International Journal of Data Engineering (IJDE)

ISSN : 2180-1274

Volume 1, Issue 5

Number of issues per year: 6

International Journal of Data Engineering (IJDE)

Volume 1, Issue 5, 2011

Edited By
Computer Science Journals
www.cscjournals.org

International Journal of Data Engineering (IJDE)

Book: 2011 Volume 1, Issue 5
Publishing Date: 08-02-2011
Proceedings
ISSN (Online): 2180 -1274

IJDE Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers

Editorial Preface

This is Second issue of volume one of the International Journal of Data Engineering (IJDE). IJDE is an International refereed journal for publication of current research in Data Engineering technologies. IJDE publishes research papers dealing primarily with the technological aspects of Data Engineering in new and emerging technologies. Publications of IJDE are beneficial for researchers, academics, scholars, advanced students, practitioners, and those seeking an update on current experience, state of the art research theories and future prospects in relation to computer science in general but specific to computer security studies. Some important topics cover by IJDE is Annotation and Data Curation, Data Engineering, Data Mining and Knowledge Discovery, Query Processing in Databases and Semantic Web etc.

This journal publishes new dissertations and state of the art research to target its readership that not only includes researchers, industrialists and scientist but also advanced students and practitioners. The aim of IJDE is to publish research which is not only technically proficient, but contains innovation or information for our international readers. In order to position IJDE as one of the top International journal in Data Engineering, a group of highly valuable and senior International scholars are serving its Editorial Board who ensures that each issue must publish qualitative research articles from International research communities relevant to Data Engineering fields.

IJDE editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJDE. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJDE provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Data Engineering (IJDE)

Editorial Board

Editor-in-Chief (EiC)

Editorial Board Members (EBMs)

Table of Content

Volume 1, Issue 5, July 2011.

Pages

Reconstruction of a Complete Dataset from an Incomplete Dataset by ARA (Attribute Relation Analysis): Some Results

Sameer S. Prabhune ssprabhune@ssgmce.ac.in

Assistant Professor & Head,
Department of Information Technology
S.S.G.M. College of Engineering,
Shegaon-444203, Maharashtra, India

Dr. S. R. Sathe srsathe@cse.vnit.ac.in

Professor,Department of E & CS,
V.N.I.T., Nagpur, Maharashtra, India

Abstract

Preprocessing is crucial steps used for variety of data warehousing and mining Real world data is noisy and can often suffer from corruptions or incomplete values that may impact the models created from the data. Accuracy of any mining algorithm greatly depends on the input data sets. Incomplete data sets have become almost ubiquitous in a wide variety of application domains. The incompleteness in the data sets may arise from a number of factors: in some cases it may simply be a reflection of certain measurements not being available at the time; in others the information may be lost due to partial system failure; or it may simply be a result of users being unwilling to specify attributes due to privacy concerns. When a significant fraction of the entries are missing in all of the attributes, it becomes very difficult to perform any kind of interpretation on the original data. And also it overall deteriorates the accuracy of any classifiers, algorithm for further analysis. For such cases, we introduce the novel idea of attribute weightage, in which we give weight to every attribute for prediction of the complete data set from incomplete data sets, on which the data mining algorithms can be directly applied. This simple but robust mechanism of weighted average gives very precise results, when tested on variety of real world datasets. This paper describes a theory and implementation of a new filter ARA (Attribute Relation Analysis) to the WEKA workbench, for finding the complete dataset from an incomplete dataset.

Keywords: Data Mining, Data Preprocessing, Missing Data

1. INTRODUCTION

Many data analysis applications such as data mining, web mining, and information retrieval system require various forms of data preparation. Mostly all this worked on the assumption that the data they worked is complete in nature, but that is not true! In data preparation, one takes the data in its raw form, removes as much as noise, redundancy and incompleteness as possible and brings out that core for further processing. Common solutions to missing data problem include the use of default [16],imputation, statistical or regression based procedures [11,15]. We note that, the missing data mechanism would rely on the fact that the attributes in a data set are not independent from one another, but that there is some predictive value from one attribute to another [1]. Therefore we used the well-known principle namely, weightage on attribute instance [11], for predicting the missing values. This paper gives the theory and implementation details of addition of an ARA filter in the WEKA workbench for estimating the missing values.

1.1 Contribution of this paper

This paper gives the theory and implementation details of ARA filter addition to the WEKA workbench. Also it gives the precise results on real datasets.

2. PRELIMENARY TOOLS KNOWLEDGE

To complete our main objective, i.e. to develop the ARA filter for the WEKA workbench we have used the following technologies. These are as follows:

2.1 WEKA 3-5-4

Weka is an excellent workbench [4] for learning about machine learning techniques. We used this tool and the package because it was completely written in java and its package gave us the ability to use **ARFF** datasets in our filter. The weka package contains many useful classes, which were required to code our filter. Some of the classes from weka package are as follows [4].

- weka.core
- weka.core.instances
- weka.filters
- weka.core.matrix.package
- weka.filters.unsupervised.attribute;
- weka.core.matrix.Matrix;
- weka.core.matrix.Eigenvalue Decomposition; etc.

We have also studied the working of a simple filter by referring to the filters available in java [9,10].

2.2 JAVA

We used java as our coding language because of two reasons:

1. As the weka workbench is completely written in java and supports the java packages, it is useful to use java as the coding language.
2. The second reason was that we could use some classes from java package and some from weka package to create the filter.

3 PSEUDO CODE

This pseudo code is designed to give the user an understanding of the ARA algorithm. ARA is a simple yet robust technique of weighted average of attribute values. [11,13].

Attribute Relation Analysis Pseudo code

Input: Dataset D with Missing attributes.

Output: Completed dataset D´ with estimated values.

ARA (I_k , J, AV^k, I_t , MA_k , PA_k , AC^k)

Where

I_K - Instant in AV^K

J - Element in AV^K

AV^K - Given attributes with Missing values

MA_K - Missing attributes

PA_k - Predictable attribute

I_t - Iteration

Step 1

Get input data:-

a. Take ARFF file as input.

b. It will be taken as one-dimensional array.

Step 2

Repeat Until (iteration > CK_j)

a. Initially all attribute and instance in iteration is null.

b. Array of instance is initially null.

c. Iteration start from first instance.

d. Initially missing values are null.

e.

Step 3

(Ai_i…….. Ai_j (MA_K) ;

Ai_i…….. Ai_j ; { Ai_i…….. Ai_j} $\neq \Omega$ [ ] ;

a. Check iteration until missing instances

b. If any missing instances is getting stop the procedure.

Step 4

$? \quad \leftarrow \quad AV_i^k \ldots\ldots ? \leftarrow \quad AV_j^k$

a. In iteration number of instances from {Ai……..Aj} so check entire instances in iteration.

b. Check any missing instance in iteration.

c. In given iteration if no missing value, so stop the procedure and print given iteration in array.

Step 5

If { correct classify (I_K.T)} ;

a. Replace the missing value or missing instance by X.
b. And existing instances by Y.

Step 6

(V[chg Num] = $AV_i^k + ... + AV_j^k$

Ω [chg Num] $\leftarrow$ { Ai_i --- Ai_J }

a. Check given instances is correctly classified or not.
b. If not correctly classify.
c. Restore instances using given dataset.

Step 7

Ai_K …. Ai_j; chg Num ++

a. Store the iteration and its changing attribute in an array.

Step 8

Iteration ++

a. Restore old value.

Step 9

if { chgnum ! = 0 }

a. If changing value contains 0, it will be replace by 0.
b. Otherwise replace it by new value.

Step 10

$$Av_i^k = \frac{\sum_{i=0}^{n}[Ai_1 \times w(Ai_n)] + [Ai_2 \times w(Ai_{n+1})] + ... + [Ai_n \times w(Ai_{n+1})]}{\sum_{i=0}^{n} w(Ai_1 + ... + Ai_n)}$$

a. Take upper sum of instances from missing instances and multiply with its weight.
b. Nearest instances get higher weight and weight will be decrease by 1 for each instance.
c. Sum of upper instances is divided by sum of weight.

Step 11

$$Av_j^k = \frac{\sum_{j=0}^{n}[Aj_1 \times w(Aj_n)] + [Aj_2 \times w(Aj_{n+1})] + ... + [Aj_n \times w(Aj_{n+1})]}{\sum_{i=0}^{n} w(Aj_1 + ... + Aj_n)}$$

a. Take lower sum of instances from missing instances and multiply with its weight.

b. Nearest instances get higher weight as compared to other and weight will be decrease by 1 for each instance.

c. Sum of lower instances is divided by sum of weight

Step12

$$\mathrm{MV}^k = \frac{AC_i^K + AC_J^K}{2}$$

Finally take average of upper instances and lower instances.

Result is replaced by new value called as predict instance in given attribute.

Step 13

After completion of one instance check next missing instance.

Procedure will be repeated until all value will be predicted.

Figure 1 Shows the ARA psedo code for prediction of the missing values.

4. IMPLEMENTATION

Coding Details

We were using datasets in ARFF format as an input to this algorithm and the ARA filter [2,7,8]. The filter would then take ARFF dataset as input and estimating out the missing values in the input dataset. After fixing out the missing values in the given dataset, it would apply the ARA algorithm and predict the missing values and also reconstruct the whole dataset from an incomplete dataset.

We have created an ARA filter class, which is an extension of the Simple Batch Filter class, which is an abstract class. Our algorithm first of all takes an ARFF format database as input then read how many attribute in given data set. It takes each attribute individually and writes it into array format. After that, it insert all instances into that array, including missing instances and find first missing instance, if it got the instance replace it by zero. After that it calculates average of all upper instances by using its weight effect on that particular instance. Nearest instance get more weight and weight will be decreases instance by instance. After that it also calculates average of all lower instances by using its weight effect on that particular instance, nearest instance get more weight and weight will be decreases instance by instance and vice-versa. Finally we are calculating the average of lower as well as upper instances.

5 EXPERIMENTAL SET UP

5.1 Approach

The objective of our experiment is to build the filter as a preprocessing step in Weka Workbench, which completes the data sets from missing data sets. We did not intentionally select those data sets in UCI

[12], which originally come with missing values because even if they do contain missing values, we don't know the accuracy of our approach. For experimental set up, we take the complete dataset from UCI repository [12], and then missing values are artificially added to the data sets to simulate MCAR missing values. To introduce m% missing values per attribute x_i in a dataset of size n, we randomly selected mn instances and replaced its x_i value with unknown i.e. ?(In WEKA, missing values are denoted as "?"). We use 10%,20% and 30% missingness for every dataset.

5.2 Results

After preprocessing steps, we use WEKA's M5Rules classifier for finding the error analysis. The classification was carried out on 10 fold cross validation technique. In Table 1, we have calculated the standard errors along with correlation coefficient on UCI[12] database repository. When observing the correlation coefficient of all the dataset i.e. CPU, Glass and Wine with missingness parameters – 10%,20% and 30%, it has been clear that, as missingness increases the accuracy of the classifier decreases.

S. N.	Error Analysis	Dataset								
		CPU			Glass			Wine		
		10% M	20% M	30% M	10% M	20% M	30% M	10% M	20% M	30% M
1	Correlation coefficient	0.88	0.66	0.47	0.62	0.59	0.62	0.78	0.78	0.75
2	Mean absolute error	36.03	53.85	60.50	0.77	0.76	0.81	144.89	140.69	147.70
3	Root mean squared error	74.03	116.22	144.70	2.31	2.39	2.18	198.67	188.66	204.41
4	Relative absolute error	43.63%	63.64%	79.63%	40.40%	39.99%	43.23%	57.76%	59.17%	62.67%
5	Root relative squared error	48.68%	75.88%	114.19%	78.06%	80.81%	79.32%	64.36%	63.10%	68.63%
6	Total Number of Instances	209	209	209	214	214	214	178	178	178

TABLE-1: After Applying the ARA Filter with M5Rules Classifier in WEKA Workbench on UCI [12] Datasets.

6. CONCLUSION

In this paper, we provided the theory and implementation details of a new filter viz. ARA in the WEKA workbench. As seen from the result, this simple but yet robust ARA filter works well to predict the missing data. We also demonstrate the efficacy of our approach by performing the analysis on real world UCI [12] data repositories. Thus it proves the extension as a preprocessing filter.

ACKNOWLEDGMENTS

Our special thanks to Mr. Peter Reutemann, of University of Waikato, fracpete@waikato.ac.nz, for providing us the support as and when required.

REFERENCES

1. S.Parthsarthy and C.C. Aggarwal, “On the Use of Conceptual Reconstruction for Mining Massively Incomplete Data Sets”,IEEE Trans. Knowledge and Data Eng., pp. 1512-1521,2003.

2. J. Quinlan, “C4.5: Programs for Machine Learning”, San Mateo, Calif.: Morgan Kaufmann, 1993.

3. http://weka.sourceforge.net/wiki/index.php/Writing_your_own_Filter

4. wekaWiki link : http://weka.sourceforge.net/wiki/index.php/Main_Page

5. S. Mehta,, S. Parthsarthy and H. Yang “Toward Unsupervised correlation preserving discretization”, IEEE Trans. Knowledge and Data Eng. pp.1174-1185 ,2005.

6. Ian H. Witten and Eibe Frank , “Data Mining: Practical Machine Learning Tools and Techniques” Second Edition, Morgan Kaufmann Publishers. ISBN: 81-312-0050-7.

7. http://weka.sourceforge.net/wiki/index.php/CVS

8. http://weka.sourceforge.net/wiki/index.php/Eclipse_3.0.x

9. weka.filters.SimpleBatchFilter

10. weka.filters.SimpleStreamFilter

11. R.J.A. Little and D. Rubin. “Statastical Analysis with Missing Data”. Ch. 3, pp-42-53,Wiley Series in Prob. and Stat., 2002.

12. UCI Machine Learning Repository, http://www.ics.uci.edu/umlearn/MLsummary.html

13. X. Zhu and X. Wu, “ Cost Constrained Data Acquisition for Intelligent Data Preparation”, IEEE Transactions on Knowledge and Data Engineering, Vol.17, Number 11, pp.1542-1556.

14. J. L. Schafer, “Analysis of Incomplete Multivariate Data”, Monographs on Stat and Applied Prob. 72, Chapman and Hall/CRC.

15. J. W. Grzymala-Busse and M.Hu. “A comparison of Several Approaches to Missing Attribute Values in Data Mining, Rough Sets and Current Trends in Computing”, 378-385, 2000.

16. C. J. Date and H. Darwen, “The Default Values approach to Missing Information,” Relational Database Writings 1989-1991, pp.343-354, 1989.

An Efficient Hybrid Successive Markov Model for Predicting Web User Usage Behavior using Web Usage Mining

V.V.R.Maheswara Rao mahesh_vvr@yahoo.com
Professor, Department of Computer Applications,
Shri Vishnu Engineering College for Women,
Bhimavaram, Andhra Pradesh, India.

Dr. V. Valli Kumari vallikumari@gmail.com
Professor, Department of CS&SE,
AU College of Engineering,
Visakhapatnam, Andhra Pradesh, India

Abstract

With the continued growth and proliferation of Web services and Web based information systems, the volumes of user data have reached astronomical proportions. Analyzing such data using Web Usage Mining can help to determine the visiting interests or needs of the web user. As web log is incremental in nature, it becomes a crucial issue to predict exactly the ways how users browse websites. It is necessary for web miners to use predictive mining techniques to filter the unwanted categories for reducing the operational scope. The first-order Markov model has low accuracy in achieving right predictions, which is why extensions to higher order models are necessary. All higher order Markov model holds the promise of achieving higher prediction accuracies, improved coverage than any single-order Markov model but holds high state space complexity. Hence a Hybrid Markov Model is required to improve the operation performance and prediction accuracy significantly.

