ISSN:0975-9646
S.Hari Ganesh et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, 1470-1473

HARIALGM: Knowledge Discovery and Data Mining in Pedagogy with DNA Finger Printing

S. HARI GANESH [1] , DR.C.CHANDRASEKAR[2]

[1] *Computer Application Department, Bharathidasan University, Bishop Heber College, Tiruchirapalli,620017, TamilNadu,India.*

[2] *Computer Science Department, Periyar University Salem, TamilNadu, India*

***Abstract*- Knowledge Discovery and Data Mining (KDD) is a multidisciplinary area focusing upon methodologies for extracting useful knowledge from data and there are several useful KDD tools to extract the knowledge. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have created an immense need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced busines intelligence and web discovery solutions. Data mining involves text mining from files, which is an important concept that is related to knowledge discovery. In this paper it proposes an innovative algorithm named HariAlgm for Knowledge discovery in pedagogy. The voice of teacher is converted into text file. The text file which contains several keywords related to subject which he or she undertaking. Making proper analysis of text spoken by the teacher and the keywords in file, this paper can predict the teaching ability and performance further more evaluation is being made with DNA finger printing, to confirm whether it is possible to make right prediction which is hypothetical.**

Keywords: **Data mining, knowledge discovery, DNA Finger printing**

I. INTRODUCTION

The knowledge discovery and data mining (KDD) [3,8] field draws on findings from statistics, databases, and artificial intelligence to construct tools that let users gain insight from massive data sets[10]. People in business, science, medicine, academia, and government collect such data sets, and several commercial packages now offer general-purpose KDD tools [2, 9] .In today's world data mining is a very important concept that used in Analysis and prediction. The paper makes an evaluation of teacher performance using the concept of data mining. The keywords of the subject are being stored in a file. The file is in the text format. The teacher's voice in the class room is being recorded and it being converted into text file that is being stored in server using the Bluetooth technology and transcription service.

The each word in the file is compared with another the keyword file, which contains the list of key words of the subject handled, count the time that is being spent in each keywords. The Analysis is being made related to Duration of time spent in each key word, until matching with all the words spoken by the teacher and keywords in the file. The paper also tries to the correlate between teacher's performance and DNA Finger printing of the teacher which is hypothetical.

II. METHODOLOGY

The Knowledge discovery and data mining in pedagogy [6], in this paper it tries resolve the problem of assessing the performance of teacher through HariAlgm algorithm model and with the student's feedback, then correlate between teacher's performance and DNA Finger printing of the teacher. Finally predict performance of the teacher both psychologically and biologically.

In this paper it uses the concept of Bluetooth technology used to record the voice in the audio files which then being converted into text file using transcription services. The key word file are created carefully using Experts in that area which has to be incorporated with staff audio file. And analysis has be made how many keywords spoken by the staff for how many minutes that helps in evaluation of the staff with the feedback of the staff that has been already taken.

Finally paper proposes DNA finger printing and to correlate the performance of the individual more accurately using the concept of VNTR's and proteomics which is able extract the protein sequence can verified in the ProteinCenter for individuality characteristics and try to find the hidden information in the DNA finger Printings which is an hypothetical idea.

A. HariAlgm Model briefly describes how the whole process being executed

Explained in Figure 1

B) HariAlgm Model-Algorithm

Step 1: Start of class session.

Step 2: Voice of the teacher being recorded in audio file using blue tooth technology to the nearest server.

Step 3: Audio file is then converted in to text file using voice to Transcription services that converts the voice into text.

Step 4: There is an existing key word file 'K' that is being opened.

Step 5: W_i where i=1,2,3,4…n word is taken that is being compared with keyword file K_i. If it matches count starts for the particular keyword 'i'

Else

Wait for the keyword matching

Step 6: The count stops at next occurrence of the next keyword K_i.

Step 7: The time Duration far each keyword K_i is T_i time which is spent for teaching is being recorded.

Step 8: Analysis made with the Time duration which is T_i which is the total number in minutes which is used for computing performance P_i which is compared with the student's feedback which is already been taken manually.

