PLAGIARISM DETECTION SYSTEMS – AN EVALUATION OF SEVERAL SYSTEMS

Ahmad Hashem
Open University Malaysia, Kuala Lumpur, Malaysia

ABSTRACT

In the Open University environment where students are not centrally located and are not under any direct supervision the potential for plagiarism definitely exists. The very technology that facilitates open learning also allows easy exchange of papers among peers. Students can also easily access “paper-mills” where essays can be quickly customized to suit requirements, for a fee. There is also the vast information residing on the Internet ready for creative reuse. In the light of all these temptations the Open University of Malaysia (OUM) is exploring the use of technology to educate students and deter plagiarism. One approach that appears promising is to use a good commercial plagiarism detection system. The system, by being able to detect cases of plagiarism would serve as a deterrent and hopefully contribute towards inculcating the culture of honesty. This paper presents the findings from a small study using two detection systems, a commercial Plagiarism Detection System, MyDropBox and a simple and free automatic file comparison system, Pl@giarism. The results show that while the automatic system is very useful, the simple and free system can be adequate for most purposes. Recommendations on a suitable system for the OUM context are then made based on these findings.

PLAGIARISM DETECTION SYSTEMS

The Internet has become a vast repository of easily accessible knowledge. Most are freely available via any Internet Search Engine while others can be accessed by members only (free or subscription). Students are availing themselves to this vast pool of knowledge. This ease of access has spawned an undesirable culture – that of the ‘cut-and-paste’. Passages from the Internet are copied and pasted to be presented as the student’s own work without substantial rewriting or due citation. The diligent teacher can look at clues such as stylistic changes to spot “cut-and-paste” portions of essays submitted by students. But maybe not all teachers are diligent or have time to be diligent, thus, to ensure uniformity in the treatment of all submitted essays some form of automatic detection of plagiarism must be employed. Deterring plagiarism and training students to abhor plagiarism would certainly contribute to improving quality in higher education.

Plagiarism Detection Systems can help find sources of online passages that occur in an essay. The assumption behind all these systems is similar. Material for essays can be downloaded from freely available Internet resources as well as proprietary online databases and presented in a student’s essay. By comparing phrases in an essay with those from online resources occurrences of online material in a student’s essay can be discovered. It is important to remember when using any Plagiarism Detection System that the discovery of similar passages in an essay to online resources must necessarily imply plagiarism can never be absolutely decided by any Plagiarism Detection System. Most systems would highlight phrases in an essay that is similar or almost similar to online resources and some sort of similarity index would be assigned depending on the extent of the similarity. Establishing plagiarism is never easy. There are issues such as intent, incorrect citation, unfamiliarity with the language and perception of what constitute plagiarism that need to be considered (Ashworth, et. al., 1997). Thus, after discovery it is best left to a human reader to decide on the next course of action.

Of the many detection systems on the market that claim to help reduce the incidence of plagiarism, perhaps the most “famous” is Turnitin (www.turnitin.com). Many universities throughout the world have formed consortia to better use this software for detecting plagiarism. One such example is the Plagiarism Detection Centre, based in the University of Northumbria that has been set up by the Joint Information Systems Committee (JISC) of all the Universities in the United Kingdom (http://www.jiscpas.ac.uk/). The Research Policy and Practice Committee
now require that all students’ essays be submitted to the JISC detection service. It should be noted that many of these systems are not free but are subscription-based according to several financial models. Also, issues such as intellectual property rights of students’ essays must be considered with some of these systems.

A COMPARISON OF TWO PLAGIARISM DETECTION SYSTEMS

At the OUM it has been felt for some time now that plagiarism cases have been on the increase. There has not been any quantitative study to support this but anecdotal evidence abound. In order to get a better understanding of the extent of plagiarism and perhaps also gain some idea on deterrent approaches a small study was carried out. A small sample of essays from several OUM Learning Centres was collected for analysis. In total 100 essays from four different subjects and from three Learning Centres were analyzed. These essays were hard-copy documents that have been handed in by the students to their tutors for marking. On average an essay is about 10 A4 pages long.