The present paper introduces An Efficient Hybrid Successive Markov Prediction Model, HSMP. The HSMP model is initially predicts the possible wanted categories using Relevance factor, which can be used to infer the users' browsing behavior between web categories. Then predict the pages in predicted categories using techniques for intelligently combining different order Markov models so that the resulting model has low state complexity, improved prediction accuracy and retains the coverage of the all higher order Markov model. These techniques eliminates low support states, evaluates the probability distribution and estimates the error associated with each state without affecting the overall accuracy as well as protection of the resulting model. To validate the proposed prediction model, several experiments were conducted and results proven this are claimed in this paper.

Keywords: Web Usage Mining, Prediction Model, Navigation Behavior, Higher order Markov Model, Web log data, Browsing Patterns, Pre-Processing.

1. INTRODUCTION

The web has become the world's largest knowledge repository. The popularity of WWW is rapidly developing and is a golden mount with a lot of valuable information. Extracting the knowledge from the web efficiently and effectively is becoming a tedious process. Towards this, web mining has been defined as the research field focused on studying the application of data mining techniques to web data. More specifically, the field of research focused on developing techniques to model and study user web usage data has been called web usage mining. When user visits web pages, data representing their navigational experience is recorded in web log. The web log consists of an unordered, semi structured and complex of web page requests from which it is possible to accurately infer user navigational sessions, usually defined as sequence of web pages.

Web Usage Mining techniques have been proposed for mining user navigation pattern from usage data. The analysis of such patterns helps to understand the user behavior when visiting web pages. The general process of web mining process includes (I) Pre-processing: It is very important task in any mining applications is the creation of suitable target data set to which mining algorithms can be applied. The primary data preparation tasks are Data Cleansing, Page View Identification, user Identification, Sessionization, Path Completion and Data Integration. (II) Pattern discovery: The goal of pattern discovery is the task of learning some general concepts from a given set of documents. In this phase, Pattern recognition and machine learning techniques, like classification, clustering and association rule mining, are usually used on the extracted information. (IV) Pattern analysis: The goal of pattern analysis is the task of understanding, visualizing, and interpreting the patterns once they are discovered in the Pattern Discovery phase as shown in Fig.1.

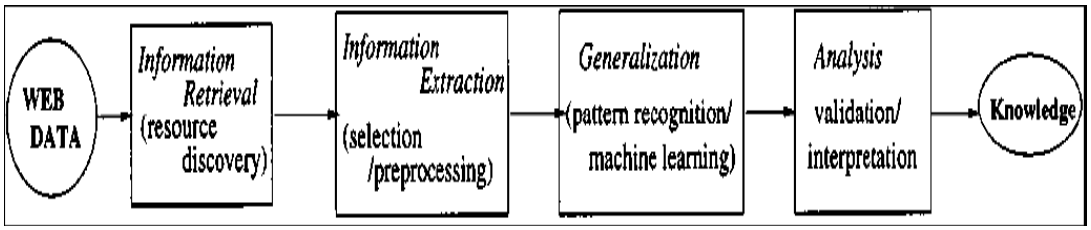


FIGURE 1: Web Mining Processes

The most recent research field focused on developing techniques to model and study users' Web navigation data. The web navigational data, which resides in weblog, consist of many more categories. In general, any web user closely associated with one or two categories. But the previous mining techniques are taking the input of all categories. This reduces operational performance of mining techniques.

Markov model have been used for studying and understanding stochastic processes, and well suited for modeling and predicting a user's browsing behavior on a web. In general, the input for these problems is the sequence of web pages that are accessed by a user and the goal is built Markov model that can be used to predict the web user usage behavior. The state space of the Markov model depends on the number of previous actions used in predicting the next action. The simplest Markov model predicts the next action by only looking at the last action performed by the user. In this model, also known as the first order Markov model, each action that can be performed by a user corresponds to a state in the model. A somewhat more complicated model computes the prediction by looking at the last two actions performed by the user. This is called the second order Markov model, and its states correspond to all possible pairs of action that can be performed in sequence. This approach is generalized to the n^{th} order Markov model, which computes the prediction by looking at the last N actions performed by the user, leading to a state space that contains all possible sequences of N actions as shown in Fig. 2.

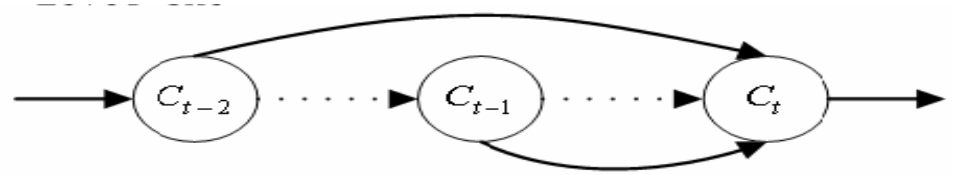


FIGURE 2: Prediction of Categories

In most of the applications, the first-order Markov model has low accuracy in achieving right predictions, which is why extensions to higher order models are necessary. All higher order Markov model holds the promise of achieving higher prediction accuracies and improved coverage than any single-order Markov model, at the expense of a dramatic increase in the state-space complexity. Hence, the authors proposes techniques for intelligently combining different order Markov models so that the resulting model has low state complexity, improved prediction accuracy and retains the coverage of the all higher order Markov model.

This paper introduces an Efficient Hybrid Successive Markov Prediction Model, HSMP. The HSMP model is initially predicts the possible required web categories using Relevance Factor, which can be determined from Similarity, Transition and Relevance Matrices to infer the users' browsing behavior between web categories. Then predict the pages in predicted categories using intelligently combining Support, Confidence and Error pruned techniques, for different order Markov models so that the resulting model has low state complexity, improved prediction accuracy and retains the coverage of the all higher order Markov model. Support-Pruned Markov model eliminates low support states without affecting the overall accuracy as well as coverage of the resulting model. Confidence-Pruned Markov Model evaluates the probability distribution of the outgoing actions before making its pruning decisions. Error pruned Markov model estimate the error associated with each state.

The rest of this paper is organized as follows: Section 2 is the related work. The Hybrid Successive Predictive Model is proposed in Section 3. Section 4 is the experimental analysis. Finally in section 5 conclusions and future work are mentioned.

2. RELATED WORK

Many of the previous authors are expressing the criticality and importance of identifying the user's browsing behavior available visiting data available in web log. Most of the works in the literature concentrates on single order Markov Model to identify the browsing behavior of the user. Several models in the literatures proposed for identifying the association between the pages without considering the category.

Myra Spiliopoulou [1] suggests applying Web usage mining to website evaluation to determine needed modifications, primarily to the site's design of page content and link structure between pages. Eirinaki et al. [2] propose a method that incorporates link analysis, such as the page rank measure, into a Markov model in order to provide Web path recommendations. Schechter et al. [3] utilized a tree-based data structure that represents the collection of paths inferred from the log data to predict the next page access. Chen and Zhang [4] utilized a Prediction by Partial Match forest that restricts the roots to popular nodes; assuming that most user sessions start in popular pages, the branches having a Non popular page as their root are pruned. R. Walpole, R. Myers and S. Myers [5] proposed Bayesian theorem can be used to predict the most possible users' next request.

There has been previous research on evaluating the ability of a Markov model to predict the next link choice of a user [14], [15]; however, there is a lack of publications on evaluating the ability of

a Markov model to represent user sessions up to a given history length. Moreover, as far as we know, the relationship between summarization ability and prediction power has not been studied before in the context of Web mining.

To extract useful browsing patterns one has to follow an approach of pre processing and discovery of the hidden patterns from possible server logs which are non scalable and impractical. Hence to reduce the operation scope there is a need of Hybrid model, which can identify the category and then the finding the association between the pages.

3. PROPOSED WORK

Because of voluminous data of web pages in Web log, the mining techniques were proved to be in efficient with respect to operation performance. To improve the operational performance of mining techniques, it is necessary to reduce the state space complexity. Towards this goal, Markov models are well suited, since they are compact, expressive and simple to understand and well established on mathematical theory. So that the resulting models has low state complexity. In the present paper the authors proposes a Hybrid Successive Markov Prediction Model HSMP using high order Markov, which can filter the unwanted categories and further it can predict the required Web pages with high prediction accuracy. Initially, the raw web log has to be preprocessed, to get formatted, integrated and actual data, which is more suitable to mining techniques as shown in Fig. 3.

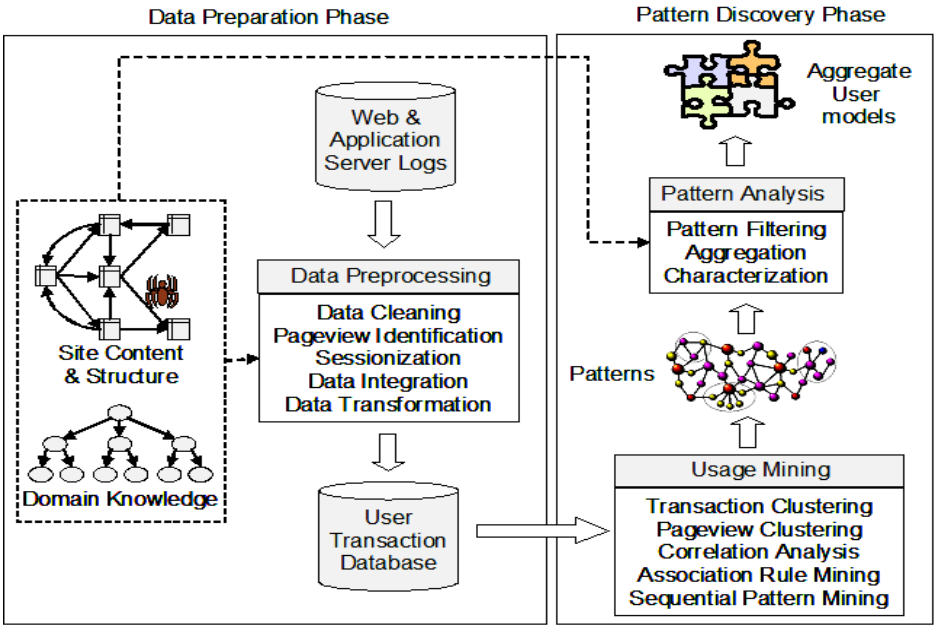


FIGURE 3: Web Usage mining Process

3.1. IMPROVED PRE-PROCESSING SYSTEM - IPS

IPS is the first stage of proposed model, which filters raw weblog data and extracts formatted, integrated and actual data which is more suitable to mining techniques. IPS is further divided into different steps which include Data Cleansing, User Identification, Session Identification, Path Completion and Data Integration. The functionality of each step explained in detail as below:

Data Cleansing

The first step of pre processing is data cleansing. It is usually site-specific, and involves tasks such as, removing extraneous references to embedded objects that may not be important for the purpose of analysis, including references to style files, graphics, or sound files as shown in Table 1. The cleansing process also may involve the removal of at least some of the data fields (e.g. number of bytes transferred or version of protocol used, etc.) that may not provide useful information in the analysis or data mining tasks.

No	Object Type	Unique Users	Requests	Bytes In	% of Total Bytes In
1	*.gif	1	46	89.00 KB	0.50%
2	*.js	1	37	753.95 KB	4.40%
3	*.aspx	1	34	397.05 KB	2.30%
4	*.png	1	31	137.67 KB	0.80%
5	*.jpg	1	20	224.72 KB	1.30%
6	Unknown	1	15	15.60 KB	0.10%
7	*.ashx	1	15	104.79 KB	0.60%
8	*.axd	1	13	274.81 KB	1.60%
9	*.css	1	8	71.78 KB	0.40%
10	*.dll	1	7	26.41 KB	0.20%
11	*.asp	1	4	1.26 KB	0.00%
12	*.html	1	3	2.17 KB	0.00%
13	*.htm	1	2	69.87 KB	0.40%
14	*.pli	1	2	24.92 KB	0.10%

TABLE1: Example of web log with different extensions

User Identification

The task of User Identification is, to identify who access web site and which pages are accessed. The analysis of Web usage does not require knowledge about a user's identity. However, it is necessary to distinguish among different users. Since a user may visit a site more than once, the server logs record multiple sessions for each user. The user activity record is used to refer to the sequence of logged activities belonging to the same user.

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:19	1.2.3.4	C	A
0:22	2.3.4.5	D	B
0:22	1.2.3.4.	A	-
0:25	1.2.3.4.	E	C
0:25	1.2.3.4.	C	A
0:33	1.2.3.4.	B	C
0:58	1.2.3.4.	D	B
1:10	1.2.3.4.	E	D
1:15	1.2.3.4.	A	-
1:16	1.2.3.4.	C	A
1:17	1.2.3.4.	F	C
1:26	1.2.3.4.	F	C
1:30	1.2.3.4.	B	A
1:36	1.2.3.4.	D	B

User 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 2

0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

User 3

0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B
1:10	1.2.3.4	E	D
1:17	1.2.3.4	F	C

FIGURE 4: Example of User Identification

Consider, for instance, the example of Fig 4. On the left, depicts a portion of a partly pre processed log file. Using a combination of IP and URL fields in the log file, one can partition the log into activity records for three separate users (depicted on the right).

Session Ordering

Sessionization is the process of segmenting the user activity record of each user into sessions, each representing a single visit to the site. Web sites without the benefit of additional

authentication information from users and without mechanisms such as embedded session ids must rely on heuristic methods for sessionization. The goal of a sessionization heuristic is to re-construct, from the click stream data, the actual sequence of actions performed by one user during one visit to the site.

Generally, sessionization heuristics fall into two basic categories: time-oriented or structure-oriented. As an example, time-oriented heuristic, h1: Total session duration may not exceed a threshold θ. Given t_0, the timestamp for the first request in a constructed session S, the request with a timestamp t is assigned to S, iff $t - t_0 \leq \theta$. In Fig 5, the heuristic h1, described above, with θ = 30 minutes has been used to partition a user activity record (from the example of Fig 3) into two separate sessions.

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

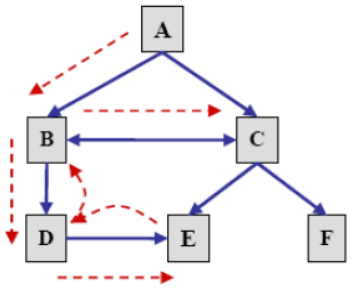
Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

FIGURE 5: Example of Sessionization

Path Completion

Another potentially important pre-processing task which is usually performed after sessionization is path completion. Path completion is a process of adding the page accesses that are not in the web log but those which is actually occurred. Client or proxy-side caching can often result in missing access references to those pages or objects that have been cached. For instance, if a user returns to a page A during the same session, the second access to A will likely result in viewing the previously downloaded version of A that was cached on the client-side, and therefore, no request is made to the server. This results in the second reference to A not being recorded on the server logs. Missing references due to caching can be heuristically inferred through path completion which relies on the knowledge of site structure and referrer information from server logs. In the case of dynamically generated pages, form-based applications using the HTTP POST method result in all or part of the user input parameter not being appended to the URL accessed by the user. A simple example of missing references is given in Fig 6.



User's actual navigation path:

A → B → D → E → D → B → C

What the server log shows:

URL	Referrer
A	--
B	A
D	B
E	D
C	B

FIGURE 6: Identifying missing references in path completion

Data Integration

The above pre-processing tasks ultimately result in a set of user sessions each corresponding to a delimited sequence of pageviews. However, in order to provide the most effective framework for pattern discovery, data from a variety of other sources must be integrated with the preprocessed

clickstream data. This is particularly the case in e-commerce applications where the integration of both user data (e.g., demographics, ratings, and purchase histories) and product attributes and categories from operational databases is critical. Such data, used in conjunction with usage data, in the mining process can allow for the discovery of important business intelligence metrics such as customer conversion ratios and lifetime values.