Step9: Use Bio-metric device (southern Blot) - DNA finger printing set of staff S_i is recorded 'i' is Incremented Perform step 1 to step 8 until all 'i' is being exhausted

Step 10: Try to bring out the correlation between the performances using the teacher's DNA finger printing and the staff Performance.

Step11: Analysis is made to predict the result with DNA similarity which is able to distinguish between the good or poor performance of the teacher. The DNA finger prints matching is only hypothetical that has to be proved using bio-metric devices.

C) Related works

In this paper it proposes HariAlgm model which moves from data sets to meaningful biology. The paper also tries to correlate between teacher's performance and DNA Finger printing of the teacher. Even after preliminary identification and validation. The scientists face the critical, time-consuming task of interpreting their proteomics data - extracting meaningful biological information from multiple, complex data sets. ProteinCenter[11] is a web-based data interpretation tool that enables scientists to compare and interpret data sets in minutes instead of months. Scientists can now:

- Reveal the biological context of data set comparisons - independent of search engine and database
- Reveal the biological context of each data set
- Reveal the biological context in quantitative studies
- Reveal the biological context for a single protein

ProteinCenter enables filtering, clustering and statistical bioinformatics analysis[5] utilizing a regularly updated, consolidated protein sequence database containing >10 million non-redundant proteins. In this paper it provides such a tool for analyzing the teacher's performances

