# DETECTING PLAGIARISM IN A STUDENT'S ASSIGNMENT BASED ON WEB PAGES REFERENCED BY A STUDENT

ESTHER SANGEEDHA D/O EZRA ANANDARAJ
(Software Engineering)
HELENA PING MERING
(Software Engineering)

This project is submitted in partial fulfillment of
the requirements of the degree of Bachelor of Computer Science with Honours

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2004

- Computer Applications

## ACKNOWLEDGEMENT

The final year project is a coursework done for duration of nine months in partial fulfillment of the requirements for the degree of Bachelor (Hons) Computer Science, University Malaysia Sarawak.

We gained much experience and knowledge by doing this project. We now have a more in-depth knowledge in the area of text analysis and similar sentence structure detection. We would like to thank our supervisor, Puan Azni Haslizan bt Abd.Halim for her guidance, support and ideas throughout this project. You have been an inspiration to us.

We would also like to express our gratitude to our ex-supervisors from our former industrial training placement companies namely Mr. Hoe Kok Meng from the Technology Assimilation and Deployment Department, Mimos Berhad for his ideas and contribution that inspired us to take up this research topic.

To these and those not mentioned, everyone who played a part in our project, we thank you for your encouragement.

# TABLE OF CONTENT

# TABLE REGISTER

## DIAGRAM REGISTER

# CHART REGISTER

# ABSTRACT

## DETECTING PLAGIARISM IN A STUDENT'S ASSIGNMENT BASED ON WEB PAGES REFERENCED BY A STUDENT

Esther Sangeedha Anandaraj

&

Helena Ping Mering

Students in institutions of higher learning refer to many sources of information to complete assignments. More often than not, these sources are copied directly or paraphrased without proper citation. This is considered plagiarism.

There are many plagiarism detection systems that are able to detect plagiarism. The techniques used are accurate, effective and use the latest technology. However, they require users to have the knowledge of using the system and purchase the system before usage.

We attempt to implement a simple yet effective method using common programming concepts to detect plagiarism in a student's assignment based on the web pages referenced. Plagiarism detection is done at two levels, sentence and word level. A measurement of similarity (Plagiarism Index) is assigned to the student assignment after the detection process has been completed.

# CHAPTER 1 INTRODUCTION

## 1.0 Introduction

The task of detecting plagiarism in a student's assignment has been undertaken by educationists worldwide in educational institutions. The Internet, being the world's largest repository of information supplies a wealth of references to a student. With this at hand, a student references various websites and collects information from different sites. After some alteration, compiling and deleting, the assignment is ready to be handed in.

Generally, there are two approaches to detecting plagiarism in a collection of student assignment. One is, to perform comparison of every pair of document in the collection to detect a co-derivative. This is referred to as the n-to-n problem. In this research, the one-to-n problem is focused on. The single document taken is the student assignment. The collection of reference documents is web pages the student referenced to compare the assignment.

## 1.1 Plagiarism

Students in institutions of higher learning reference sources of information and use them for tasks assigned to them. Plagiarism is defined as using other's ideas and words without clearly acknowledging the source of that information. (Writing Tutorial Services, Indiana University, Bloomington, IN)

Below is an excerpt from Software Engineering: A Practitioner's Approach by Roger S Pressman, "Once the software is implemented in machine – executable form, it must be tested to uncover defects in function, in logic and in implementation", pp 90.

There are two ways of rewriting this excerpt. One is an acceptable paraphrase the other is not.

Acceptable paraphrase:

As soon as the software is changed into a form accepted by the machine, it must go through a testing phase to detect functional, logic and implementation faults.

(Pressman, 1992)

Unacceptable paraphrase:

The software must be tested once it is implemented in machine – executable form. This is done to uncover defects in function, in logic and in implementation.

The above phrase is unacceptable because:

i) The writer has only changed a few words and phrases, or changed the order of the original sentences.

ii) The writer has failed to cite a reference for any of the ideas or facts used.

Plagiarism is on the rise and occurs in the field of education and other industries like music and software. Plagiarism is committed for various reasons. But the fact remains that plagiarizing one's work and passing it off as your own is unethical. Plagiarism must be detected and measures should be taken to prevent it from reoccurring. In educational institutions, students plagiarize when doing their assignment because of a few factors. Among them are bad time management, a lack of confidence in their ability to author the assignment, and laziness. They feel the effort put in to acquire the references justifies the act of copying it with minor alterations. In reality though, students with higher GPAs are less likely to cheat than those with lower GPAs. (Rittman, 1996, Penn State Pulse Survey)

Plagiarism can be detected in two ways: manual and automatic. Manual checks performed to detect plagiarism, require the examiner to look through the document for the following signs: "lack of recent reference sources, unusual references, a mysteriously improved writing style between paragraphs and un-cited quoting of references". (Cyberplagiarism: Detection and Prevention, 2003)

Automated checking involves the use of a computer to detect plagiarism. In most cases, the sentences structure and words in the text are compared and manipulated. There

are many techniques developed to detect plagiarism in a student's assignment. They come in the form of Plagiarism Detection Systems (PDS). These techniques have been tested and proven effective in detecting plagiarism.