In addition to user and product data, e-commerce data includes various product-oriented events such as shopping cart changes, order and shipping information, impressions (when the user visits a page containing an item of interest), click through (when the user actually clicks on an item of interest in the current page), and other basic metrics primarily used for data analysis. The successful integration of these types of data requires the creation of a site-specific "event model" based on which subsets of a user's clickstream are aggregated and mapped to specific events such as the addition of a product to the shopping cart. Generally, the integrated e-commerce data is stored in the final transaction database. To enable full-featured Web analytics applications, this data is usually stored in a data warehouse called an e-commerce data mart. The e-commerce data mart is a multi-dimensional database integrating data from various sources, and at different levels of aggregation. It can provide pre-computed e-metrics along multiple dimensions, and is used as the primary data source for OLAP (Online Analytical Processing), for data visualization, and in data selection for a variety of data mining tasks

3.2. HYBRID SUCCESSIVE MARKOV PREDICTION MODEL USING HIGHER ORDER - HSMP

The Hybrid Successive Markov Predictive Model HSMP has been used for investigation and understanding stochastic process and it was to be well suited for modeling and predicting users browsing behavior in the Web log Scenario. In most of the applications, the first-order Markov model has low accuracy in achieving right predictions, which is why extensions to higher order models are necessary. All higher order Markov model holds the promise of achieving higher prediction accuracies and improved coverage than any single-order Markov model, at the expense of a dramatic increase in the state-space complexity. Hence, the authors proposes techniques for intelligently combining different order Markov models so that the resulting model has low state space complexity, improved prediction accuracy and retains the coverage of the all higher order Markov model.

The input for Hybrid Successive Markov Predictive Model HSMP is preprocessed web log, it is designed and implemented in stages, namely (A) Prediction of categories and (B) Prediction of pages in the predicted categories.

3.2. A. PREDICTION OF CATEGORIES

The steps in prediction of categories are described as follows. At first, the Similarity Matrix S of category is established. The approach of establishing similarity matrix is to gather statistics and to analyze the users' browsing behavior which can be acquired from web log data. In step two, it is to establish the first-order transition matrix P and second-order transition matrix P^2 and n-order of higher order Markov model. Secondly, the Transition Matrix of Markov is established by the same approach, statistical method, from web log. Finally, the Relevance Matrix R is computed from first-order and second-order (or n-order) transition matrix of Markov model and similarity matrix. In the proposed method, the relevance is an important factor of prediction. Relevance can be used to infer the users' browsing behavior between web categories.

It is assumed that D denotes a web log data, which contains m users' usage record. It means that the users' session is recorded and $D = \{session_1, session_2, ..., session_m\}$ is obtained. Each user's session can also be recorded as a sequential pattern of n web pages which is browsed by time order, and $session^p = \{page_1, page_2, ..., page_n\}$, where page i represents the user's visiting page at time j, is obtained. If a web site has k categories, then the user's session can be reorganized by $session^c = \{c_1, c_2, ..., c_k\}$, where $c_i = 0$. After giving the definitions, more details for the prediction model are described in the following sections.

Concept of Similarity Matrix of web categories

Step one of proposed prediction of categories model is to create the similarity matrix from web log file. At first, the situation of categories in each user's session has to be understood. The vector$_i$ = <$v_{1,i},\ldots,v_{h,i},\ldots,v_{m,i}$> for each category i is gather the i^{th} element of session c from all m user sessions, $v_{h,i}$ = 1 means user h visited web page of category i otherwise $v_{h,i}$ = 0 . Two categories can be calculated the Set similarity and Euclidean distance, respectively. Euclidean distance is further normalized. The results are computed by similarity and Euclidean distance. They are combined into a weight total similarity equation.

$$SetSim(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

Euclidean distance:

$$D(A,B) = \sqrt{\sum_{i-1}^{m}(A_i - B_i)^2} \qquad (2)$$

Normalization:

$$N(D(A,B)) = 1 - \sqrt{\frac{\sum_{i=1}^{m}(A_i - B_i)^2}{m}} \qquad (3)$$

Weight total similarity:

$$S(A,B) = SetSim(A,B).W_{ss} + N(D(A,B)).W_D \qquad (4)$$

Where $W_{ss} + W_D = 1$, $W_D = 1 - W_{ss}$

After the similarity is calculated, the similarity matrix S is a k x k matrix of categories similarity, where S_{ij} is the similarity between C_i and C_j that is established by above steps.

$$S = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{kk} \end{bmatrix} \end{array}$$

Design of Similarity Matrix of web categories

The web log pre-processing results in a set of n pageviews, P = {p_1, p_2, ⋯, p_n}, and a set of m user transactions, T = {t_1,t_2,⋯,t_m}, where each t_i in T is a subset of P. Pageviews are semantically meaningful entities to which mining tasks are applied. Each transaction t can be conceptually viewed as an l-length sequence of ordered pairs

$$t = \langle (p_1^t, w(p_1^t)), (p_2^t, w(p_2^t)), \ldots, (p_l^t, w(p_l^t)) \rangle \qquad (5)$$

where each p_i^t =for some j in {1, 2, ⋯, n}, and $w($ is the weight associated with pageviews in transaction t, representing its significance. In web usage mining tasks the weights are, either binary representing the existence or non-existence of a pageview in the transaction or they can be a function of the duration of the pageview in the user's session. In the case of time durations, it should be noted that usually the time spent by a user on the last pageview in the session is not

available. Hence for the proposed frame work weights are associated based on binary representation.

Pageviews

Sessions / users	A	B	C	D	E	F
user0	15	5	0	0	0	185
user1	0	0	32	4	0	0
user2	12	0	0	56	236	0
user3	9	47	0	0	0	134
user4	0	0	23	15	0	0
user5	17	0	0	157	69	0
user6	24	89	0	0	0	354
user7	0	0	78	27	0	0
user8	7	0	45	20	127	0
user9	0	38	57	0	0	15

FIGURE 7: An example of a Hypothetical UPVM

For association rule mining the ordering of pageviews in a transaction is not relevant, one can represent each user transaction as a vector over the n-dimensional space of pageviews. Given the transaction vector t (bold face lower case letter represents a vector) as:

$$\boldsymbol{t} = \left(w_{p_1}^{t}, w_{p_2}^{t}, \ldots, w_{p_n}^{t}\right) \qquad (6)$$

where each $w_{p_i}^{t} = w(p_j^{t})$ for some j in {1, 2, ···, n}, if $p_j$appears in the transaction t, and $w_{p_i}^{t} = 0$ otherwise. Thus, conceptually, the set of all user transactions can be viewed as an m×n User-PageView Matrix, denoted by UPVM. An example of a hypothetical user-pageview matrix is depicted in Fig. 7. In this example, the weights for each pageview is the amount of time (e.g., in seconds) that a particular user spent on the pageview.

An association or sequential pattern mining techniques can be applied on UPVM as described in example to obtain patterns and in turn these patterns are used to find important relationships among pages based on the navigational patterns of users in the site.

As noted earlier, it is also possible to integrate other sources of knowledge, such as semantic information from the content of web pages with the web usage mining process. Generally, the characteristic from the content of web pages reflects behavior of web user. Each pageview p can be represented as an r-dimensional characteristics vector, where r is the total number of extracted characteristics from the site in a global dictionary. The vector, denoted by p, can be given by:

	A.html	B.html	C.html	D.html	E.html	F.html
User1	1	0	1	0	1	0
User2	1	1	0	0	1	0
User3	0	1	1	1	0	1
User4	1	0	1	1	1	1
User5	1	1	0	0	1	0
User6	1	0	1	1	1	1

FIGURE 8: Examples of a UPVM

$$\boldsymbol{p} = \left(fw^{p}(f_1), fw^{p}(f_2), \ldots, fw^{p}(f_r)\right) \qquad (7)$$

Where $fw^{p}(f_j)$ is the weight of the j[th] characteristic in pageview p, for 1 ≤ j ≤ r. For the whole collection of pageviews in the site, one can represent an n×r PageView - Characteristic Matrix $PVCM = \{p_1, p_2, \ldots, p_n\}$.

Concept of Transition Matrix of web categories: Step two of proposed prediction of categories model is to create the transition matrix of Markov model P, which is based on web log file as well as similarity matrix. The P matrix is first-order transition matrix of Markov model and it is presented as follows:

$$P = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{21} & P_{22} & \cdots & P_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1} & P_{k2} & \cdots & P_{kk} \end{bmatrix} \end{array}$$

Each element in the P matrix presents a transition probability between any two categories. P_{ij} presents a transition probability, which is calculated between category i and category j. The numerator is the number of transition times between category i and category j, and the denominator is the total number of transition times between category i and every category k.

Design of Transition Matrix of web categories

The integration process, involve the transformation of user transactions in UPVM into "content-enhanced" transactions containing the semantic characteristics of the pageviews. The goal of such a transformation is to represent each user session as a vector of characteristics rather than as a vector over pageviews. In this way, a user's session reflects not only the pages visited, but also the significance of various characteristics that are relevant to the user's interaction. While, in practice, there are several ways to accomplish this transformation, the most direct approach involves mapping each pageview in a transaction to one or more content characteristics. The range of this mapping can be representing the set of characteristics. Conceptually, the transformation can be viewed as the multiplication of UPVM with PVCM. The result is a new matrix $TCM = \{t_1, t_2, \ldots, t_m\}$, where each t_i is a r-dimensional vector over the set of characteristics. Thus, a user transaction can be represented as a content characteristics vector, reflecting the user's interests.

As an example of content-enhanced transactions consider Fig.9 which shows a hypothetical matrix of user sessions (UPVM) as well as an index for the corresponding Web site conceptually represented as a term-pageview matrix (TPVM). Note that the transpose of this TPVM is the pageview-characteristic matrix (PVCM). The UPVM simply reflects the pages visited by users in various sessions. On the other hand, the TPVM represents the concepts that appear in each page. For simplicity the weights are assumed with binary values.

The corresponding Characteristics-Enhanced Transaction Matrix CETM (derived by multiplying the UPVM and the transpose of the TPVM) is depicted in Fig. 10.

	A.html	B.html	C.html	D.html	E.html	F.html
Web mining	0	0	1	1	1	1
Data mining	0	1	1	1	0	0
Business	0	1	1	1	0	0
Marketing	1	1	0	0	0	1
Education	1	1	0	0	1	0
Oracle Applications	1	1	0	0	1	0
ecommerce	0	1	1	0	0	1
Intelligence	1	0	1	0	0	1
DBMS Material	1	0	1	1	1	0
Information Retrieval Notes	1	0	1	1	1	1

FIGURE 9**:** Examples of TPVM

	Web mining	Data minig	Business	Marketing	Education	Oracle Applications	Ecommerce	Intelligence	DBMS Material	Information Retrieval Notes
User1	2	1	1	1	2	2	1	2	3	3
User2	1	1	1	2	3	3	1	1	2	2
User3	2	3	3	1	1	1	2	1	2	2
User4	3	2	2	1	2	2	1	2	4	4
User5	1	1	1	2	3	3	1	1	2	2
User6	3	2	2	1	2	2	1	2	4	4

FIGURE 10**:** The Characteristics-Enhanced Transaction Matrix from matrices of Fig. 9

Concept of Relevance Matrix of web categories

Step three of proposed prediction of categories model is to create the Relevance Matrix. The element R_{ij} of relevance matrix is equal to product of S_{ij} and P_{ij} , which are acquired from similarity matrix and transition matrix of Markov model respectively. In this paper, the relevance is an important factor of prediction between any two categories. The relevance can be used to infer the users' browsing behavior between categories. The relevance matrix is presented as follows:

$$R^n = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_3 \end{array} \\ \begin{bmatrix} R^n_{11} & R^n_{12} & \cdots & R^n_{1k} \\ R^n_{12} & R^n_{22} & \cdots & R^n_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ R^n_{k1} & R^n_{k2} & \cdots & R^n_{kk} \end{bmatrix} \end{array}$$

Where $R^n_{ij} = S_{ij}.P^n_{ij}.$ (8)

R^n_{ij}, presents a relevance, which is calculated, between category i at time t – n and category j at time t. More high the value of R^n_{ij} means more relevance between category i and category j.

Design of Relevance Matrix of Web Categories: The above Transition Matrix shows, the significance of various characteristics that are relevant to the users interaction. If the content features include relevance attributes associated with categories on the Web, then the discovered patterns may reveal user interests at the deeper semantic level reflected in the underlying properties of the categories that are accessed by the users on the Web. For example, that users 4 and 6 are more interested in concepts related to DBMS material and Information Retrieval notes, while user 2 is more interested in Education category. Therefore, the integration of

semantic content with Web usage mining can potentially provide a better understanding of the underlying relationships among categories.

	Web mining	Data minig	Business	Marketing	Education	Oracle Applications	E-Commerce	Intelligence	DBMS Material	Information Retrieval Notes
User1	5	5	5	4	6	6	4	4	7	7
User2	4	3	3	5	8	8	3	4	7	7
User3	9	8	8	5	8	8	5	6	12	12
User4	11	9	9	6	10	10	6	8	15	15
User5	4	3	3	5	8	8	3	4	7	7
User6	11	9	9	6	10	10	6	8	15	15

FIGURE 11**:** Relevance Matrix

3.2. B. PREDICTION OF PAGES IN THE PREDICTED CATEGORIES

A user navigation session within predicted category can be represented by the sequence of pages requested by the user. First-order Markov models have been widely used to model a collection of user sessions. In such context, each Web page in the category corresponds to a state in the model, and each pair of pages viewed in sequence corresponds to a state transition in the model. A transition probability is estimated by the ratio of the number of times the transition was traversed to the number of times the first state in the pair was visited. Usually, artificial states are appended to every navigation session to denote the start and finish of the session.

A first-order Markov model is a compact way of representing a collection of sessions, but in most cases, its accuracy is low, which is why extensions to higher order models are necessary. In a higher order Markov model, a state corresponds to a fixed sequence of pages, and a transition between states represents a higher order conditional probability. For example, in a second-order model, each state corresponds to a sequence of two page views. The serious drawback of higher order Markov models is their exponentially large state space compared to lower order models.

A Hybrid Successive Markov Prediction Model, HSMP is a model extension that allows variable length history to be captured. By the refined analysis on web page content of HSMP, it is possible to predict the association among the pages, and it can able to summarize web user usage behavior accurately. For example, each state definition can include a vector of keywords representing the contents of the corresponding Web page. As a result, it would be possible to identify high probability trails that are composed of pages that are relevant to a given topic. The proposed model illustrates sequence of increasing order Markov models with an example as below.

session	*freq.*
S,1,2,3,4,F	2
S,1,3,5,F	7
S,1,3,4,F	2
S,2,3,5,6,F	2
S,2,3,4,F	3
S,4,6,7,F	1
S,4,6,F	5
S,4,F	6
S,5,6,7,F	4
S,5,6,F	1
S,2,3,4,6,F	3
S,2,3,F	2

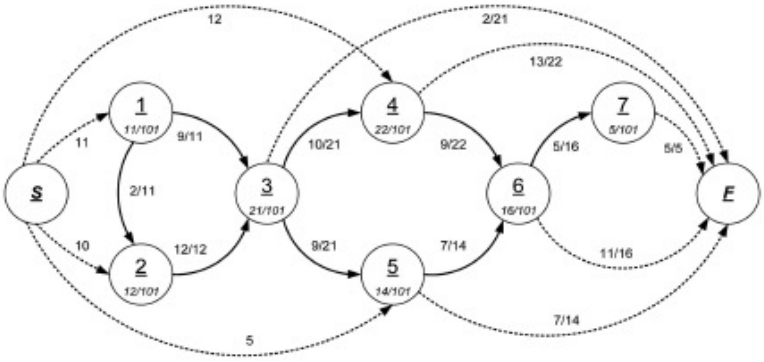


FIGURE 12: a. Sessions b. First-Order Model

First-Order HSMP Model Construction

Fig. 12.a shows an example of a collection of navigation sessions in category. The session start and finish at an artificial state; freq. denotes the number of times the corresponding sequence of pages was visited. Fig. 12.b presents the first-order model for these sessions. There is a state corresponding to each Web page and a link connecting every two pages viewed in sequence. For each state that corresponds to a Web page, give the page identifier and the number of times the page was viewed divided by the total number of page views. This ratio is a probability estimate for a user choosing the corresponding page from the set of all pages in the category. For example, page 4 has 22 page views from a total of 101 page views. For each link, indicates the proportion of times it was followed after viewing the anchor page. For example, page 5 was viewed 14 times, five of which were at the beginning of a navigation session. After viewing page 5, the user moved to page 6 in 7 of the 14 times and terminated the session seven times. The probability estimate of a trail is given by the product of the probability of the first state in the trail (that is, the initial probability) and the probabilities of the traversed links (that is, the transition probabilities). For example, the probability estimate for trail (3,4) is 21/101 * 10/21 = 0.099, and for trail (1, 3, 5), it is 11/101 * 9/12 * 9/21 = 0.035.

