Global Plagiarism Management Through Intelligence of Hawk Eye

Puri Karuna and Mulay Preeti
Department of Computer Science & Engineering, Symbiosis Institute of Technology,
Near Lupin Research Park, Gram: Lavale,
Tal: Mulshi, Pune - 412 115, Maharashtra, India; karuna.puri@sitpune.edu.in, preeti.mulay@sitpune.edu.in

Abstract

Background/Objectives: Rising Plagiarism incidences among students has become a common source of stress for academic community. This research is undertaken to reduce plagiarism in programming labs of universities across globe, faculty evaluation system development, and to remove social media use in labs. Methods/Statistical Analysis: Students of 21st century copy and exchange software codes available online with each other and at times even pay third-parties for timely submission of lab-assignments. They feel exchanging assignments with others maintain their social rapport/image among their friends. Our research initiated with Hawk-Eye that’s a mobile based plagiarism detection system easy to handle and maintain. By further research we had observed that Hawk-Eye combination with Cohort-Intelligence (CI) would lead to design of more reliable plagiarism evaluation system. As Hawk’s penetrating eyesight help in successful plagiarism detection and CI self-supervising nature helps to prevent students plagiarized behavior. Findings: The key motivation behind writing this paper is to highlight the fact that plagiarism in today’s cyber age is contributing as a global business option in academic areas. So the design of evaluation systems in Universities should be flexible enough to lessen the overall incidence of plagiarism and motivating its students for not breaking University rules and regulations. Applications/Improvements: Hawk Eye along with CI would prove to be an efficient system to design an incrementally learning evaluation system as a measure in Universities to stop plagiarism that would continue to evolve every year with new semester.

Keywords: CI, OCR, Optimization, Plagiarism, Tokenizing

1. Introduction

Incidence of Plagiarism is ‘growing faster’, ‘growing stronger’; and ‘growing sustainably’ which are the three key requirements to survive and thrive in today’s dynamic global marketplace. The increasing concern of Plagiarism is growing among Universities these days. Most common form of plagiarism frequently observed among students these days in programming labs of universities is code based plagiarism also called code-cloning. Increase in availability of third party sources for contract based plagiarism is helping students to solve lab assignments is another source of this kind of plagiarism in universities apart from digital resources. Plagiarism is contributing to businesses as a global option in this way in this cyber age. Hence all this urges for a mechanism to keep a check as well as prevent growing incidences of code-cloning in universities labs. As it is visible from [Figure 1] a broad overview of complete system in form of system mind map. [Figure 2] gives flow of execution for complete system.

1A Plagiarism Detection System is a convenient mobile based plagiarism detection system easy to carry and manage. This system uses OCR (Optical Character Recognition) capability for image preprocessing and to remove disturbance in order to extract required code syntax keywords from image of lab-assignment submitted by student that is clicked by faculty to check plagiarism. Plagiarism detection Algorithms plays a major role to remove all unnecessary information like comments etc. from submitted lab-assignment. Further Plagiarism
detection systems like Viper or Plagiarism Checker mobile based applications could be used to detect level of plagiarism. [Figure 3, 4] given an overview of flow of execution of Hawk Eye System.

After plagiarism detection next step is prevention hence the concept of CI comes into play. For prevention different Cohort Analysis tools can serve the purpose and based on students behavior analysis CI procedure would be implemented in order to categorize behavioral patterns of different students. After classification of students behavior into different categories like low, medium and high level plagiarism appropriate measures could be taken to remove roots of plagiarism from students. [Figure 5, 6]

The overall Equation for complete System can be formulated as:

\[
\text{Hawk Eye + Cohort Intelligence} = (\text{Efficient} + \text{Flexible} + \text{Preventive}) \times \text{PLAGIARISM DETECTION SYSTEM} \quad (1)
\]

2. Proposed Methodology

In this section the flow of execution of Hawk Eye mobile plagiarism detection system as well as Cohort Intelligence System for analysis of behavioral distribu-
Figure 3. Part-I: Flowchart for Hawk-Eye System.

Figure 4. Execution Flow for Hawk-Eye System.

Figure 5. Part-II: Flowchart for Cohort Intelligence System.
tion patterns of students in a cohort on basis of detected percentage of plagiarism from Hawk Eye System is discussed in brief.

The system architecture for the required Hawk Eye system along with Cohort Intelligence concept can be seen in [Figure 7].