In local institutions of higher learning, students commit plagiarism. Even though there are clear guidelines and legislations against plagiarism, this does not hinder students from practicing the trade. In a survey done in University Malaysia Sarawak, almost 90% of the student body named the Internet as a main source of reference for their assignments. The candidates were given a test to do. The test required them to paraphrase an excerpt from an article in The Reader's Digest. Please refer to the questionnaire in the appendix. Of the 10 candidates tested, 8 out of 10 of them were caught plagiarizing the source. They copied verbatim blocks of text and failed to cite their source of reference.

## 1.2 Problem Statement

Currently, plagiarism is common in institutions of higher learning. Students do assignments as part of their coursework. As stated above, it was found that almost 90% of those interviewed relied on the Internet as the main source of reference. The Internet, being the world's largest repository of information contains web pages that cover a wide range of topics. The fact that information can be acquired just by a click of the mouse proves that students need not go through the hassle of searching for information manually.

For example, they do not need to go to the library and search for copies of journals in print. Almost everything is available in electronic format, online.

Students refer to web pages to complete their assignments. There is a tendency to copy directly from the web pages. This, and the failure cite their source of reference categorizes the work as plagiarized work.

In the local campus, lecturers assess student assignments for each course they conduct. In many cases, the lecturers detect plagiarism manually. A survey conducted among the lecturers found that 99% of the lecturers have never used any plagiarism detection system (PDS) before. The lecturers detected plagiarism in a student's assignment manually by comparing the similarity between web pages referenced by the student and comparing one student assignment to other student assignments.

This was done using word-by-word comparison. If a block of text proved to be identical, the lecturer was able detect the plagiarism committed. The lecturers also took into account the style of writing. An assignment is made up of text in a few paragraphs. In some cases, some paragraphs were written grammatically correct and were coherent. Other paragraphs had a haphazard layout and were unorganized. This created a sharp contrast between the paragraphs. Lecturers were able to detect the difference in the written structure of the assignment.

It was concluded that if there was a difference in the style of presentation of ideas between paragraphs, therefore, it was proven that a student committed plagiarism. Students were also not analytical and critical in their work.

The manual method of detecting plagiarism is time-consuming and not as effective as an automated method of checking. The lecturers have plenty of student's assignments to assess. Manual detection of plagiarism requires much time and energy. This taxes the lecturers. A lecturer may become exhausted after assessing a few student assignments. Exhaustion reduces his capability of effectively detecting plagiarism.

There are many PDS systems in the market. These systems can effectively detect plagiarism in a student assignment. However, the techniques used in these systems are complex. For example, a number of PDS systems use text classifiers and implement artificial neural networks to detect plagiarism. These systems require the user to have already acquired the knowledge beforehand to operate it and interpret the results returned by the system. It costs much to purchase such a system to detect plagiarism.

## 1.3 Project Objectives

The research project was undertaken to do an in-depth study on how common programming concepts could be used to detect plagiarism. A method that is simple yet effective in detecting plagiarism was proposed. The method proposed should be able to successfully detect automatically, the similarities between a student assignment and a

webpage referenced by the student. After detecting the similarities between the documents, the method should be able to assign a measurement of similarity to the student assignment. The measurement of similarity assigned to the document is called a Plagiarism Index (PI). The PI helps the lecturer assessing a student assignment determine the severity of plagiarism committed by the student.

## 1.4 Proposed Solution

As stated in the problem statement, the current approaches taken to detecting plagiarism are effective but complex. Therefore, the research project was undertaken to revise the approach taken to detecting plagiarism. As in earlier methods, plagiarism is detected based on the level of similarity between the text and reference text. This algorithm is also used in the proposed method. The sentence structures in a student assignment are studied and compared to a webpage referenced by the student. Common programming concepts are used to detect the similarity between both documents. The similarity detection between both documents was done at two levels: Sentence level and word level. At the sentence level, a sentence alignment technique was used. At the word level, a keyword comparison technique was used. The similarity detection was done at two levels because this increased the chances of detecting similar sentences structures and allowed a thorough checking to be performed on the documents.

## 1.5 Project Scope

The project scope is to propose a method that will be able to detect the similarities between a student assignment and a collection of web pages referenced by the student. Based on the similarities detected between both documents, the level of plagiarism committed in the student assignment can be measured. The method proposed must be effective or a little less effective than the techniques currently used to detect plagiarism.

The methodology used to detect similar sentence structures between two documents was done at two levels: Sentence level and word level. Sentence alignment was used to detect the similarity between two documents at the sentence level. Sentence alignment aligns sentences in the student assignment to sentences in the webpage referenced by the student. At the word level, similarity was detected using the keyword comparison method. In the keyword comparison process, words in the student assignment were compared to words in the webpage. Identical words were identified and eliminated from the student assignment.

A measurement of similarity was given to the student assignment. This measurement is called the plagiarism index.