Higher Order HSMP Model Construction

The first-order model does not accurately represent all second-order conditional probabilities. For example, according to the input data, the sequence (1, 3) was followed nine times, that is, #(1, 3) = 9, and sequence (1, 3, 4) was followed twice, that is, #(1, 3, 4) = 2. Therefore, the probability estimate for viewing page 4 after viewing 1 and 3 in sequence is p(4|1,3) = #(1, 3, 4) / #(1, 3) = 9. The error of a first-order model in representing second-order probabilities can be measured by the absolute difference between the corresponding first and second-order probabilities. For example, for state 3, |p(4|1,3) – p(4|3)| = |2/9 – 10/21 | = 0.254 and |p(4|2,3) – p(4|3)| = |8/12 – 10/21| = 0.190. Thus, state 3 is not accurately representing second-order conditional probabilities. The accuracy of transition probabilities from a state can be increased by separating the in-paths to it that correspond to different conditional probabilities. To increase the accuracy in the example by cloning state 3 (that is, creating a duplicate state 3^1) and redirecting the link (2, 3) to state 3^1. The weights of the out-links from states 3 and 3^1 are updated according to the number of times the sequence of three states was followed.

For example, since #(1, 3, 4) = 2 and #(1, 3, 5) = 7 in the second-order model, the weight of the link (3, 4) is 2 and the weight of (3, 5) is 7. The same method is applied to update the out-links from the clone state 3^1. Fig. 13 shows the resulting second-order model after cloning four states in order to accurately represent all second-order conditional probabilities.

In the extended model given in Fig. 13, all the out-links represent accurate second-order probability estimates. The probability estimate of the trail (1, 3, 5) is now 11/101 * 9/11 * 7/9 = 0.069. The probability estimate for trail (3, 4) is (9/101 * 2/9) + (12/101 * 8/12) = 0.099, which is equal to the first-order estimate. Therefore, the second-order model accurately models the conditional second-order probability estimates while keeping the correct first-order probability estimates.

In order to provide control over the number of additional states created by the method, use of a parameter β that sets the highest admissible difference between a first-order and the corresponding second-order probability estimate. In a first-order model, a state is cloned if there is a second-order probability whose difference from the corresponding first-order probability is greater than β. Alternatively, to interpret β as a threshold for the average difference between the first-order and the corresponding second-order probabilities for a given state. In the later, the state is cloned if the average difference between the first and second-order conditional probabilities surpasses β. Moreover, if to set β > 0 and the state has three or more in-links, use of the Associations rules to identify in-links inducing identical conditional probabilities. When β is measuring the maximum probability of divergence, denote it by β_m, and when it is measuring the average probability divergence, denote it by β_a.

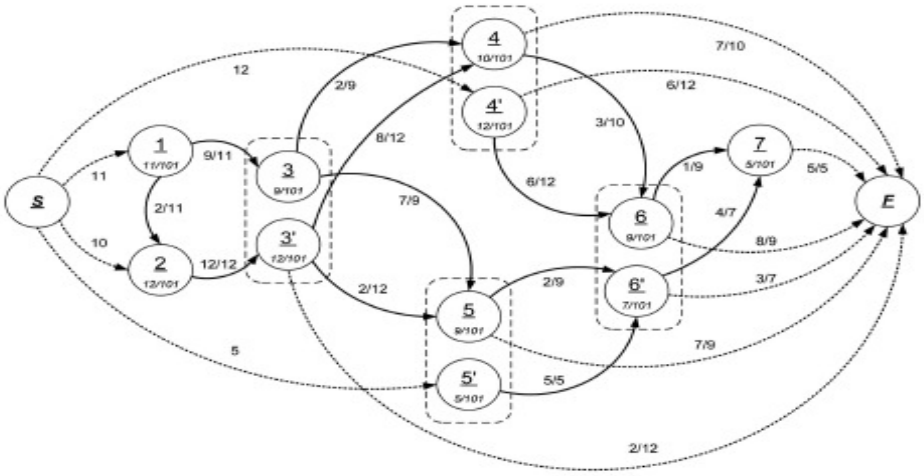


FIGURE 13**:** Second-Order Model

The method to extend a model to higher orders is identical. N-order conditional probability estimates are compared to the corresponding lower order estimates, and cloning is applied to states that are not accurate in order to separate their n-state length in-paths. Experimental results showing that the running time is approximately linear with respect to the order of the model. The above two example shows that construction of first-order and second-order Markov model, one can construct the higher Markov model in the similar way at the expense of dramatic increase in state space complexity. To overcome this problem pruning models are introduced as below.

3.2. C. PRUNED MARKOV MODELS

As discussed in the previous section, all high order Markov model holds the promise of achieving higher prediction accuracies and improved coverage than any single-order Markov model, at the expense of a dramatic increase in the state-space complexity. To overcome, develop techniques for intelligently combining different order Markov models so that the resulting model has low state complexity, improved prediction accuracy and retains the coverage of the all high order Markov model. Based on this observation, to start from the all high order Markov model and eliminate many of its states that are expected to have low prediction accuracy. This will allow reducing the overall state complexity without affecting the performance. The goals of this pruning step are primarily to reduce the state complexity and secondarily improve the prediction accuracy of the resulting model.

Here the authors present three different schemes with an increasing level of complexity. The first scheme simply eliminates the states that have very low support. The second scheme uses statistical techniques to identify states for which the transition probabilities to the two most prominent actions are not statistically significant. Finally, the third scheme uses an error-based pruning approach to eliminate states with low prediction accuracy.

Support-Pruned Markov Model

The Support-Pruned Markov Model (SPMM) is based on the observation that states that have low support and low prediction accuracies. Consequently, these low support states can be eliminated without affecting the overall accuracy as well as coverage of the resulting model. The amount of pruning in the SPMM scheme is controlled by the parameter Φ referred to as the frequency threshold. In particular, SPMM eliminates all the states of the different order Markov models that are supported by fewer than Φ instances.

First, the same frequency threshold is used for all the models regardless of their order. Second, this pruning policy is more likely to prune higher-order states as higher order states have less support; thus dramatically reducing the state-space complexity of the resulting scheme. Third, the

frequency threshold parameter Φ specifies the actual number of instances that must be supported by each state and not the fraction of instances as it is often done in the context of association rule discovery. This is done primarily for the following two reasons: (i) the trust-worthiness of the estimated transition probabilities of a particular state depend on the actual number of instances and not on the relative number; (ii) the total number of instances is in general exponential on the order of the Markov model, thus the same fractional pruning threshold will have a completely different meaning for the different order Markov models.

Confidence-Pruned Markov Model

One of the limitations of the SPMM scheme is that it does not capture all the parameters that influence the accuracy of the state. In particular the probability distribution of outgoing actions from a state is completely ignored. For example, consider a Markov state which has two outgoing actions/branches, such that one of them is substantially more probable than the other. Even if the overall support of this state is somewhat low, the predictions computed by this state will be quite reliable because of the clear difference in the outgoing probabilities. On the other hand, if the outgoing probabilities in the above example are very close to each other, then in order for that difference to be reliable, they must be based on a large number of instances. Ideally, the pruning scheme not only considers the support of the state but also weigh the probability distribution of the outgoing actions before making its pruning decisions.

This observation is to develop the confidence-pruned Markov model (CPMM) scheme. CPMM uses statistical techniques to determine for each state, if the probability of the most frequently taken action is significantly different from the probabilities of the other actions that can be performed from this state. If the probability differences are not significant, then this state is unlikely to give high accuracy and it is pruned. In contrast, if the probability differences are significant the state is retained.

The CPMM scheme determines if the most probable action is significantly different than the second most probable action by computing the 100(1-α) percent confidence interval around the most probable action and checking if the probability of the second action falls within that interval. If this is true, then the state is pruned, otherwise it is retained. If $\hat{p}$ is the probability of the most probable action, then its 100(1- α) percent confidence interval is given by

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (9)$$

where $z_{\alpha/2} = 2$ is the upper α/2 percentage point of the standard normal distribution and n is the frequency of the Markov State.

The degree of pruning in CPMM is controlled by α (confidence coefficient). As the value of α decrease the size of the confidence interval increases, resulting in more pruning. Also note that if a state has a large number of examples associated with it, then Equation 9 will compute a tighter confidence interval. As a result, even if the difference in the probabilities between the two most probable actions is relatively small, the state will most likely be retained.

Error-Pruned Markov Model

In the previous schemes either the support of a state or the probability distribution of its outgoing branches to gauge the potential error associated with it. However, the error of each state can be also automatically estimated and used to decide whether or not to prune a particular state. A widely used approach to estimate the error associated with each state is to perform a validation step. During the validation step, the entire model is tested using validation set that was not used during the model building phase. Since the actual actions performed by the sequences in the validation set, can easily determine the error-rates and use them for pruning. To develop the Error Pruned Markov Model (EPMM) scheme, two different error-based pruning strategies that use a different definition as to what constitutes the error-rate of a Markov state. To refer these schemes as overall error pruning and individual error pruning.

First, for each sequence in the validation set use each one of the single-order Markov models to make a prediction. Record each prediction whether that is correct or not. Once all the sequences in the validation set have been predicted, use these prediction statistics to calculate the error-rate of each state. Next for each state of the highest-order Markov model, identify the set of states in the lower-order models that are its proper subsets. For example, if the higher-order state corresponds to the action-sequence $\{a_5, a_3, a_6, a_7\}$, then the lower-order states that are identified are $\{a_3, a_6, a_7\}$ (third-order), $\{a_6, a_7\}$ (second-order) and $\{a_7\}$ (first-order). Now if the error-rate of the higher-order state is higher than any of its subset lower-order sates, it is pruned. The same procedure of identifying the subset sates and comparing their error-rates is repeated for all the states in the lower-order Markov models as well, except the first-order Markov model. The states form the first-order Markov model is never pruned so as not to reduce the coverage of the resulting model.

In the second scheme, at first iterate over all the higher-order states, and for each of them find its subset states (as described in the previous scheme). Then, identify all the examples in the validation set that can be predicted using the higher-order state (i.e., the validation examples which have a sequence of actions corresponding to the higher-order state). This set of examples is then predicted by the higher-order state and its subset states and the error-rates on these examples for each one of the states are computed. If the error-rate of the higher-order state is greater than any of its subset states, the higher-order Markov state is pruned. The same procedure is repeated for all the lower-order Markov models except the first-order Markov model.

Though both schemes follow a similar procedure of locating subset states and pruning the ones having high error rates, they differ on how the error-rates for each state are computed. In the first scheme, every lower-order Markov state has a single error-rate value that is computed over the entire validation set. In the second scheme, each of the lower-order Markov states will have many error-rate values as it will be validated against a different set of examples for each one of its superset higher-order states.

These techniques introduced by the authors, combining intelligently different order Markov models reduce state complexity and improves prediction accuracy. As a result HSMP have the high operational performance towards predicting the web user usage behavior.

4. EXPERIMENTAL ANALYSIS

The server side web log data is experimented over a period of six months under standard execution environment. In the preprocessing stage erroneous and image requests were eliminated, set a session length limit of requests and therefore very long sessions were split in to two or more shorter session. The total size of 238 requests made by a single user is 2.14 MB. Out of that the size of requests like .gif, .jpg, .css, .dll and so on is 1.3 MB (61%). Hence cleansing is an important phase in the process of pre processing and reduces the human user accesses web log by 60% approximately.

From the collection of sessions, a first-order model was inferred. This model was then evaluated for second and higher order conditional probabilities and, if needed, a state was cloned to separate the in-paths due to differences in the conditional probabilities. As described above, the β parameter sets the tolerance allowed on representing the conditional probabilities. In addition, there is a parameter that specifies the minimum number of times a page has to be requested in order to be considered for cloning. In these experiments, we set num visits ≥ 30. Fig. 14 shows the variation of the model number of states when the order of the model increases while having the accuracy threshold set to $\beta = 0$.

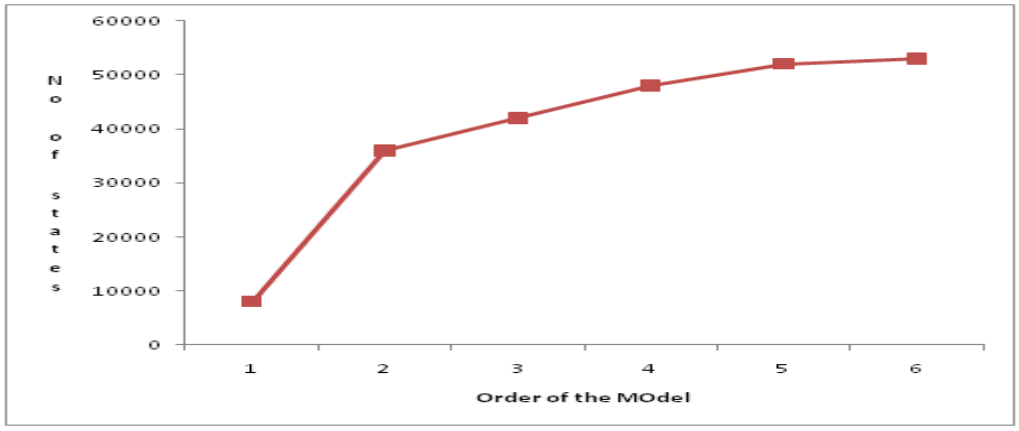


FIGURE 14**:** Performance comparison of Different order Markov Models w.r.t. No. of States

A) The HSMP model compared with the 1st, 2nd, 3rd and 4th order Markov Models with respective number of states. The experimental results indicate that the higher order Markov Models, the number of states increases at a slower rate, which is an indication of gain in accuracy as shown in Fig 14.

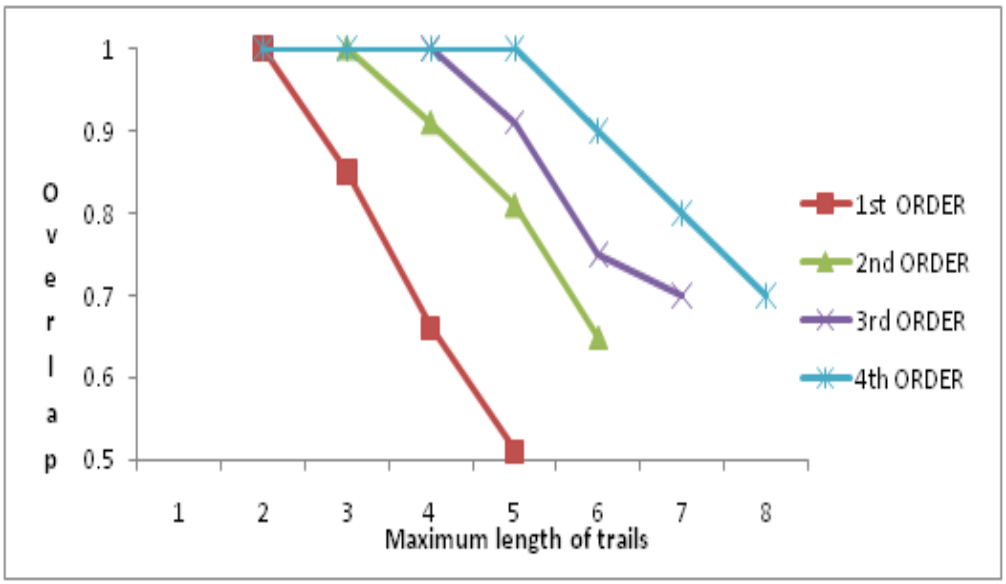


FIGURE 15**:** . Performance comparison of Different order Markov Models w.r.t. Overlap

B) The HSMP model compared with the 1st, 2nd, 3rd and 4th order Markov Models with respective overlap. The experimental results indicate closed to linear decrease as length of trails increases. as shown in Fig 15.