2.1 Hawk-Eye System

2.1.1 OCR Engine Working

Smartphone's inbuilt scanning capability using apps like OCR Instantly, Cam-Scanner etc. could play a vital role
in proposed Plagiarism Detection System for extraction of relevant text/keywords from clicked snapshot of lab-assignment by faculty. Thus to summarize the complete process of OCR Engine:-

1. Image Preprocessing/Enhancement for better image clarity & removal of all image disturbances as far as possible.
2. OCR for keywords extraction editable text format from clicked image.
3. PDF, JPEG etc. format could be used as one of the appropriate format for saving and further editing.
4. Option of sharing also makes easy exchange of processed content from one device to another.

Students may also opt. for copying and exchanging of handwritten plagiarised codes. To deal with such plagiarism incidences OCR concept need to extended to IWR \(^3\) (Intelligent Word Recognition) which is used basically for recognition of handwritten words.

### 2.2 Plagiarism Detection Methods\(^6\)

#### 2.2.1 Abstract Syntax Tree

A program hierarchical representation or tree like structure can be given using AST \(^7\). Mathematical operators, function calls or other programming structures represents nodes of an AST, and variables or constants are the leaves of an AST. Optimization is done by compilers on tree like structure of program even before generation of a lower-level code. Hence this property makes AST a useful method in plagiarism detection.

##### 2.2.1.1 Limitations of AST

Tree like structure complexity increases with increase in number of lines of code, hence process of comparison between trees becomes more challenging and time-consuming process. Few generic modifications in code like code insertion or reordering makes AST \(^4\) based systems insignificant for plagiarism check.

#### 2.2.2 Tokenizing String based system

After pre-processing phase which removes comments, white spaces etc. the complete program variables could be divided into common token. Then a comparison is performed using string sequences. It performs a string-based comparison using the Karp-Rabin algorithm \(^9,10\).

This algorithm mainly focuses on concept of hashing to compute hash value for more accurate/precise detection.

### 2.3 Plagiarism Detection Tool

#### Analysis Results for Submitted Lab Assignment

#### 2.3.1 PlagScan Plagiarism Detection Results

Detected Plagiarism Level/Percentage is 48% as per PlagScan report for sample submitted lab assignment.

Figure 8 gives summary report and Figure 9 shows comparison\(^11\) results of PlagScan for the submitted lab assignment by a student for analysis and detects level of plagiarism in submitted lab assignment.

### 2.4 Cohort Intelligence System

CI \(^12\) is a self-supervised learning mechanism of a candidate which may be inspired from other candidates in a cohort. A cohort \(^13\) is a group of candidates with certain similar behavior/characteristic over a certain time period. And Cohort analysis mainly focuses its attention on the behavior/activities of a particular cohort. It gets more interesting as cohorts are compared over a period of time. Cohort analysis performs relationships \(^14\) identification using characteristics of a population and behavior being followed by population. This makes every candidate to improve/evolve its and eventually the entire cohort behavior. The complete procedure of Cohort Intelligence to compute student plagiarized behavior for categorization of students into different cohorts is explained below:

**Procedure for Cohort Intelligence Algorithm\(^12\)**

- Initialize the different student behavior of code-plagiarism for detailed analysis.
- Initialize all other parameters like convergence parameter $\epsilon$, number of iterations $i$, sampling interval $S_i$, number of rounds $r$, and number of solutions each actor generates $t$.
- Calculate the probability $\text{prob.}$ of every code-plagiarism case that’s associated with the behavior being followed by every student in student cohort.
- Apply roulette wheel approach which selects randomly index from vector based on probability distribution to decide the behavior and qualities to follow from C available choices.
- Every student shrinks/grow its sampling interval of quality based on saturation condition.
• If there is no change in behavior of each cohort, the cohort behavior can be considered as saturated.
• If cohort converges to same saturated behavior even after maximum number of attempts then current cohort behavior for particular case of code-plagiarism can be accepted as final behavior, confirming the convergence of solution.
• Stop if number of iterations equals to cohort or cohort is saturated.

Hence the Use Case diagram for the complete system can be formulated as shown in [Figure 10].

2.4.1 Cohort Analysis Tools Results

Different Cohort Analysis Tools like RJMetrics, Excel etc. could be used to design Cohort Buckets as visible in [Figure 11]. Then using CI Algorithms suitable evaluation measures can be taken. The proposed evaluation system
can prevent increasing plagiarism incidences among students of different streams in different universities.

