**NIRAV BHUT**

**Combining and Comparing Multiple Algorithms for Better Learning and Classification: A Case Study of MARF**

This case study is intended to show the general pattern recognition pipeline design methodology, interfaces, classes and data structures to test and compare more than one algorithms and their combinations at the pipeline’s stages. The main problem was to develop a tools to test variety of patterns and NLP algorithms and their combinations for any task given. To compare the tool must allow us to add external plug-ins into the framework. The system need to have the supportive data types and structures that allow for the scripting of the recognition tasks for all potential processing.

Researchers wants this tools to be available publically as a open source project such that everyone from the community can contribute for the enhancement of the project.

At the beginning the MARF framework evolved for stand-alone, sequential and limited support for multithreading. Next step was to make it distributed. But there was a problem with DMARF as used to require lot of manual management. So further step was to make it autonomic.

MARF approach is applied to a variety experiments. Some approaches were text independent- speaker identification while other was writer indentification from scanned hand written docs. Researchers have used first and second guess statistics to find the accurate results. The framrwork is good overall but for more number of algorithms it is more difficult to adjust.

**HIREN**

**Case study: #14**

**Study of Best Algorithm Combinations for Speech**

**Processing Tasks in Machine Learning**

**Using Median vs. Mean Clusters in MARF**

**Presented by:**Serguei A. Mokhov

Concordia University

**Summarized by:**

Hirenkumar R. Tarsadiya

Email Id: hrtarsadiya647@gmail.com

Student Id: 6742904

This case study concentrates on combinations of best algorithms to identify the different features of speaker like their gender, their accent etc. It uses concept of mean clusters and median clusters to find out the best possible combinations of algorithms. By using mean approach this study gives many best possible combinations does not give assurance that chosen combination is only the best one. There may be any other best solution left. This study has adapted SpeakerIdentApp of MARF (where MARF is collection of algorithms for audio and natural language text analysis) to get the results of various speech processing tasks. SpeakerIdentApp uses the disjoint set of training and test set to study the robustness of the algorithms.This application gathers many results of comparison between successful and unsuccessful guesses so second best approach came into picture because of possibility of inaccurateresults at first time. To get the statistics of first and second guess, samples were properly trained and were tested on new features and then all these results and implementations are validated and verified to make it more efficient and flawless. In this study all speaker identification features are analyzed using median and mean clusters. This both clusters are used to get the clear picture and selection of median or mean is completely based on quality on gathered data and chosen algorithms.This study is helpful in many applications where human authentication is necessary and can be useful to get the best area of study according to our combination.

**SOHAN ARGULWAR**

**Topic 16**

**Towards Syntax and semantics of Hierarchical Contexts in Multimedia Processing Applications using MARFL**

**Original Paper by**

Serguei Mokhow

Concordia University, Montreal

**Summary by**

SOHAN ARGULWAR

Student ID: 6757871

sohanargulwar@gmail.com

MARF (Modular Audio Recognition Framework) is a collection of voice/text/speech and Natural Language Processing (NLP) algorithms written in Java and arranged in modular and extensible framework which provides addition of new algorithms. This research paper focus on syntax and semantics for MARF Language. Scripting in required application is difficult task by providing context of all parameters. The context expressions provides scripting MARF based applications as context aware. To make the syntax simpler the overloaded context operators to accept various types of arguments and return types @ and # which are taken from Generic Intensional Programming Language (GIPL) helps to achieve this task. MARF plays and important role in the field of Image Processing and Pattern Recognition. It provides APIs in Java with implementation of unsupervised learning. For example, if user misses any dimension to provide in scripting MARFL it will consider the default value. The paper illustrates the practical application of MARFL as SpeakerIdentApp. The definations of @ and # are overridden to introduce the concept of dot operator for object membership.

**References:**

<http://marf.sourceforge.net/>

**KISHAN SHAH**

A MARF APPROACH TO DEFT’2010

*Analysis of MARF-based approach to classification problems of decades*

*& place of origin of various French publications in DEFT 2010 challenge.*

*Kishan Shah (#6819230)*

*Master of Software Engineering (Student)*

*Concordia University*

*Montreal, Canada*

[*shah.kishan0007@yahoo.co.uk*](mailto:shah.kishan0007@yahoo.co.uk)

*SOEN-6611 Software Measurement*

1. BACKGROUND

MARF provides great usefulness to researchers to decide different combinations of algorithm. It also provides facility to choose best suited algorithm combination for each task. The estimate of algorithm combination is based on statistical estimators and NLP parsing and many other modules.Two types of approach were applied for the challenge of two tracks of identification within francophone press. Classical MARF and NLP MARF pipeline approach. Main usefulness of these 2 approaches is to divide the audio in to sub framework using different combinations of algorithm.

Throughout this experiment there are number of permutations of the variable parameters like, -piestel, -journal -text-only, -title text, simple loading and interpretation, preprocessing, extraction, classification, means and median cluster, etc.All algorithm in this study was not debugged together due to slowness. By using each other’s data in another, using francophone websites additional testing was done. To the result of these experiments highest precision and recall results come from title only processing while journal leading casesexperiments give 48% in their best macro precision.

**TIRTH PATEL**

The Use of NLP Techniques in Static Code Analysis to Detect Weaknesses and Vulnerabilities.