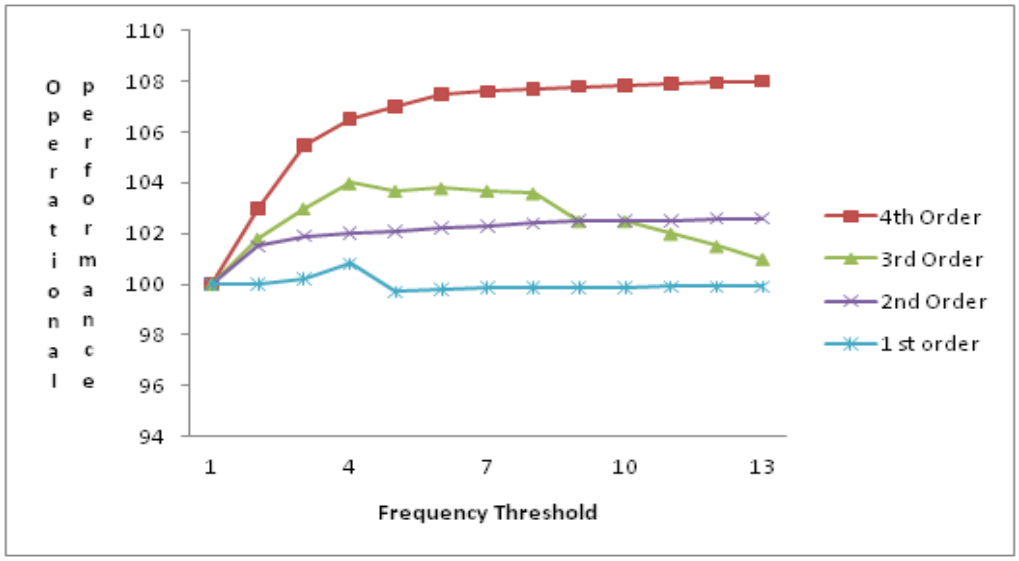


FIGURE16**:** Performance comparison of different order Markov model w.r.t. Frequency Thershold

C) The HSMP model compared with the 1st, 2nd, 3rd and 4th order Markov Models with respective frequency threshold. The experimental results indicate that noticeable improvement of HSMP operational performance over the low order Markov Models as shown in Fig 16.

D) In addition, the standard analysis algorithms are applied on the collective output (desired patterns) generated by HSMP, the web user usage interests are identified as shown in Fig 17.

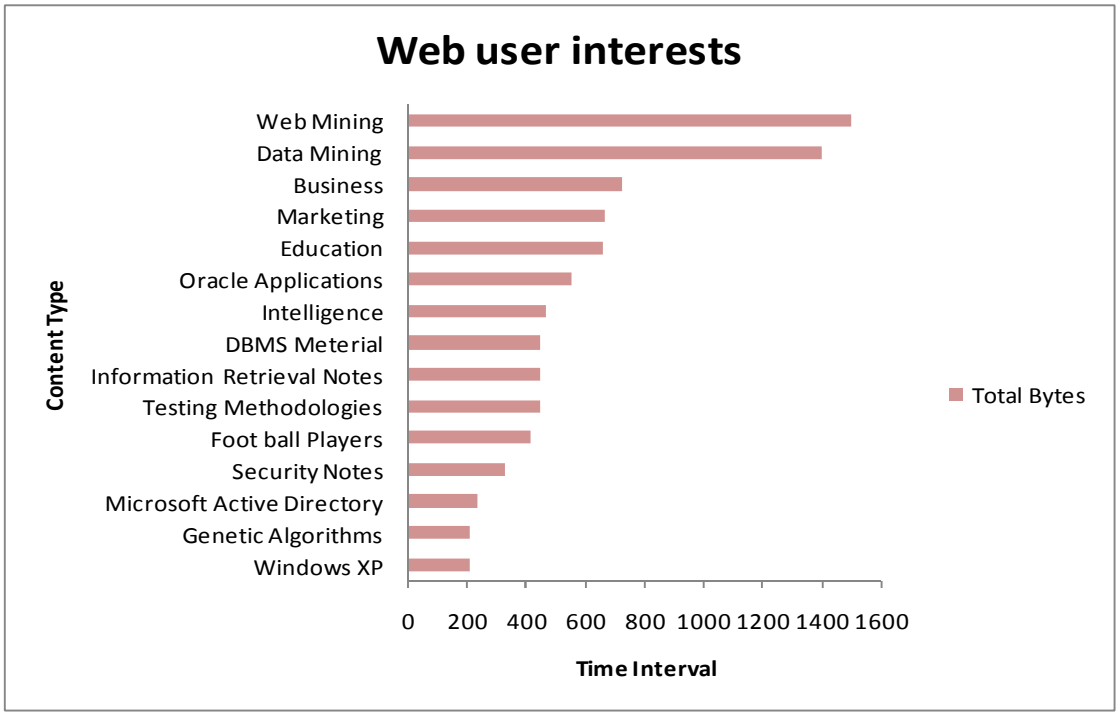


FIGURE 17**:** Web User Usage interests

5. CONSLUSION & FUTURE WORK

Because of the huge quantity of data of web pages on many portal sites, for convenience, are to assemble the web page based on category. In this paper, users' browsing behavior will be predicted at two levels to meet the nature of the navigation. One is category stage and the other is web page stage. In stage one is to predict category. The unnecessary categories can be excluded. The scope of calculation is massively reduced. Next, using pruned Markov models using higher order in the level two to predict the users' browsing page is more effectively and high operational performance. The results of experiment prove the low state complexity and predictive power is well in both stages. Even though these techniques are developed in the context of web usage data, we have successfully used these techniques for prediction in different applications, as well.

6. REFERENCES

1. M. Spiliopoulou, "Web Usage Mining for Site Evaluation," Comm. ACM, vol. 43, no. 8, 2000, pp. 127–134.

2. M. Eirinaki, M. Vazirgiannis, and D. Kapogiannis, "Web Path Recommendations Based on Page Ranking and Markov Models," Proc. Seventh Ann. ACM Int'l Workshop Web Information and Data Management (WIDM '05), pp. 2-9, 2005.

3. S. Schechter, M. Krishnan, and M. Smith, Using Path Profiles to Predict HTTP Requests," Computer Networks and ISDN Systems, vol. 30, pp. 457-467, 1998.

4. X. Chen and X. Zhang, "A Popularity-Based reduction Model for Web Pre fetching," Computer, pp. 63-70, 2003.

5. R. Walpole, R. Myers, S. Myers and K. Ye, "Probability and Statistics for Engineers and Scientists," in Paperback, 7 ed., Pearson Education, 2002, pp.82-87.

6. R. M. Suresh and R. Padmajavalli," An Overview of Data Preprocessing in Data and Web Usage Mining," Digital Information Management IEEE, pp. 193-198, 2006.

7. Yan Wang,Web Mining and Knowledge Discovery of Usage Patterns, February, 2000.

8. R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the World Wide Web. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997.

9. M.-S. Chen, J.S. Park, and P.S. Yu., "Efficient Data Mining for Path Traversal Patterns," IEEE Trans. Knowledge and Data Eng., vol. 10, no. 2, 1998.

10. M. S. Chen, J. S. Park and P. S. Yu, "Efficient Data Mining for Path Traversal Patterns in a Web Environment", IEEE Transaction on Knowledge and Data Engineering, 1998.

11. Wang Shachi, Zhao Chengmou. Study on Enterprise Competitive Intelligence System.[J]. Science Technology and Industrial, 2005.

12. Tsuyoshi, M and Saito, K. Extracting User's Interest for Web Log Data. Proceeding of IEEE/ACM/WIC International Conference on Web Intelligence (WI'06), 2006.

13. UCI KDD archive, http://kdd.ics.uci.edu/.

14. M. Deshpande and G. Karypis, "Selective Markov Models for Predicting Web Page Accesses," ACM Trans. Internet Technology, vol. 4, pp. 163-184, May 2004.

15. J. Borges and M. Levene, "Testing the Predictive Power of Variable History Web Usage," J. Soft Computing, special issue on Web intelligence, 2006.

16. Proceedings of the IEEE International Conference on Tools with Artificial Intelligence, 1999.

17. M. S. Chen, J. S. Park and P. S. Yu, "Efficient Data Mining for Path Traversal Patterns in a Web Environment", IEEE Transaction on Knowledge and Data Engineering, 1998.

18. M. Craven, S. Slattery and K. Nigam, "First-Order Learning for Web Mining", In Proceedings of the 10th European Conference on Machine Learning, Chemnitz, 1998.

19. Tsuyoshi, M and Saito, K. Extracting User's Interest for Web Log Data. Proceeding of IEEE/ACM/WIC International Conference on Web Intelligence (WI'06), 2006.

Performance Assessment of Faculties of Management Discipline From Student Perspective Using Statistical and Mining Methodologies

Chandrani Singh singh.chandrani@gmail.com
Associate Professor, MCA Department
Sinhgad Institute of Business Administration and Research
Pune, Maharashtra - 411048

Arpita Gopal arpita.gopal@gmail.com
Director,MCA Department
Sinhgad Institute of Business Administration and Research
Pune, Maharashtra-411048

Santosh Mishra ssantosh.k.mishra@gmail.com
Sinhgad Institute of Business Administration
and Research
Pune,Maharashtra-411048

Abstract

This paper deals with Faculty Performance Assessment from student perspective using Statistical Analysis and Mining techniques .Performance of a faculty depends on a number of parameters (77 parameters as identified) and the performance assessment of a faculty/faculties are broadly carried out by the Management Body ,the Student Community ,Self and Peer faculties of the organization .The parameters act as performance indicators for an individual and group and subsequently can impact on the decision making of the stakeholders. The idea proposed in this research is to perform an analysis of faculty performance considering student feedback which can directly or indirectly impact management's decision, teaching standards and norms set by the educational institute, understand certain patterns of faculty motivation, satisfaction, growth and decline in future. The analysis depends on many factors, encompassing student's feedback, organizational feedback, institutional support in terms of finance, administration, research activity etc. The statistical analysis and mining methodology used for extracting useful patterns from the institutional database has been used to extract certain trends in faculty performance when assessed on student feedback. The paper compares first the traditional approach with the statistical approach and then justifies the usage of data mining classification technique for deriving the results.

Keywords: Data Analysis, Mining, Clustering, Trend Extraction, Performance Prediction

1. INTRODUCTION

The applications of Data Mining in the field of higher education can truly be justified because typical type of data mining questions used in the business world has counter part questions relevant to higher education [2]. The need of Data Analysis and Mining in higher education is to mine faculty and students data from various stakeholders' perspective [7]. The methodology adapted to design the system is dealt extensively in the previous paper [16].Initially 77 parameters were considered,50 faculties performance was assessed based on the feedback obtained from various segments and averaged out to show the mean performance of Faculties using traditional approach. The ongoing research on Faculty Assessment has enabled us to increase our data size and implement segment slicing. The result generated in this paper is

strictly from student feedback. Around 3000 student records were taken into consideration. The data was smoothened and profiled, inconsistent data was removed and the operational data included two consecutive years student feedback from two institute's of management discipline. This data was then analyzed using conventional MS-Excel and the following pattern was derived as shown in Figure 1.

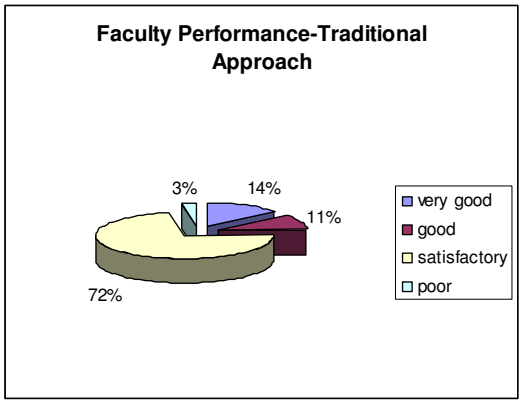


FIGURE 1: Faculty Performance – Traditional Approach

The accuracy of the result then was taken into rigorous consideration because the influence of the other parameters on the faculty performance was missing considerably. The justification of implementing statistical analysis and mining algorithms was required to extract intelligent information and to perform complex calculations, trend analysis and sophisticated data modeling, and reporting. The need was to identify critical information on the not so obvious data and extract mission critical information and intelligence that would enable better decision by the academia. This is an ongoing research work so comparative evaluation is behind the scope of this paper since similar work has been performed only on monitoring student academic performance using data mining technique.

2. DATA ANALYSIS AND MINING RATIONALE

The goal of higher education is to continually maintain quality and standards with the most efficient procedures implemented for growth and the degree of quality teaching involves the pertinent issues of how to enhance and evaluate it through overt and covert processes. Hence the Data Mining processes for knowledge discovery is to subject various classification and prediction procedures on the data. This helps institutes to predict certain trends of faculties in terms of intellectual contribution, administrative services, and standards followed which cannot be meted out using traditional approach.

3. CLASSIFICATION AND CLUSTERING

The classifier model used was the full training set and ZeroR algorithm was used to predict the classified instances. The results of classification are as shown in Table 2. Initially incorrectly classified instances was found to be around 28 % hence the data was again profiled to increase the percentage of correctly classified instances.

=== Classifier model (full training set) ===	
ZeroR predicts class value: satisfactory	
Time taken to build model: 0.02 seconds	
=== Evaluation on training set ===	
=== Summary ===	
Correctly Classified Instances	2890
Kappa statistic	0
Mean absolute error	0.0254
Root mean squared error	0.112
Relative absolute error	100%
Root relative squared error	100 %
Total Number of Instances	2890

TABLE 2: Classifier model (full training set)

Then clustering of the correctly classified data was performed using EM algorithm where clusters were generated based on the parameter value and for every parameter cluster the percentage of ratings were found out as shown in the table 4 and then the cumulative value was averaged out to find out the mean and the ratings were represented using percentages. A snapshot of the cluster formation which is an intermediate process is also shown in the table 3.

Row Id.		Cluster id	Dist clust-1	Dist clust-2	Subject_ Knowledge	Teaching Ability_with_use_of_new_Teaching_Aids	Motivation_Self_Students	Aggre_per
84		1	1.1136	4.3082	8	8	4	40
85		1	1.9408	2.024	9	9	5	43
86		2	3.5222	0.27612	10	10	5	48
87		2	3.5222	0.27612	10	10	5	48
88		2	3.2658	0.81656	10	10	5	45
89		2	2.9407	1.901	10	8	5	47
90		2	3.7414	0.3378	10	10	5	50
91		2	3.7414	0.3378	10	10	5	50
92		2	3.7414	0.3378	10	10	5	50

TABLE 3: A snapshot of the intermediate cluster generation process

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
very good	2.839	6.0199	59.6441	218.529	50.2789	16.6368	2.0933	45.8842
Good	67.6326	11.8359	17.4102	2.0227	45.9184	37.7739	10.66	56.7013
Satisfactory	636.33	4.9919	115.724	1.7551	195.544	312.472	148.407	223.753
Poor	7.6643	6.0078	1.0183	1.3386	3.4087	8.7138	14.303	3.3781

Cluster 9	Cluster 10	Cluster 11	Cluster 12	Cluster 13	Cluster 14	Average
1	1.9775	40.0564	10.0671	1.0022	5.9714	32.99999
3.0049	4.0798	39.2667	40.6924	4.997	1.0043	24.50001
44.8842	8.9885	180.93	183.253	6.017	6.9517	147.8572
9.3508	2.663	1.0011	10.1516	1.0006	1.0005	5.071443
58.2399	17.7088	261.254	244.164	13.0168	14.9279	
[total]	714.466	28.8554	193.796	223.646	295.15	375.597	175.463	329.717

Table 4: Clustered data of Faculty Performance based on student feedback

The ratings were then represented using pie chart which is shown in the figure below and the representation reveals more accuracy than the traditional approach because the clusters generated are influenced heavily by all the attributes which had been taken into consideration.