[Figure 11] gives design of Cohort Buckets for analysis of student behavior based on detected level of plagiarism. [Figure 12] discusses in brief about various evaluation measures taken by faculties based on analysis of particular category of student behavior from different cohort buckets in order to prevent plagiarism.

As we can see in [Figure 13] with each year of upcoming batch of students in C Lab the trend in plagiarism will decrease if Hawk Eye system is followed systematically. This learning curve will help faculties also to analyze the effectiveness of the evaluation measures taken based on analyzed student behavior formulated from different cohort buckets that evolves every year with new batch of students.

3. Integration of Cohort Intelligence with Hawk-Eye

Cohort Intelligence basically helps in analysis of behavior which in turn can have applications in fields like data mining and big data analysis like students data in Universities. CI self-supervising mechanism is actually a learning and influence mechanism from other candidates in a cohort that can modify a student behavior in a cohort. CI also has great potential as an optimization strategy on the data set under analysis.

Contemporary approaches like Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Honey Bee Mating Algorithm are limited to natural behavior of the living organisms whereas CI procedure is used solve complex optimization problems by considering behavior of most complex living organism i.e. humans. CI self-supervising learning mechanism makes it a better optimization strategy compared to other traditional approaches from performance point of view. The biggest advantage CI has over other algorithms is it is reasonable with respect to computation cost. The only disadvantage of CI could be it converges slowly and requires more parameters for tuning during sampling compared to others like PSO etc.

4. Conclusion

Hawk Eye System with Cohort Intelligence concept could prove to be valuable system to keep a check on growing incidence of plagiarism in universities. The growing cases of fraudulent representation of someone else work as one’s own need to be prevented before plagiarism becomes a source of global business opportunity for third party. This
transition of businesses from age of industrialization to present cyber age of plagiarism need to be analyzed before it spreads across complete education system. Apart from outsourcing various discussion forums, auction sites, feed aggregator are also constant sources of plagiarism for students for completing their lab assignments. This prevents students from developing problem solving and coding skills which in turn leads to high rates of unemployment in country. The growing trend of socialism among students is also a key motivator of plagiarism which needs to be constantly checked. Hawk Eye as an incrementally learning evaluation system. Algorithms like CI, Particle swarm optimization and M-CFBA could definitely play a significant role to prevent increasing cases of plagiarism across different universities. Even though plagiarism has great potential to grow as a global business option and contribute as a major factor for high unemployment rates and reduction in overall economic development of country. But Hawk’s keen observant eyesight along with indicative behavior for students as an outcome of Cohort Intelligence would prove as a concrete evaluation system across universities to lessen the overall impact of plagiarism. The incremental approach of assignments design by faculties with every upcoming batch would be an iterative challenge which could be resolved with Hawk Eye Cohort Intelligence concept. The learning curve could help faculties to draw inferences regarding the effectiveness of this evaluation system over time with every new batch in comparison to previous batch. All these factors urges for the need of an intelligent and incrementally learning evaluation system like Hawk Eye in academic areas in order to keep a constant check and prevent growing issues of plagiarism. Hence Hawk Eye plus CI concept as a system is beneficial for education industry in general; for students to improve their learning skills, for faculty to design effective evaluation measures and for universities to lessen overall impact of plagiarism in particular.

Figure 11. Cohort Buckets for Student Behavior Analysis based on Level of Plagiarism.

Figure 12. Designed Evaluation Measures based on designed category of Cohort Bucket.
5. Future Work

Future extension to this Plagiarism detection cum prevention system could be check for technologically advanced hidden layers like WhatsApp, Smartphones Camera, Bluetooth or other social networks which are possible sources of increasing cases of plagiarism among students.

Also this mobile based plagiarism detection system need to be extended to a desktop compatible system as evaluation process becomes difficult to analyze with length of codes on a mobile based system.

This concept of plagiarism detection and prevention could be extended to individual software industries in their respective firm as a mechanism for analyzing and checking of cloning of codes in their respective products before launching it into market. This in turn would create industry long term healthy relations with customers.

A group of expert academic researchers with expertise in field of Plagiarism can be formed for analyzing of all possible sources, types and cases of plagiarism. This would enable universities to effectively update their existing laws against plagiarism based on detailed analysis of all possible reasons and contributors of plagiarism.

Validations like F-measure, Precision, Recall, ROC Curve, and Confusion Matrix etc. can provide detailed analysis over successive semesters to universities regarding effectiveness of evaluation system to curb frequent incidences of plagiarism these days among students in universities.

6. References