The essays were analyzed by a typical automatic Plagiarism Detection System and in the process aspects such as ease of use, adaptability to the OUM workflow and the extent of plagiarism in the OUM were assessed. An automatic system that can detect similarities between pairs of essays in a database was also tested.

There have been many studies comparing the many plagiarism detection systems. The Virtual Academic Integrity Laboratory (VAIL, 2004) report is quite comprehensive. Many reports and discussions on plagiarism detection systems can be found on the web. For this study a subscription-based detection system, MyDropBox (www.mydropbox.com), a plagiarism detection system that has received some good reviews was compared with the free and semi-automatic Pl@giarism (www.plagiarism.tk) developed by the University of Maastricht, Belgium. It is used extensively by the Law faculty there. The free system is being evaluated here because free is usually the most cost-effective.

MyDropBox provides 150 days free use for testing purposes. Submitted essays are first compared automatically against a group of selected essays or against every essay in its proprietary databases, as well as resources on the Internet. MyDropBox claimed it has access to nearly 400,000 previously submitted and stored essays as well as access to certain subscription-based digital libraries. The MyDropBox submission process is as follows:

1. Tutor submit essays to MyDropBox
2. MyDropBox checks essays against internal collection of essays, sources on the Internet, open and subscription based online libraries and other proprietary sources.
3. MyDropBox generates a report highlighting sections of the essays with similar passages found from other sources.
4. Tutor scrutinizes report and decides whether plagiarism has taken place.

Pl@giarism is a simple program that automates the process of determining similarities between pairs of essays by comparing three word phrases in each. An essay is paired with each essay in the folder in turn. Pl@giarism does not automatically check against sources on the Internet, but there is a provision for selecting and submitting one phrase at a time to an Internet search engine for comparison with Internet resources. Pl@giarism is a useful standalone detector for peer-to-peer copying.

METHOD

All the essays were scanned and converted to text document using an Optical Character Recognition program. The scanned essays were then edited to correct for inaccurate recognition and saved to one folder. Preliminary scrutiny indicated that the most likely sources of copied material would be the Internet and the peers. The scanned essays were treated as follows:

1. Submitted to the MyDropBox Plagiarism Detection System for comparison against each other, against all essays in the MyDropBox database, against all subscription databases accessible to MyDropBox and against resources on the Web.
2. Submitted to Pl@giarism for cross-comparison against each other in order to detect similarities among the essays in the sample.

The number of essays with similarity to Internet resources for MyDropBox gave an indication of the extent of plagiarism from the Internet. Also, the number of pairs of essays in the database with some similarities identified by each of the system gave an indication of plagiarism among peers.

RESULTS

Both programs proved easy to use and most cases of shared contents among peers were found. Since MyDropBox compare essays against a database as well as the Internet, a report for each essay can take some time. In this study, completion times were quite short because comparison was limited to a small selection of files (the scanned essays) as well as contents on the web only. Typical times were about 2 minutes per essay. When the scope of comparison was extended to include its proprietary databases and online libraries then the turn-around time can be anywhere between 10 hours to 24 hours, depending on the length of the essay. Typically a thorough report for one essay in this study can be produced after about 30 minutes of analysis. Pl@giarism does not compare against online resources (unless a passage is manually selected and submitted), thus, cross comparison of pairs of files typically take about 5 minutes for thirty essays. The three-word comparing algorithm used by Pl@giarism proved adequate to detect cases of minor structural modification of sentences. Pl@giarism discovered all files that show some similarity just as well as MyDropBox.

The essays for this study are in Bahasa Malaysia and English. Some subjects at the OUM are taught in Bahasa Malaysia and some in English. It was expected that the paucity of technical contents in Bahasa Malaysia may result in many more original work in this language compared to essays in English. However, this was not the case. It was observed that almost all the essays in areas of technology and information technology contain material from the Internet. The difference is in the extent. In some essays almost all the material are copied from the Internet without any attempt at modification. One such example for the subject Basic Concepts in Information Technology is shown in Example 1 below:

EXAMPLE 1

Wikipedia http://ms.wikipedia.org/wiki/Perbualan_Pengguna:Rizman_Suzaidi

Perisian Sumber Terbuka merujuk kepada keterbukaan untuk mengkaji, mengubah, dan memperbaiki reka bentuk sesuatu perisian komputer (termasuk kod sumbernya) menerusi lesen sumber terbuka

Student Essay

Perisian Sumber Terbuka merujuk kepada keterbukaan untuk mengkaji, mengubah, dan memperbaiki reka bentuk sesuatu perisian komputer (termasuk kod sumbernya) menerusi lesen sumber terbuka

This student copied material for his essay from several Wikipedia articles and passed it off as his own work without any attempt at rewriting or citation.