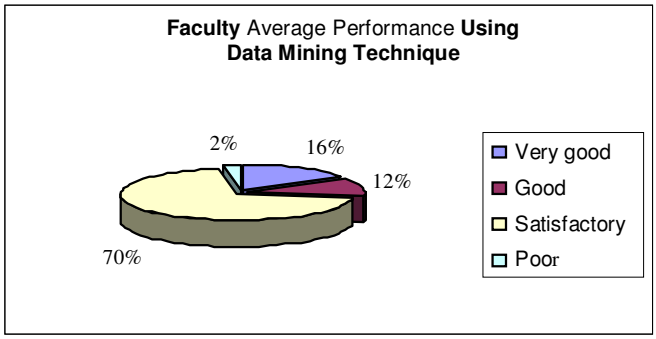


FIGURE 2: Faculty Average Performance Using Data Mining Technique

4. FACULTY PERFORMANCE TREND EXTRACTION USING OLAP STATISTICAL TOOL

NAME OF FACULTY	Arpita Gopal	Chandrani Singh	Netra Patil		Amresh Nikam		Gauri Jha		Rizwana Khan	Gen. Deepak Azwani	Uday Oak	Meghraj Chakur		Rubina Sheikh	Sonali Patil	
TEACHING Semester	5th	5th	5th		5th		5th		5th	5th	5th	5th		5th	5th	
TEACHING Subject	Advance Unix Programming	Software Project Management	Advance Internet Technologies	Mini Project	Human Computer Interface	Mini Project	Emerging Trendes In Information Technology	AUP Lab	Technical Trainning	Soft Skills	Appitude Trainning	ADT Lab	.Net	Appitude Trainning	Technical Trainning	AUP Lab
WORK EXPERIENCE-TEACHING	10	12	9	9	7	7	1	1	5	5	5	2.8	2.8	8	8.2	8.2
RESEARCH PUBLICATIONS-INTERNATIONAL	12	5	5	5	12	12	0	0	2	0	0	0	0	0	6	6
DESIGNATION																

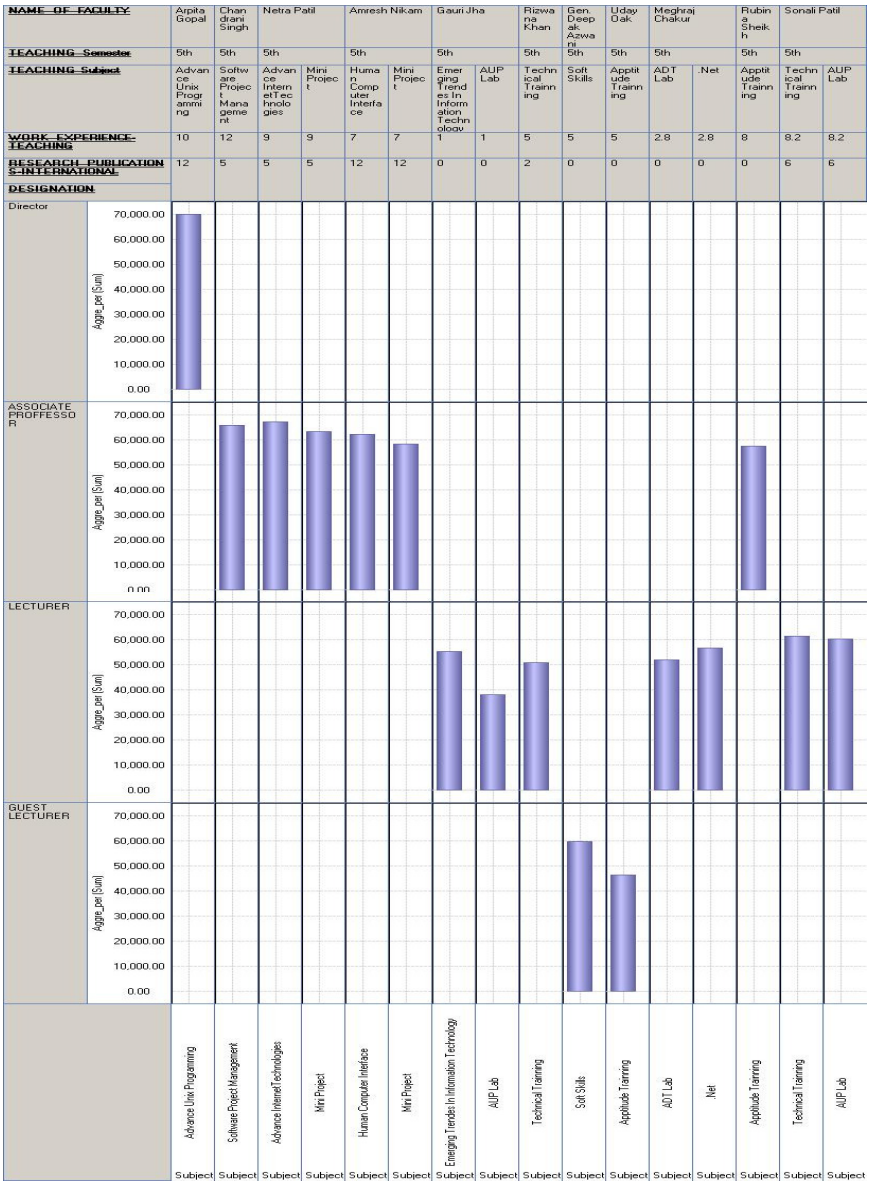


TABLE 5: Performance trend- management faculty

In the above section the overall performance of the management faculties based on student feedback have been shown first using traditional approach and then by using mining methodology to provide a more accurate result. In this section we have used OLAP statistical tool to extract certain trends in faculty performance and also to assess individual faculty performance across several parameters represented in the cube form and extracted it in to the grid and synchronized with the chart. The patterns as identified are as follows:

- The consistency in performance of faculties in the Associate Professor level was found to be more leveled than the faculties at the lecturer level.
- Also the dip in the performance when analyzed across the lecturer level was found to be more than that at the Associate Professor Level.
- The performance of the visiting faculties showed a subsequent drop in spite of them having considerable industry and teaching experience.

5. CONCLUSION AND FUTURE WORK

The future work will contain association rule mining on the student feedback database and dependencies will be analyzed to draw some meaningful conclusions.

6. REFERENCES

1. Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J., 1984. Classification and RegressionTree Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California.

2. A.K. Jain and R. C. Dubes. [1988]. Algorithms for Clustering Data. Prentice Hall.

3. R Agrawal, R Srikant Fast Algorithms for Mining Association rules in Large Databases (1994) by Proceedings of the VLDB.

4. Ganti, V., Gehrke, J. and Ramakrishnan, R. 1999a. CACTUS-Clustering Categorical Data Using Summaries. In Proceedings of the 5th ACM SIGKDD, 73-83, San Diego, CA.

5. GUHA, S., RASTOGI, R., and SHIM, K. 1999. ROCK: A robust clustering algorithm for categorical attributes. In Proceedings of the 15th ICDE, 512-521, Sydney, Australia.

6. Zaki, M.J. Scalable algorithms for association mining Knowledge and Data Engineering, IEEE Transactions on Volume 12, Issue 3, May/Jun 2000 Page(s):372 390 Digital Object Identifier 10.1109/69.846291

7. Chiu, T., Fang, D., Chen, J., and Wang, Y. 2001. A Robust and scalable clustering algorithm for mixed type attributes in large database environments. In Proceedings of the 7th ACM SIGKDD, 263-268, San Francisco, CA.

8. Luan J. [2002] “Data Mining and Knowledge Management in higher Education” Presentation at AIR Forum, Toronto, Canada.

9. Fathi Elloumi, Ph.D., David Annand. [2002] Integrating Faculty Research Performance Evaluation and the Balanced Scorecard in AU Strategic Planning: A Collaborative Model.

10. Raoul A. Arreola, Michael Theall, and Lawrence M. Aleamoni [2003] Beyond Scholarship: Recognizing the Multiple Roles of the Professoriate." Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
11. M.R.K. Krishna Rao. [2004] Faculty and Student Motivation: KFUPM Faculty Perspectives

12. Karin Sixl-Daniell, Amy Wong, and Jeremy B. Williams.[2004] The virtual university and the quality assurance process: Recruiting and retaining the right faculty. Proceedings of the 21st ASCILITE Conference.

13. Emmanuel N. Ogor.[2007] Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques. Electronics, Robotics and Automotive Mechanics Conference, 2007. CERMA 2007 Volume, Issue, 25-28 Sept. 2007 Page(s): 354 – 359 Digital Object Identifier 10.1109/CERMA.2007.4367712.

14. Amy Wong and Jason Fitzsimmons [2008] Student Evaluation of Faculty: An Analysis of Survey Results. U21GlobalWorking Paper Series, No. 003/2008.

15. Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás.[2008], Data Mining Algorithms to Classify Students. The 1st International Conference on Educational Data Mining Montréal, Québec, Canada, June 20-21, 2008 Proceedings.

16. Chandrani Singh ,Dr. Arpita Gopal Performance Analysis of Faculty Using Data Mining Techniques,IJFCSA-2010,1st edition

Subjective Probabilistic Knowledge Grading and Comprehension

A.Suresh Babu asureshjntu@gmail.com
Asst. Professor
Department of Computer
Science Engineering JNTUA,
Anantapur

Dr P.Premchand p.premchand@uceou.edu
Professor, Department of Computer
Science Engineering Osmania
University Hyderabad

Dr A.Govardhan govardhan_cse@yahoo.co.in
Professor, Department of Computer
Science Engineering
JNTUHCEJ, Jagityala

Abstract

Probabilistic Comprehension and Modeling is one of the newest areas in information extraction, text linguistics. Though much of the research vested in linguistics and information extraction is probabilistic, the importance is disappeared in 80's. This is just because of the input language is noisy, ambiguous and segmented. Probability theory is certainly normative for solving the problems related to uncertainty. Perhaps human language processing is simply non-optimal, non-rational process. Subjective Probabilistic approach fixes this problem, through scenario, evidence and hypothesis.

Keywords: Probability & Statistics, Probabilistic Comprehension, Subjective Grading, Subjective Probability.

1. INTRODUCTION

It is not possible to stop grading which is an objective standardized measure for valuating the texts. Subjective and objective measures declare the values in the texts. Subjective grading is percussive than objective grading, as the previous researches prove to be mechanistic towards grading the knowledge when objective measure are used.

One of the central problems in the field of knowledge extraction or discovery is the availability and the development methods to determine good measures of interestingness of discovered patterns. Such measures of interestingness are divided into objective measures - those that depend only on the structure of a pattern and the underlying data used in the discovery process, and the subjective measures - those that also depend on the class of users who examine the pattern. Objectiveness describes absolute grading of knowledge, whereas subjectiveness describes relative grading. Probabilistic belief is the most important concern that paves out the clarity in selection of measures.

1.1 Important Definitions

Objective – is a statement that is completely unbiased. It is not touched by the speaker's previous experiences or tastes. It is verifiable by looking up facts or performing mathematical calculations.

Subjective – is a statement that has been colored by the character of the speaker or writer. It often has a basis in reality, but reflects the perspective through with the speaker views reality. It cannot be verified using concrete facts and figures.

1.2 When to Be Objective and Subjective

Objective – it is important to be objective when you are making any kind of a rational decision. It might involve purchasing something or deciding which job offer to take. You should also be objective when you are reading, especially news sources. Being objective when you are meeting and having discussions with new people helps you to keep your concentration focused on your goal, rather than on any emotions your meeting might trigger. In essence, accomplishing the goal without any distractions is objective oriented.

Subjective – can be used when nothing tangible is at stake. When you are watching a movie or reading a book for pleasure, being subjective and getting caught up in the world of the characters makes your experience more enjoyable. If you are discussing any type of art, you have to keep in mind that everyone's opinions on a particular piece are subjective. In essence, accomplishing the goal with the complete knowledge and the supplements are not distractions rather support enriching the knowledge to fulfill the goal.

Natural language degree expressions denote degrees or relations between degrees and norms of expectation that lie on a scale [3]. What has not been clarified yet is what kind of relations is expressed by degree expressions. Degree expressions are not only a means of denoting graded properties, but they also allow for the adaptation to varying precision requirements as well as for very efficient communication by referring to entities that are only available implicitly or derivable from the context of the phrase.

1.3 Basis

The Dempster-Shafer theory, also known as the theory of belief functions, is a generalization of the Bayesian theory of subjective probability. Whereas the Bayesian theory requires probabilities for each question of interest, belief functions allow us to base degrees of belief for one question on probabilities for a related question. These degrees of belief may or may not have the mathematical properties of probabilities; how much they differ from probabilities will depend on how closely the two questions are related [4].

The Dempster-Shafer theory is based on two ideas: the idea of obtaining degrees of belief for one question from subjective probabilities for a related question, and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence.

Implementing the Dempster-Shafer theory in a specific problem generally involves solving two related problems. First, we must sort the uncertainties in the problem into a priori independent items of evidence. Second, we must carry out Dempster's rule computationally. These two problems and their solutions are closely related. Sorting the uncertainties into independent items leads to a structure involving items of evidence that bear on different but related questions, and this structure can be used to make computations feasible.

2. RELATED WORK

2.1 Dempster-Shafer theory

The method of reasoning with uncertain information known as Dempster-Shafer theory arose from the reinterpretation and development of work of Arthur Dempster and by Glenn Shafer in his book *a mathematical theory of evidence*, and further publications. Dempster-Shafer theory is a belief system that deals with the evidence available for a hypothesis on uncertainty. The uncertainty is represented as $\boldsymbol{Bel}(\mathbf{U})$, the belief which is an evidence for hypothesis, $Plaus(\mathbf{U})$, the plausibility which is an evidence that does not contradict the hypothesis.

Suppose, for example, that Betty and Sally testify independently that they heard a burglar enter my house. They might both have mistaken the noise of a dog for that of a burglar, and because of this common uncertainty, it is not possible to combine degrees of belief based on their evidence directly by the Dempster's rule. But if to consider explicitly the possibility of a dog's presence, then three independent items of evidence can be identified: evidence for or against the presence of a dog, evidence for Betty's reliability and the evidence for Sally's reliability. These items of evidence can be combined by Dempster's rule and the computations are facilitated by the structure that relates the different questions to arise.

Belief and Plausibility can be viewed as providing a lower and upper bounds respectively on the likelihood of $\mathbf{U}$. Over all the possible states and worlds the Dempster and Shafer belief and plausibility functions are defined.

$$Bel(x) : 2^W \rightarrow [0..1]$$
$$Plaus(x) : 2^W \rightarrow [0..1]$$

Where $W = \{w_1, w_2, ..., w_n\}$

Since belief and plausibility encode evidence, they cannot be defined solely on individual states

$$\sum_i Bel(w_i) \leq 1$$
$$\sum_i Plaus(w_i) \geq 1$$

2.2 Grading Knowledge

Information Extraction is akin to "Knowledge Extraction" in text maps natural language onto a formal representation of the facts contained in the texts. Common text knowledge extraction methods show a severe lack of methods for understanding natural language "degree expressions", which describe gradable properties like price and quality, respectively [3]. However, without an adequate understanding of such degree expressions it is often impossible to grasp the central meaning of a text. Degree expressions describe gradable attributes of objects, events or other ontological entities. Complex Lexical semantics is to be instrumented when describing what remains constant when a word is put into different contexts. Fuzzy Logic is even implemented for defining various grades and scales for measuring quantitative text. The special calculi for the text are available to derive valid conclusions from a representation that sticks near to the surface of the utterance and can also handle non-gradable text. Ontological research degree expressions are denoted as their own entities, the upper ontology consisting of the categories such as, Physical-Object and Action augmented by the primitive concept Degree. Some researchers believe that ontological entities are too unparsimonious; the lexical approach only seems to be the one that overcomes the problems. Adjectives of the text propose much easier method of grading. The gradability of (most) adjectives is more obvious than the gradability of members of the other word classes. Besides the fine frequency of gradable adjectives another reason to focus on them is that any advancement of more general mechanisms for degree expressions can easily

be contrasted with other authors' research on adjectives, while — but as a fact — degree expressions in general have not been considered in any approach towards deep understanding of language.