HariAlgm: Model

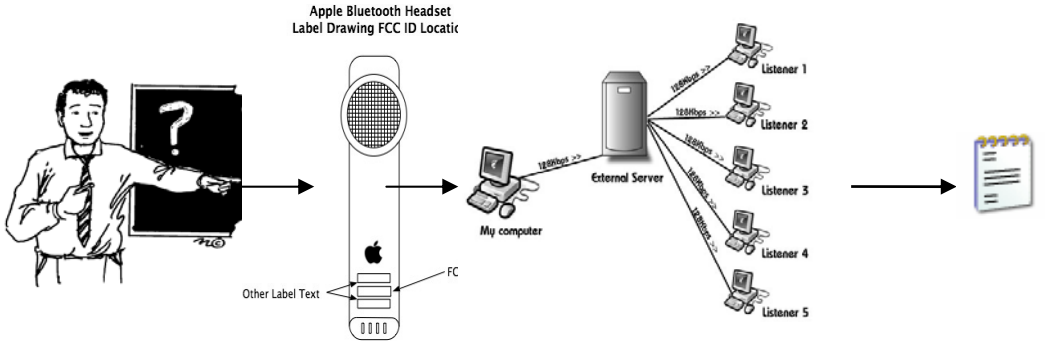


Fig:1 HariAlgm Processing Model

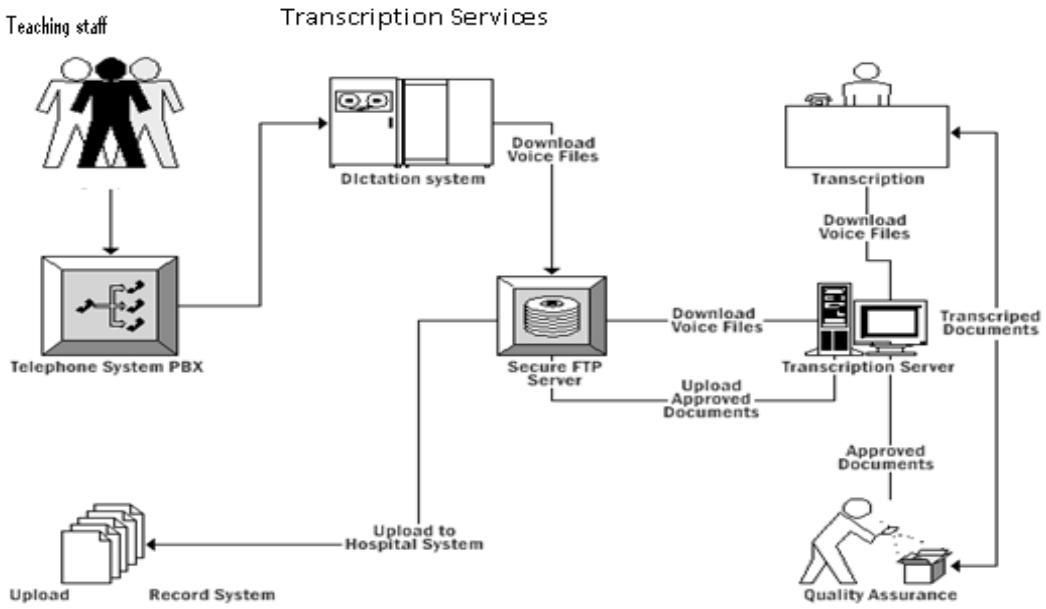


Fig:2 Transcription Service Model

III RESULTS AND DISCUSSIONS

Sample tables (Table1 and Table 2) that depict how to make the analysis. Table 1 consists of staff name and subject handled and table 2 consists of keywords related to respective subject. HariAlgm explains the process involved that being depicted in Table 3.

Table 1 Teacher table (Sample Data)

Name of staff	Subject handled
John	Data mining and data ware housing.
Peter	Compiler design
George	Networks
Britto	Mobile Computing
Stella	Graphics

Table 2 Keyword Table (Sample Data)

Subject	Keywords	Time in minutes
Data mining and data ware housing	Classification	15
	Clustering	15
	Supervised	10
	unsupervised	10
Compiler design	Compiler Definition	10
	Translators	10
	Interpreters	10
	Assemblers	10
	Preprocessors	10
Networks	Networks Definition	10
	LAN	10
	MAN	10
	WAN	10
	Topology definition	10
Mobile Computing	Mobile device definition	10
	Guided media	10
	Unguided media	10
	WI-FI	15
	Emulators	5
Graphics	Graphics definition	10
	Input devices	10
	Output devices	10
	Frame buffer	5
	CRT	15

Table 3 Staff Performance table (Sample Data)

Staff Name	Subject handled	keywords	Staff Time management
John	Data mining and data ware housing.	Classification	15-10
		Clustering	15-15
		Supervised	10-5
		unsupervised	10-9
Peter	Compiler design	Compiler Definition	10-15
		Translators	10-5
		Interpreters	10-10
		Assemblers	10-8
		Pre-processors	10-12
George	Networks	Networks Definition	10-7
		LAN	10-0
		MAN	10-5
		WAN	10-0
		Topology definition	10-20
Britto	Mobile Computing	Mobile device definition	10-10
		Guided media	10-9
		Unguided media	10-12
		WI-FI	15-13
		Emulators	5-5
Stella	Graphics	Graphics definition	10-8
		Input devices	10-7
		Output devices	10-10
		Frame buffer	5-4
		CRT	15-15

Evaluation=ABS (Time management)
Performance=evaluation/10

In fig.3. Depicts staff performance related hariAlgm algorithm. Fig 4: represents the student's feedback which being conducted manually.

