources to the endeavor will not be competitive over time.

REFERENCES

[7] *Robert H. Beck, Spring1985, vol 35 no 2@1985 by the board of trustees of the university of Illinois, Plato's views on teaching.*

[8] *Allen, I E, and J Seaman. Staying the Course: Online Education in the United States, MA: Sloan Consorium, 2008.*

[9] *The Interdependence of the Factors Influencing the Perceived Quality of the Online Learning Experience: A Causal Model, James W. Peltier, John A. Schibrowsky and William Drago, Journal of Marketing Education 2007; 29; 140 http://jmd.sagepub.com/cgi/content/abstract/29/2/140 accessed on 20-10-2010*

[10] *Gallagher, S. (2004, March). Online distance education market update: A nascent market begins to mature. Edventures, 1-13 Journal of Marketing Education, 25, 208-217.*

[11] *Drago,W., Peltier, J. W., & Sorensen, D. (2002). Course content or instructor: Which is more important in online teaching? Management Research News, 25(6/7), 69-83.*

[12] *Abernathy, D. J. (1999). www.online.learning. Training & Development, 53(9), 36-42.*

[13] *Boumedyen, Victor V. Alexandrov ,Prof. Rafael ,"Student relationship in Higher Education using Data Mining Techniques",Global Journal of Computer Science and Technology ,Vol.10 Issue 11(Ver.1.0) October 2010 Page 71.*

[14] *Delavari N, Beikzadeh M. R. A New Model for Using Data Mining in Higher Educational System, 5th International Conference on Inforcnation Technology based Higher Education and Training: ITEHT '04, Istanbul, Turkey, 31 st May-2nd Jun 2004.*

[15] *Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery, Two Crows Corporation. Third Edition, U.S.A, 1999.*

[16] *Kalles D., Pierrakeas C., Analyzing student performance in distance learning with genetic algorithms and decision trees, Hellenic Open University, Patras, Greece,2004.*

[17] *Varapron P. et al. Using Rough Set theory for Automatic Data Analysis. 29th Congress on Science and Technology of Thailand. 2003.*

[18] *Benefits and Advantages of Online Education written by: Marina Moore; article published: year 2006, month 08; http://e-articles.info/e/a/title/Benefits-and-Advantages-of-Online-Education/ Accessed on 20-11-2010*

[19] *Han J, Kamber M. Data Mining.- Concepts and Techniques. Morgan Kaufmann Publishers ,2001.*

[20] *Mehta M, Agrawal R, Rissanen J. SLIQ: A Fast Scalable Classifier for Data*

[21] *Mining,in proc.1996 int.conf. extending database technology(EDBT'96),Avignon,*

[22] *France,Mar 1996.*

[23] *Murthy K. Automatic Construction Of Decision trees from Data : Multi-Disciplinary Survey. Siemens Corporate Research, Princeton,NJ 08540 USA.*

[24] *[16].Peng W,Chen J, Zhou H , An Implementation of ID')-Decision Tree Learning Algorithem , Universty of New South Wales,Australia.*

[25] *Allen, M., Mabry, E., Mattrey, M., Bourhis, J., Titsworth, S., & Burrell, N.(2004). Eval* SPIRAN University *uating the effectiveness of distance learning: A comparison using meta-analysis. Journal of Communication, 54, 402-420.*

[26] *Chyung, S. Y., & Vachon, M. (2005). An investigation of the satisfying and dissatisfying factors in e-learning. Performance Improvement Quarterly, 18, 97-114.*

[27] *Marks, R. B., Sibley, S. D., & Arbaugh, J. B. (2005). A structural equation model of predictors for effective online learning. Journal of Management Education, 29, 531-565.*

[28] *Peltier, J. W., Drago,W., & Schibrowsky, J. A. (2003). Virtual communitie sand the assessment of online marketing education. Journal of Marketing Education, 25, 260-276.*

[29] *Morrison, M., Sweeney, A., & Heffernan, T. (2004). Learning styles of oncampus marketing students: The challenge for marketing educators.*

[30] *Close, A. G., Dixit, A., & Malhotra, N. K. (2005). Chalkboards to cybercourses: The Internet and marketing education. Marketing Education Review, 15(2), 81-94.*

[31] *Grandzol, J. R. (2004). Teaching MBA statistics online: A pedagogically sound process approach. Journal of Education for Business, 80, 237-244.*

[32] *Hunt, L., Eagle, L., & Kitchen, P. J. (2004). Balancing marketing education and information technology: Matching needs or needing a better match? Journal of Marketing Education, 26, 75-88.*


©IJITCE PUBLICATION