*Tirth Patel (#6893783)*

*Master of Software Engineering (Student)*

*Concordia University*

*Montreal, Canada*

*tirth\_it@yahoo.com*

*SOEN-6611 Software Measurement*

1. BACKGROUND

NLP techniques is good for routine testing and static analysis of code. It can be used automatically to analysis of source code with safety to find vulnerabilities and weakness in code. MARF’s NLP framework and MARCRAFT application is used for this task.For quick scan in known type of file in large collection of file authors also did binary analysis. Approach for static code analysis was in its core principles,the knowledge base, machine learning categories, and the high-level algorithm. Core methodology includes *n*-gram and smoothing techniques. Primary knowledge base was CVE selected test cases. To validate the approach they have used NIST data set, so they can ensure their work. This was the first time NLP and machine learning‘s combination is used for static code analysis and the first result is published during the SATE2010 workshop. This approach uses character model for tokenization, so each printed character is a token. The matrices are learnedand stored from the known vulnerable code samples. The testing then does the sameof looking up the *n*-gram frequencies in the learned matrices instead of updating thematrices for the likely CVE or CWE class. In a way this approach was similar with classical natural language identification task. In these experiments unigram alone was used because it has produced good precision and they are the fastest among all other but signal pipeline.

**CHILAT SHAH**

THE USE OF MACHINE LEARNING WITH SIGNAL AND NLP PROCESSING OF SOURCE CODE TO FINGERPRINT, DETECT, AND CLASSIFY VULNERABILITIES AND WEAKNESSESS WITH MARFCAT

*Use MARF framework and MARF application to present a machine learning approach to static code analysis and fingerprinting to find weaknesses related to security and software engineering, and NIST’s SATE 2010 static analysis tool exposition.*

*Chilat Shah (6780024)*

*Master of Software Engineering (Student)*

*Concordia University*

*Montreal, Canada*

*Chilatshah1@gmail.com*

*SOEN-6611 Software Measurement*

1. BACKGROUND

MARFCAT is a MARF-based code analysis tool which is presented at the Static Analysis tool exposition (SATE) workshop 2010 and collocated with the Software Assurance forum.

The methodology behind static source code analysis contains two core principles: Machine learning and Spectral and NLP techniques.

This approach use Signal processing techniques, with source code as signal. Here it shows system example of file with weakness and uses signal processing techniques from CVE-selected test cases to compute spectral signatures which is learn by MARFCAT. At the test to find similarity or distant between each file in the known trained on weakness laden files. With methodology can find that some signature based antivirus or IDS software systems detect bad signature, while using machine learning and signal processing algorithms can find which combination gives the highest precision and best run time.

**AVI MODI**

# Choosing Best AlgorithmCombinations for Speech Processing Tasks in Machine Learning Using MARF

This article provides information about the different possible algorithm combination that can be used to process speech in Machine Learning using MARF. The core purpose of this study is to identify speaker, gender and accent through Machine Learning.MARF uses SpeakerIdentApp as a testbed which can be used as a tool for comparing different algorithms as well as it allows dynamic module selection based on available configuration options.The process starts by loading a sample, pre-processing it, extracting important features and finally it classifies who/what the subject is. The SpeakerIdentApp has wide range of algorithms for pre-processing, feature extraction and classification.Some most commonly used algorithms for this system are FFT, LPC, cosine-similarity, Chebyshevetc .The output for this process is a unique 32-bit integer which reflects the subject details.. All the algorithms(modules) are tested using the default parameters.

**MOHIT PUJARA**

**Writer Identification Using Inexpensive Signal Processing Techniques**

In this paper, it proposed to use novel and classical audio and text signal-processing otherwise it will be “inexpensive” for fast writer identification tasks of scanned hand-written documents “visually”. “inexpensive” is refers to the efficiency of the process of identification in terms of CPU cycles during preserving accuracy for preliminary identification. This comparative study of combinations of multiple algorithms in a pattern recognition pipeline implemented in Modular Audio Recognition Framework (MARF). Instead of extracting fine-grained features out of the classification, we can identify “visual” by “looking” at the hand-written document.

Most of the writer identification techniques rely on classical tools, methodologies and algorithms in handwriting recognition such as skeletonizing, contouring, line-based and angle-based feature extraction and many more. These techniques are highly accurate but its time consuming for large volume of digital data of handwritten material for its preliminary or secondary identification of who may have written.

To overcome from this problem, by looking at hand written text as a process of identification we simulate “quick visual identification” of the hand writing of the writer. For this process we build sample pages either in 1D or 2D arrays of data apply 1D or 2D loading using various loading methods. Similarly, for filtering we flatten a 2D array into 1D prior feature extraction after that we continue the classical feature extraction, training and classification tasks using comprehensive algorithm set within Modular Audio Recognition Framework (MARF)’s implementation by treating as a wave form for each hand-written image sample. Here, 1D as it is baseline storage mechanism for MARF which consumes less storage during sufficient to achieve high accuracy in the writer identification task.

To enable the experiments in this work we need to modify some options such as MARF’s Pipeline, WriterIdentApp, Resolution.

In conclusion, some testing are still underway and expected to complete soon because it consists combinations of two writing due which is about 600 runs per loader of the experiments. Some of the results came fast but accuracy is very low that is 20%. Therefore, authors are reviewing them as they resolve the faults in the implementation and data and summarize positive or negative outcomes.