A well accepted and approved classification of adjectives is much more important for classifying minor and major groups. Metonymous collection of adjectives plays a fair role in describing the adjective classes. Multiple word senses, even disregarding the metonymous figurative interpretations many relative adjectives still cannot be attributed a single meaning. A common example for this observation is the dichotomy between "physical — mental" and "physical — logical". Depending on the context "physical" is put into, the appropriate conceptual scale must be chosen. Word sense disambiguation, hypallage, deserves far more attention to be developed technologically.

3. PROPOSED WORK

3.1 Probabilistic Grading

Degree expressions, gradability on comparisons, sub-categorizations, ontological significances, metonymy, word senses and sense disambiguation are some of the concrete and qualified methods for grading text. These methods are employed with priori information about the characteristics of the text corpus in the experiments related to determining the knowledge. The preponderance on the methods of grading has evidence in assigning or grading text or knowledge to some degree by means of comparisons and with probabilistic distinctions is the quintessential area of the research that has been working around by IE personnel waiting with eagerness. Current models for this problem have been studied mostly from a linguistic perspective and less so from one of real-world text understanding. The semantic significance is bought into the picture of research only to a small extent that promises the development of rich cohesive frameworks as a future work. A large number of implausible readings that deal with realistic outputs on the literature generate a negative impact on the natural language analysts. Syntax oriented approaches methodically fail to account for interpretations that depend entirely on semantic or conceptual criteria.

In the present work, probabilistic grading employs statistical and probabilistic methods for grading knowledge. There are several works that have been implemented for measuring correlation to express the relationship between two or more variables. Canonical correlation is an additional procedure for assessing the relationship between variables. As there are computational issues in canonical correlation which are viewed as limitations for determining the degree by comparison for grading the text, the subjective probability is applied for the successive discovery of grading the text and knowledge from the large corpus. The method proposed is feasible to implement in the theories of "discourse and corpus" that endorses the quality results in grading.

3.2 Dempster-Shafer Algorithm

Dempster-Shafer evidence theory and subjective probability provides a useful computational scheme for integrating uncertainty information from multiple sources [4]. The algorithm initiates with; the frame of discernment of the problem domain that is, to determine the grade of text. Assume U be the set of mutually and exhaustive hypotheses i.e., all possibilities of expressing the grade of text. The degree of evidence is used to determine the degree of the text. In general, the evidence is calculated for the individual of the hypotheses, but in the context the evidence determines the grade probabilistically for the valued text. That the more of its value the text is graded as good; evidence holds the fact of good grade for the text. The evidence is the evidence is observed using a function m(x) that provides the following Basic Probability Assignments on U.

Given a certain piece of evidence about the value of the text, the belief that one is willing to commit for the specified value for the text exactly to A is represented by m(A), this holds for any A□U. The subset A of frame U is called the focus element of m, if m(A)>0. That is; to determine grade of a particular text is within the scope of a set of specified values.

$$\begin{cases} m(\phi)=0 \\ \sum_{A\subset U} m(A)=1 \end{cases}$$

The DS theory starts by the assumption of Universe of Discourse θ, also called a Frame of Discernment. This is a set of mutually exclusive alternatives. Considering the current case study of determining the grade of a text, θ would be the set consisting of all possible grades.

Elements of 2^{θ}, i.e., subsets are the class of general propositions in the domain. For example, the proposition "The grade is highly starred" corresponds to the set of the elements of θ which are high graded stars.

A function $m:2^{\theta}\rightarrow[0,1]$ is called a basic probability assignment if it satisfies $m(\phi)=0$ and

$$\sum_{A\subseteq\theta} m(A) = 1$$

The quantity m(A) is defined as A's basic probability number. It represents the strength of some evidence; our exact belief in the proposition represented by A.

A function $m:2^{\theta}\rightarrow[0,1]$ is called a belief function if it satisfies $Bel(\phi)=0$, $Bel(\phi)=1$, and for any collection $A_1,\ldots A_n$ of subsets of θ.

$$Bel(A_1\cup,\ldots\cup A_n) \geq \sum_{\substack{1\subseteq\{1\ldots n\} \\ I\neq\phi}} (-1)^{|I|+1} Bel\left(\bigcap_{i\in 1} A_i\right)$$

A belief function assigns to each subset of θ a measure of our total belief in the proposition represented by the subset.

There corresponds to each belief function one and only one basic probability assignment. Conversely, there corresponds to each basic probability assignment one and only one belief function. They are related by the following two formulae:

$$Bel(A) = \sum_{B\subseteq A} m(B) \quad \text{, for all A } \subseteq\theta$$

$$\text{m(A)} = \sum_{B\subseteq A} (-1)^{|A-B|} Bel(B)$$

Thus a belief function and a basic probability assignment convey exactly the same information. Corresponding to each belief function are three other commonly used quantities that convey the same information:

A function $Q:2^{\theta}\rightarrow[0,1]$ is called a commonality function if there is a basic probability assignment, m, such that

$$Q(A) = \sum_{A\subseteq B} m(B) \quad \text{for all A}\subseteq\theta$$

The doubt function is given by

$$Dou(A) = Bel (\sim A)$$

And the upper probability function is given by

$P^*(A) = 1 - Dou(A)$
This expresses how much we should belief in A if all currently unknown facts were to support A.

This the true belief in A will be somewhere in the interval $\lfloor Bel(A), P^*(A) \rfloor$

3.3 Probabilistic Comprehensive Grading
It ought to be certainly counted that probabilistic approaches focus a paradox in process modeling and data comprehension. And it is simultaneously one of the oldest and one of the newest research areas in linguistics, where already much of the research that was done is of statistical and probabilistic in nature [1][2]. Actually, the probabilistic comprehension and modeling is drawn from early Bayesian precursors.

Probability theory is certainly the best normative model for solving problems of decision making under uncertainty [1]. But perhaps it is a good normative model, but a bad descriptive one. Despite the fact that probability theory was originally invented as a cognitive model of human reasoning under uncertainty, perhaps people do not use probabilistic reasoning in cognitive tasks like language production and comprehension.

Probabilistic modeling provides evidence for knowledge comprehension and as well as for knowledge grading. Since, they are not definitely descriptive, a fuzzy set of graded values are attributed to the knowledge for subject grading.

Since one of the oldest and most robust effects of the linguistic texts is the word frequency effect, word frequency plays an important role in auditory, comprehensive, productive and visual modalities [2]. Since, variations of word frequency induce noise, grading of knowledge which includes word frequency effect requires probabilistic comprehension to determine the suitable importance of the knowledge. Grammatical subjects make more impact in frequency identification. In an instance of research experiments, “When subjects made recognition errors, they responded with words that were higher in frequency than the words that were presented”. In an another instance of research experiment “gating paradigm”, in which subjects hear iteratively more and more of the waveform of a spoken word, to show that high-frequency words were recognized earlier (i.e. given less of the speech waveform) than low frequency words [2].

Using the Dempster-Shafer algorithm, it is perfectly possible to perform an experiment for subjective grading of knowledge. The different components that are involved in the experimentation are scenario, evidence, hypothesis and their relationships. Various probabilistic factors are assigned to the attributes that belong to the knowledge. In different hypothesis various grades are attributed to the knowledge, in a selected scenario and an instance of hypothesis the knowledge is graded with select set of qualities. While working with Dempster-Shafer Engine, it is evidently found that a graphical assessment of grading knowledge is possible. In a corpus of linguistic text where a select set of nuggets can be applied with scenario, evidence and hypotheses may be performed to determine the suitable grade.

4. CONCLUSIONS

In this paper the work is aimed at comparison of the probabilistic approaches. The probability theory works on crispy results, where grading requires a scalable demarcation to qualify the knowledge units. The scalable demaracative units are described as evidences and various hypotheses are prepared. The Dempster-Shafer algorithm using the Dempster-Shafer Engine (DSE) has become more practicable to implement the concept of evaluating the grade of the knowledge units using subjective probability.

5. REFERENCES

[1] Rieko Matsuda,Yuzuru Hayashi,Chikako Yomota,Yoko Tagashira,Mari Katsumine and Kazuo Iwaki, "Statistical and probabilistic approaches to confidence intervals of linear calibration in liquid chromatography", The Analyst, www.rsc.org/analyst, (C) 2001.

[2] Dan Jurafsky, "Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production", appeared in 'Probabilistic Linguistics', edited by Rens Bod, Jennifer Hay, and Stefanie Jannedy, MIT Press, (C) 2002.

[3] Steffen Staab, "Grading Knowledge: Extracting Degree Information from Texts", ISBN 3-540-66934-5 (C) Springer-Verlag Berlin Heidelberg 1999.

[4] Nic Wilson, "Algorithms for Dempster-Shafer Theory", School of Computing and Mathematical Sciences, Oxford Brookes University.

An Intelligent Analysis of Crime Data for Law Enforcement Using Data Mining

Malathi. A malathi.arunachalam@yahoo.com
Research Scholar
Bharathiar University
Coimbatore, 641 046, India.

Lt. Dr. S. Santhosh Baboo santhos2001@sify.com
Reader, Department of Computer Science
D. G. Vaishnav College
Chennaiy, 600 106, India

Abstract

The concern about national security has increased significantly since the 26/11 attacks at Mumbai. However, information and technology overload hinders the effective analysis of criminal and terrorist activities. Data mining applied in the context of law enforcement and intelligence analysis holds the promise of alleviating such problem. In this paper we use a clustering/classify based model to anticipate crime trends. The data mining techniques are used to analyze the city crime data from Tamil Nadu Police Department. The results of this data mining could potentially be used to lessen and even prevent crime for the forth coming years.

Keywords: Crime, Data Mining, Classification, Cluster and Crime Analysis

1. INTRODUCTION

The concern about national security has increased significantly since the terrorist attacks on November 26, 2008 at Mumbai. Any intelligent system [15][16][17] as crime analysis tool for police, it is required to understand Indian Police structure, responsibilities of the police, key changes and challenges the police is forcing [14].

Intelligence agencies such as the CBI and NCRB (National Crime Record Bureau) are actively collecting and analyzing information to investigate terrorists' activities [12]. Local law enforcement agencies like SCRB(State Crime Record Bureau) and DCRB(District Crime Record Bureau)/CCRB (City Crime Record Bureau) have also become more alert to criminal activities in their own jurisdictions. One challenge to law enforcement and intelligence agencies is the difficulty of analyzing large volumes of data involved in criminal and terrorist activities. Data mining holds the promise of making it easy, convenient, and practical to explore very large databases for organizations and users. Different kinds of crime patterns are clustered together instead of using geographical clustering. Based on these crime patterns, a classifier model is applied to predict the crime trend. However, much literature on crime trends focuses only on violence [13] In this paper, we review data mining techniques applied in the context of law enforcement and intelligence analysis.

2. AN OVERVIEW OF DATA MINING

In this paper we review the Crime Data Mining in two directions

1. Crime Types and security concerns
2. Crime Data Mining Approaches and techniques

2.1 CRIME TYPES AND SECURITY CONCERNS

Crime is defined as "an act or the commission of an act that is forbidden, or the omission of a duty that is commanded by a public law and that makes the offender liable to punishment by that law" (Webster Dictionary). An act of crime encompasses a wide range of activities, ranging from simple violation of civic duties (e.g., illegal parking) to internationally organized crimes (e.g., the 9/11 attacks). The following are the different types of crimes

- Property crime
- Violent Crime
- Crime against Women and Child
- Traffic Violations
- Cyber Crime and
- Others

2.2 CRIME DATA MINING APPROACHES AND TECHNIQUES

Data mining is defined as the discovery of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data [1]. Data mining in the framework of crime and intelligence analysis for national security is still a young field.

The following describes our applications of different techniques in crime data mining. *Preprocessing* has been used to keep the data set ready for the process. *Entity extraction* has been used to automatically identify person, address, vehicle, and personal properties from police narrative reports [2]. *Clustering* techniques has been used to cluster the city crime data mining depends on the crimes. *Classification* has been used to detect criminal data from the city crime data base. *Social network analysis* has been used to analyze criminals' roles and associations among entities in a criminal network [9].

3. DATA MINING TASKS

3.1 PREPROCESSING

The data set was made available by the department of Police. The range of years available and utilized was between 2000 and 2009.

Data Attributes

The following yearly attributes were presented and used in the data set for the city crime statistics [7][8]

- Property - # of property crimes (sum of next 6 attributes)
  - Murder for gain
  - Dacoity
  - Prep.&Assembly For Dacoity
  - Robbery
  - Burglary
  - Theft
- Violent - # of violent crimes (sum of next 5 attributes)
  - Murder
  - Attempt to commit murder
  - C.H.Not Amounting to murder
  - Hurt/Grievous Hurt
  - Riots
- Crime against Women and Child - # of women & child crimes (sum of next 5 attributes)
  - Rape

  - Dowry Death
  - Molestation
  - Sexual Harassment
  - Cruelty by Husband and her relatives.
- Others – (sum of next 6 attributes)
  - Kidnapping & Abduction of others
  - Criminal Breach of Trust
  - Arson
  - Cheating
  - Counterfeiting
  - Others IPC crimes

3.2 PREDICTION OF MISSING VALUES

The first task is the prediction of the size of the population of a city [6]. The calculation of per capita crime statistics helps to put crime statistics into proportion. However, some of the records were missing one or more values. Worse yet, half the time, the missing value was the "city population size", which means there was no per capita statistics for the entire record. Over some of the cities did not report any population data for any of their records. To improve the calculation of "yearly average per capita crime rates", and to ensure the detection of all "per capita outliers", it was necessary to fill in the missing values. The basic approach to do this was to cluster population sizes, create classes from the clusters, and then classify records with unknown population sizes [3]. Why use clustering to create classes? Classes from clusters are more likely to represent the actual population size of the cities. The only value needed to cluster population sizes was the population size of each record. These values were clustered using

"weka.clusterers.EM -I 100 -N 10 -M 1.0E-6 -S 100"

in Weka Table 1 shows the results. Ten clusters were chosen because it produced clusters with mean values that would produce per capita calculations close to the actual values [4].

Cluster	Percent	Mean	StdDev
0	1	46.012	3.938
1	13	3.488	863
2	9	7.830	894
3	8	10.233	1.185
4	7	11.444	0.0012
5	20	13.885	3.570
6	13	24.197	3.648
7	4	35.477	3.521
8	14	5.732	810
9	12	1,975	816

TABLE 1: Weka.clusters.EM

4. PREDICTION OF CRIME TRENDS

The next task is the prediction of future crime trends. This meant we tracked crime rate changes from one year to the next and used data mining to project those changes into the future. The basic method here is to cluster the cities having the same crime trend, and then using ”next year” cluster information to classify records [11]. This is combined with the state poverty data to create a classifier that will predict future crime trends. Eight "delta" attributes were applied to city crime clustering: Murder for gain, Dacoity, Prep.&Assembly For Dacoity, Robbery, Burglary, Theft, Murder, Attempt to commit murder, C.H.Not Amounting to murder, Hurt/Grievous Hurt, Riots, Rape, Dowry Death, Molestation, Sexual Harassment, Kidnapping & Abduction of others, Criminal Breach of Trust, Arson, Cheating, Counterfeiting, and Others IPC crimes. These attributes were clustered using

'Weka 3.5.8's, Simple EM (expectation maximization)'

with parameters of "EM -I 100 -N 4 -M 1.oE-6 -S 100" [4]. EM is a deviation of K-Means clustering. Four clusters were chosen because it produced a good distribution with a relatively easy to interpret set of clusters [5]. Usually, the high level interpretation of clusters from an unsupervised algorithm is not easily defined. However, in this case, the four clusters produced had the following attributes: Note: The clusters are ordered from best to worst.