## 1.6 Expected Outcome

The plagiarism detection method proposed should be able to detect the similarity between a student assignment and webpage referenced. Based on this, the severity of plagiarism committed can be determined. The proposed method is to be as effective or a little less effective than the current methods used to detect plagiarism in a student assignment. The lecturers can use the method to hasten the process of assessment of student assignments. The lecturers have a clear-cut way of detecting plagiarism. The method is more systematic than the current manual method used by the lecturers to detect plagiarism.

## 1.7 Chapter Overview

Each chapter documented in this report covers an aspect of the entire research done.

Chapter 1: Introduction.

This chapter gives a short tour on the entire project done. The main idea is painted: Propose a method that is simple yet effective in detecting plagiarism in a student assignment based on web pages referenced by the student. The method is implemented using common programming concepts.

## Chapter 2: Literature Review.

A review is done on existing PDS systems in the market. The method used by these systems to detect plagiarism is studied. The effectiveness of the systems are noted and compared. Works done by others in the field of plagiarism detection is also reviewed. The techniques used in other works of research are taken into account. They are applied to the research project undertaken where necessary. Examples of techniques used by others are preprocessing, sentence alignment and identical word comparison.

## Chapter 3: Experimental Method

The whole research structure is presented in finite detail. The research is divided into 5 modules. The first module is the preprocessing mechanism. The input used in the research is processed into a format easy for the computer to handle. The second module extracts keywords from the collection of webpages used by the student. Keyword lists are created in this phase. The keyword lists are used in the fourth module for comparison purposes.

The third module implements the sentence alignment technique. It aligns the sentences in a student assignment to sentences in a webpage. Similar and identical sentences are identified. The fourth module is keyword comparison. The keyword lists created in the previous module are used to remove words in the student assignment that match the words in the keyword list. The fifth module calculates a PI and assigns it to the student assignment. This allows one to measure the level of plagiarism committed by the student in his work.

Chapter 4: Method Proposed.

This chapter highlights the work pioneered in this research project. The keyword extraction and keyword comparison method are introduced here. These methods were proposed to save computational time. These methods are based on identical word comparisons used by other researchers in this field. A precision index formula and a formula to calculate the plagiarism index are also proposed.

Chapter 5: Experimental Results.

Datasets of student assignments and their accompanying web pages are used to test the efficiency and accuracy of the method proposed. All the sets are processed and assigned a PI. The results are studied. All observations made based on the results are justified.

Chapter 6: Conclusion.

Based on the findings, a conclusion is drawn from the whole research project. The conclusion may be the method proposed is as effective or a little less effective compared to current plagiarism detection methods or otherwise. Limitations, constraints and future work that can be done on the research project are mentioned.

# CHAPTER 2 LITERATURE REVIEW

## 2.0 Introduction

This chapter focuses on plagiarism detection systems (PDS) currently in the market. Many educational institutions use these systems to detect plagiarism. The various plagiarism detection system features are highlighted. The advantages and disadvantages of the system are identified. However, only PDS systems that detect plagiarism in a student assignment based on web resources are reviewed here. This chapter also introduces a new approach taken toward detecting plagiarism. It is a combination of a sentence alignment technique and keyword comparison. The keyword comparison technique is a new technique pioneered in this project. Further information on this technique can be obtained in Chapter 4: Method Proposed.

## 2.1 Review on Plagiarism Detection Systems

There are many commercial plagiarism detection systems available in the market. They are efficient and accurate in detecting plagiarism. Plagiarism using web resources as a source is common. Many steps have been taken to reduce the number of plagiarism cases. One of the steps taken is the creation of plagiarism detection systems. These systems are based on the latest technology and most efficient algorithm used to detect similarities between documents. Reviewed below are the PDS systems that detect plagiarism based on web sources referenced by a student to complete an assignment. These systems are commercial systems that can be purchased or require a license to use on them.

### 2.1.1 EVE2

EVE2 is a plagiarism detection system that enables educationalists at all levels of the education system to determine if plagiarism has been committed using material from the World Wide Web (WWW). EVE2 accepts student assignments in the format of pure text, Microsoft Word documents, and Corel Word Perfect.

This system implements the most advanced searching tools available to locate suspect sites. It searches the entire net for suspect sites. If it is successful in finding such a site, direct comparison is performed between contents in the site

and the student assignment. If evidence is strong, this indicates that a major part of the website has been copied. In this case, the site's Universal Reference Link (URL) is recorded.

This system returns a report on the results acquired after searching the web. The percentage of plagiarism committed plus an annotated copy of student assignment highlighting the suspected plagiarism in red is displayed in the report. The report also returns links to webpages from suspect sites, which a student may have referenced to complete the assignment. (Essay Verification Engine (EVE))

### 2.1.2 CopyChecker

The PDS helps prevent accidental plagiarism. This means it helps the user avoid the temptation of copying verbatim from the source. This approach is taken based on the assumption that *"plagiarism and incorrect paraphrase is generally caused by working too closely with the source material"*. ( CopyChecker)

The system allows the user to work closely with the sources. The interface provides two view windows. The workspace of the user is displayed in a window on the left of the screen. The source text is displayed on the right. To prevent verbatim copying of the text, the keyboard shortcuts: CTRL-C and CTRL-V, which stands for copy and paste are disabled. This encourages reinterpretation