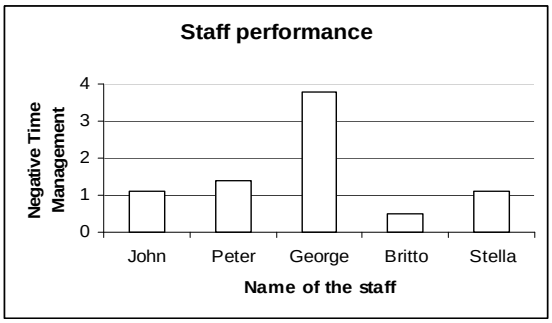


Fig. 1. Shows the staff performance

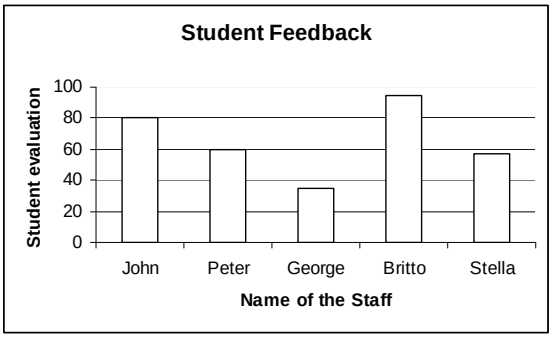


Fig.2. Shows the student feed back

According to the analysis of HariAlgm Performance is directly proportional to student feedback otherwise there is a problem in the pedagogy. This paper tries to analyse the biologically Terminology, Is there any correlation between the performance and DNA Finger printing [7] which is hypothetically placed. Genetic Fingerprinting (also called DNA testing, DNA typing, or DNA profiling) is a technique used to distinguish between individuals of the same species using only samples of their DNA. Although two individuals will have the vast majority of their DNA sequence in common, DNA profiling exploits highly variable repeat sequences called VNTRs.[1,4] These loci are variable enough that two unrelated humans are unlikely to have the same alleles. The technique was first reported in 1984 by Dr. Alec Jeffreys at the University of Leicester, and is now the basis of several national DNA identification databases. A Variable Number Tandem Repeat (or VNTR) is a location in a genome where a short nucleotide sequence is organized as a tandem repeat. These can be found on many chromosomes, and often show variations in length between individuals. Each variant acts as an inherited allele, allowing them to be used for personal or parental identification. Their analysis is useful in genetics and biology research, forensics, and DNA research.

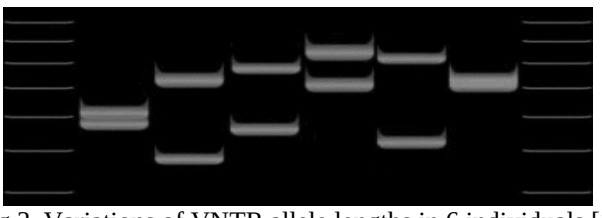
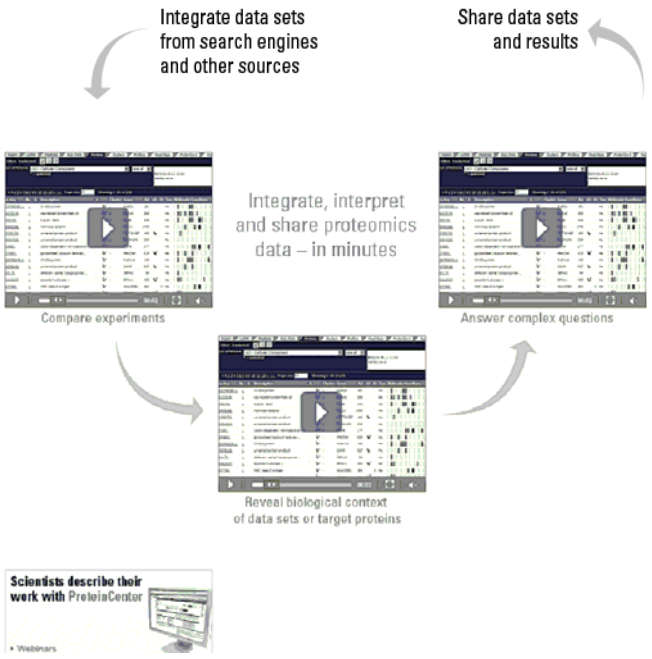


Fig.3. Variations of VNTR allele lengths in 6 individuals.[1]

Using this technique the staff's DNA finger printing can be analyzed and with their similarities the existing results can be compared and the staff DNA finger printing is how far it's able to predict the good or poor performance of the staff. Even after preliminary identification and validation, scientists face the critical, time-consuming task of interpreting their proteomics data - extracting meaningful biological information from multiple, complex data sets. ProteinCenter is a web-based data interpretation tool that enables scientists to compare and interpret data sets in minutes instead of months. Scientists can now:

- Reveal the biological context of data set comparisons - independent of search engine and database
- Reveal the biological context of each data set
- Reveal the biological context in quantitative studies
- Reveal the biological context for a single protein