1) C0: Crime is steady or dropping. The Sexual Harassment rate is the primary crime in flux. There are lower incidences of: Murder for gain, Dacoity, Preparation for Dacoity, rape, Dowry Death and Culpable Homicide.
2) C1: Crime is rising or in flux. Riots, cheating, Counterfeit, and Cruelty by husband and relatives are the primary crime rates changing. There are lower incidences of: murder and kidnapping and abduction of others.
3) C2: Crime is generally increasing. Thefts are the primary crime on the rise with some increase in arson. There are lower incidences of the property crimes: burglary and theft.
4) C3: Few crimes are in flux. Murder, rape, and arson are in flux. There is less change in the property crimes: burglary, and theft. To demonstrate at least some characteristics of the clusters,

5. CITY CRIME ANALYSIS

Looking at the number of property crimes, it looks like crime at the city has been going down since 2004 except in 2009. But the number of violent crimes has been in flux.

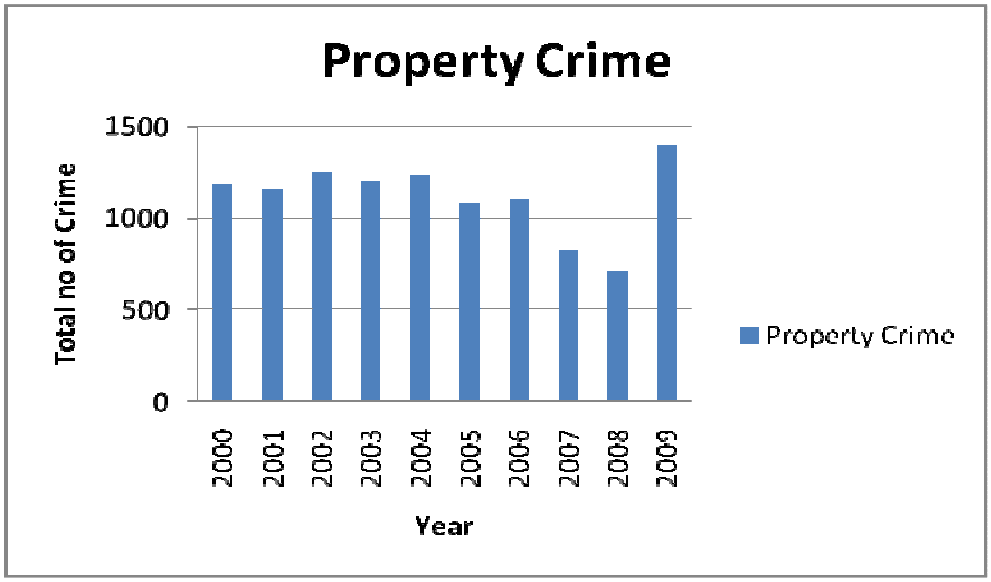


FIGURE 1: Property Crime Analysis.

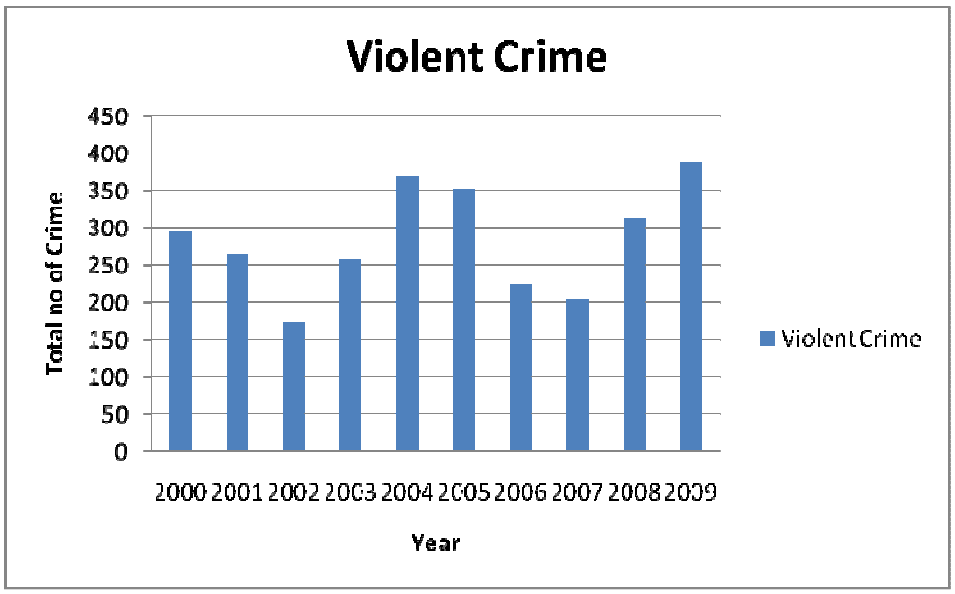


FIGURE 2: Violent Crime Analysis.

Crime against women includes Rape, Dowry Death, Molestation, Sexual Harassment, Cruelty by husband or relatives, Kidnapping & Abduction of women & Girls. Crimes against the women have been going down since 2000, but once again the same thing has been increased since 2004 till 2008, finally it started coming down.

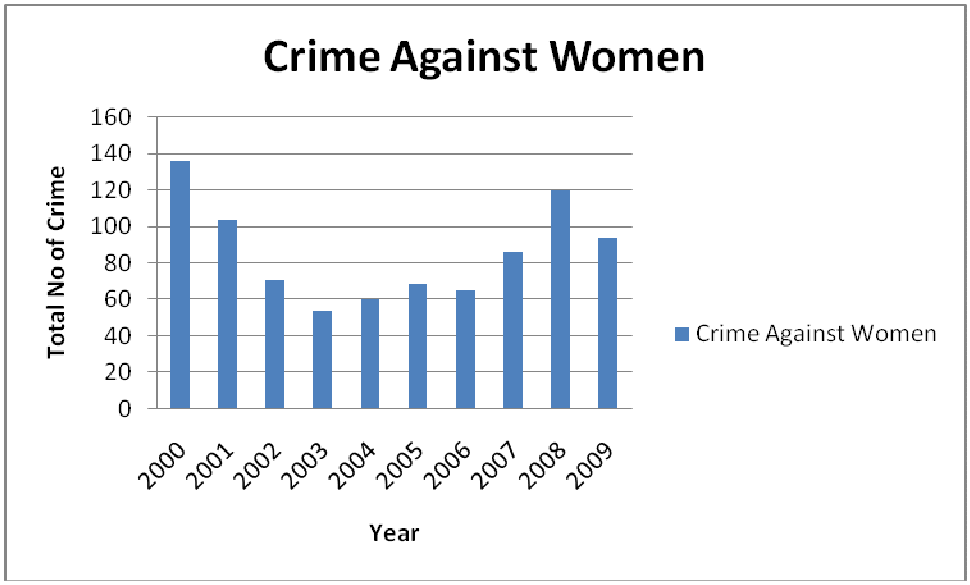


FIGURE 1: Crime Against Women – Analysis.

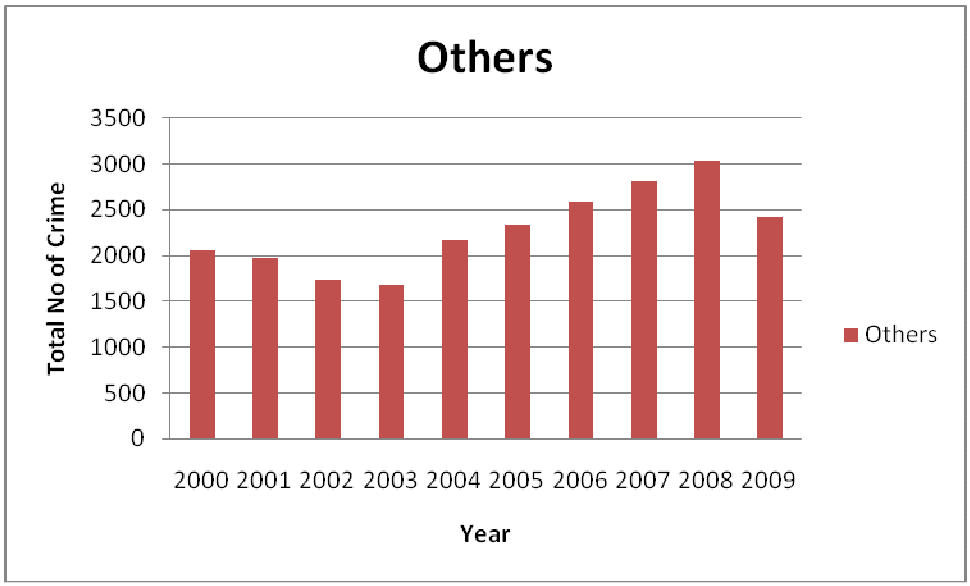


FIGURE 1: Others Crime Analysis

The other crimes includes Kidnapping & Abduction of others, Criminal Breach of Trust, Arson, Cheating and Counterfeiting. These crimes are in flux.

5. CONSLUSION & FUTURE WORK

Data mining applied in the context of law enforcement and intelligence analysis holds the promise of alleviating crime related problem. In this paper we use a clustering/classify based model to anticipate crime trends. The data mining techniques are used to analyze the city crime data from Tamil Nadu Police Department. The results of this data mining could potentially be used to lessen and even prevent crime for the forth coming years. From the encouraging results, we believe that crime data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis. Many future directions can be explored in this still young field. Visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern [10].

Acknowledgements

We are highly indebted to Dr. Sylendra Baboo, Commissioner of Police, Coimbatore. We are also thankful to him and his team members for sharing their valuable knowledge in this field

6. REFERENCES

[1] Fayyad, U.M., & Uthurusamy, R. (2002). Evolving data mining into solutions for insights. Communications of the ACM, 45(8), 28-31.

[2] Chau, M., Xu, J., & Chen, H. (2002). Extracting meaningful entities from police narrative reports. In: Proceedings of the National Conference for Digital Government Research (dg.o 2002), Los Angeles, California, USA.

[3] Classification via Decision Trees in weak, Depaul University, http://maya.cs.depaul.edu/~classes/ect584/WEKA/classify.html

[4] Frank Dellaert, The expectation Maximization Algorithm, Technical Report, Georgia Institute of Technology, 2002

[5] Weka 3 - Data Mining with Open Source Machine
Learning Software in Java. 7 Dec. 2008 http://www.cs.waikato.ac.nz/ml/weka/

[6] Stephen Schneider, Predicting crime: a review of the research,
Department of Justice Canada, 1-2, 2002

[7] Major Crime Trends – Tamil Nadu
http://www.tnpolice.gov.in/crimeprofile.html

[8] Major Crime Trends – Tamil Nadu
http://www.tnpolice.gov.in/CAWChart.html

[9] David S. Coppock, Why Lift? Data Modeling and Mining, DM review online, 2002

[10] Shyam Varan Nath, Crime Pattern Detection Using Data
Mining, Proceedings of the 2006 IEEE/WIC/ACM international
conference on Web Intelligence and Intelligent Agent Technology , 41-44, 2006

[11] University Crime Data Mining
http://www.cse.msu.edu/~kingstua/Team3/

[12] Manish Gupta et al./ Crime Data Mining for Indian Police Information System, Proceeding of the 2008 Computer Society of India.

[13] Zimring, F., & Hawkins, Crime is not the problem, Oxford
University press, 1997

[14] Krishnamorthy, S. (2003). Preparing the Indian Police for 21st Century. *Puliani and Puliani*, Bangalore, India.

[15] Tuchinda, R., Szekely P. & Knoblock, C. A (2007). Building Data Integration Queries by Demonstration. *Proceedings of the 12th international conference on Intelligent user interfaces*

[16] Michelson, M. & Knoblock, C.A. (2006). Phoebus: A System for Extracting and Integrating Data from Unstructured and Ungrammatical Sources. In *Proceedings of AAAI*.

[17] Kumar, M., Gupta, A. & Saha, S. (2006). Approach to Adaptive User Interfaces using Interactive Media Systems. *Proceedings of the 11th international conference on Intelligent user interfaces*.

CALL FOR PAPERS

Journal: International Journal of Data Engineering (IJDE)
Volume: 2 **Issue:** 1
ISSN:2180-1274
URL: http://www.cscjournals.org/csc/description.php?JCode=IJDE

About IJDE

Data Engineering refers to the use of data engineering techniques and methodologies in the design, development and assessment of computer systems for different computing platforms and application environments. With the proliferation of the different forms of data and its rich semantics, the need for sophisticated techniques has resulted an in-depth content processing, engineering analysis, indexing, learning, mining, searching, management, and retrieval of data.

International Journal of Data Engineering (IJDE) is a peer reviewed scientific journal for sharing and exchanging research and results to problems encountered in today's data engineering societies. IJDE especially encourage submissions that make efforts (1) to expose practitioners to the most recent research results, tools, and practices in data engineering topics; (2) to raise awareness in the research community of the data engineering problems that arise in practice; (3) to promote the exchange of data & information engineering technologies and experiences among researchers and practitioners; and (4) to identify new issues and directions for future research and development in the data & information engineering fields. IJDE is a peer review journal that targets researchers and practitioners working on data engineering and data management.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJDE.

IJDE List of Topics

The realm of International Journal of Data Engineering (IJDE) extends, but not limited, to the following:

- Approximation and Uncertainty in Databases and Pro
- Autonomic Databases
- Data Engineering
- Data Engineering Algorithms
- Data Engineering for Ubiquitous Mobile Distributed
- Data Engineering Models
- Data Integration
- Data Mining and Knowledge Discovery

- Data Ontologies
- Data Query Optimization in Databases
- Data Warehousing
- Database User Interfaces and Information Visualiza
- Metadata Management and Semantic Interoperability
- Personalized Databases
- Scientific Biomedical and Other Advanced Database
- Social Information Management
- Data Privacy and Security
- Data Streams and Sensor Networks
- Database Tuning
- Knowledge Technologies
- OLAP and Data Grids
- Query Processing in Databases
- Semantic Web
- Spatial Temporal

Important Dates

Volume: 2
Issue: 1
Paper Submission: January 31, 2011
Author Notification: March 01, 2011
Issue Publication: March/April 2011

CALL FOR EDITORS/REVIEWERS

CSC Journals is in process of appointing Editorial Board Members for ***International Journal of Data Engineering (IJDE)***. CSC Journals would like to invite interested candidates to join **IJDE** network of professionals/researchers for the positions of Editor-in-Chief, Associate Editor-in-Chief, Editorial Board Members and Reviewers.

The invitation encourages interested professionals to contribute into CSC research network by joining as a part of editorial board members and reviewers for scientific peer-reviewed journals. All journals use an online, electronic submission process. The Editor is responsible for the timely and substantive output of the journal, including the solicitation of manuscripts, supervision of the peer review process and the final selection of articles for publication. Responsibilities also include implementing the journal's editorial policies, maintaining high professional standards for published content, ensuring the integrity of the journal, guiding manuscripts through the review process, overseeing revisions, and planning special issues along with the editorial team.

A complete list of journals can be found at http://www.cscjournals.org/csc/byjournal.php. Interested candidates may apply for the following positions through http://www.cscjournals.org/csc/login.php.

Please remember that it is through the effort of volunteers such as yourself that CSC Journals continues to grow and flourish. Your help with reviewing the issues written by prospective authors would be very much appreciated.

Feel free to contact us at coordinator@cscjournals.org if you have any queries.

Contact Information

Computer Science Journals Sdn BhD
M-3-19, Plaza Damas Sri Hartamas
50480, Kuala Lumpur MALAYSIA

Phone: +603 6207 1607
+603 2782 6991
Fax: +603 6207 1697

BRANCH OFFICE 1
Suite 5.04 Level 5, 365 Little Collins Street,
MELBOURNE 3000, Victoria, AUSTRALIA

Fax: +613 8677 1132

BRANCH OFFICE 2
Office no. 8, Saad Arcad, DHA Main Bulevard
Lahore, PAKISTAN

EMAIL SUPPORT
Head CSC Press: coordinator@cscjournals.org
CSC Press: cscpress@cscjournals.org
Info: info@cscjournals.org