ProteinCenter enables filtering, clustering and statistical bioinformatics analysis utilizing a regularly updated, consolidated protein sequence database containing >10 million non-redundant proteins. Fig: data analyzed in ProteinCenter. Understanding the relationship between protein sequence, structure and function is fundamental to life science, as well as the healthcare and drug discovery industry. The main aim is to address important questions and problems in proteomics, structural biology, and bimolecular dynamics, using high performance computing (HPC), such as IBM Blue Gene supercomputers. It pursues a broad but well grounded approach to leverage and expand upon our existing techniques for scientific and technological advancement. This paper hypothesis of evaluating the teacher performance with the VNTR has to be checked using bio-metric devices for more specific results.



IV) CONCLUSIONS

The Methodology adopted in this paper has provided opportunities to evaluate the performance of a teacher using the concept of KDD in pedagogy. In pedagogy the voice of the teacher is being recorded and converted into text files and stored in a nearest server using the concept of Bluetooth technology and transcription services. The finger printings of the staffs are being evaluated using protonomics and predictions are then made about the performance of the staff. This will help the teaching fraternity to a greater extent in identifying the performance of a teacher. There are several KDD tools that have been developed to solve several problems. In this paper it develops a model called HariAlgm Model which helps in evaluating a teacher's performance psychologically and biologically. Biologically to prove this concept, this is a hypothesis that has to be tested with bio-metric device to support the prediction made.

REFERENCES

1. B Devlin and N Risch, "*Ethnic differentiation at VNTR loci, with special reference to forensic applications*". American journal of human genetics, 1992 September; 51(3): 534–548.
2. C. Diamantini, D. Potena, and J. Cellini. "*UDDI registry for Knowledge Discovery in Databases services"*. In Proc. Of the International Symposium on Collaborative Technologies and Systems, Orlando, FL, USA,May 21-25 2007. IEEE, PP 321–328.
3. Gregory Piatetsky-Shapiro "*Data mining and knowledge discovery 1996 to 2005 : overcoming the hype and moving from "university" to "business" and analytics"* Data Mining and Knowledge Discovery archive Volume 15 Issue 1, August 2007 Kluwer Academic Publishers Hingham, MA, USA table of contents doi>10.1007/s10618-006-0058-2 pp 605-608
4. H. Halleland, A.J. Lundervold, et ul "*Association between Catechol O-methyltransferase (COMT) haplotypes and severity of hyperactivity symptoms in Adults*" American Journal of Medical Genetics Part B: Neuropsychiatric Genetics Volume 150B, Issue 3, 5 April 2009, PP 403–410.
5. Harlow, H. F. (1983). "*Fundamentals for preparing psychology journal articles"*. Journal of Comparative and Physiological Psychology, 55, PP 893-896.
6. Hasseler, Terri A., "*Out of the Reference Section and onto the Student's Desk: Classroom Uses forder's Companions*" 1964-Pedagogy, Volume 6, Issue 3, Fall 2006, pp. 539-543 (Review)
7. Jörg T. Epplen and Thomas Lubjuhn (eds). Birkhäuser Verlag, Basel. 1999. "*DNA Profiling and DNA Fingerprinting"*.. ISBN 3 7643 6018 6. PP. 252.
8. 8. Li, Xue; Zaiane, Osmar R.; Li, Zhanhuai (Eds.) "*Advanced Data Mining and Applications"* Second International Conference, ADMA 2006, Xi'an, China, August 14-16, 2006, Proceedings Series: Lecture Notes in Computer Science, Vol. 4093, PP XXI, 1110 p.
9. Shu-hsien Liao. "*Knowledge Management Technologies and Application Literature review from 1995-2002*". Expert Systems with Applications. Vol.25, Issue 1. August 2003, pp 155-164.
10. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From "*Data Mining to Knowledge Discovery in Databases*", AI Magazine 17(3): Fall 1996, pp 37-54
11. Wilkins, Marc (Dec. 2009). "*Proteomics data mining*". Expert review of proteomics: Doi:10.1586/epr.09.81.