Example 2, also for the subject Basic Concepts in Information Technology, shows what could probably be peer-to-peer copying, albeit with some slight modification. These two essays were found by Pl@giarism using its three-word matching strategy. The two students are from the same region but are registered at two different learning centres. The matching phrase is not exactly similar but upon careful reading the two passages can be seen to be similar.

EXAMPLE 2

Student from Learning Centre Sarawak

Tidak dapat dinafikan bahawa kepentingan perisian sumber terbuka (OSS) dapat membantu menjimatkan perbelanjaan syarikat atau kerajaan dan ia semakin mendapat perhatian sebahagian besar kerajaan di seluruh dunia kerana ia dapat menjana
KEPENTINGAN perisian sumber terbuka (OSS) tidak dapat dinafikan terutamanya bagi membantu menjimatkan perbelanjaan syarikat atau kerajaan dan ia semakin mendapat perhatian sebahagian besar kerajaan di seluruh dunia.

Whether the similarity of these two passages imply plagiarism must be further investigated. Also at issue would be who plagiarized from whom?

It was indeed surprising to note that there were many resources in Bahasa Malaysia on the Internet. The most popular source of material for Information Technology and most technology-based essays is Wikipedia in Bahasa Malaysia and online Bahasa Malaysia newspapers.

Example 3 shows an exact match between a student’s essay and an article in Wikipedia.

**EXAMPLE 3**

**Students Essay**

**Passive solar techniques make use of the steady supply of solar energy by means of building designs that carefully balance their energy requirements with the building's site and window orientation.** However, it seems that we don’t have much choice if we bought an intermediate house that does not have a steady supply of sunlight.

**Source of text**

http://www.newenergy.org/sesci/publications/pamphlets/passive.html

Passive solar heating and cooling represents an important strategy for displacing traditional energy sources in buildings. Anyone who has sat by a sunny, south-facing window on a winter day has felt the effects of passive solar energy. **Passive solar techniques make use of the steady supply of solar energy by means of building designs that carefully balance their energy requirements with the building's site and window orientation.** The term “passive” indicates that no additional mechanical equipment is used, other than the normal building elements. All solar gains are brought in through windows and minimum use is made of pumps or fans to distribute heat or effect cooling.

This simple survey was not meant to be exhaustive. Rather, it was meant to provide some insight into the type of plagiarism taking place and to help define probable strategy for further work in plagiarism detection. Table 1 shows a summary of the results for MyDropBox and Pl@giarism.

<p>| <strong>Table 1 Comparison between MyDropBox and Pl@giarism</strong> |</p>
<table>
<thead>
<tr>
<th>MyDropBox</th>
<th>Pl@giarism</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Number of essays using freely available material from the Internet without proper citation.</td>
<td>23</td>
</tr>
<tr>
<td>2. Number of pairs of essays containing some similar material, with perhaps minor changes.</td>
<td>12</td>
</tr>
<tr>
<td>3. Number of essays using material from proprietary databases and subscription-based online libraries</td>
<td>0</td>
</tr>
</tbody>
</table>

MyDropBox managed to find 23 essays that have material similar to those on the Internet. It also found 12 pairs of essays in the sample that have similar passages whereas Pl@giarism found 13 pairs. The one extra pair found by Pl@giarism is the case shown in Example 2 above where a slight degree of rewriting has been carried out and the phrases are no longer exact. This is perhaps important in situations where students would modify their material slightly to differentiate from the original source. It can be seen that when comparing between pairs of essays in a
There could perhaps be cases of using material from books but for the present checking against all books is not yet possible. MyDropBox claim to have access to EBSCO database and the OUM students are also provided access to this and other library databases. The finding that no similar passages were taken from these proprietary databases could imply that for this group of students it easier to use material that is freely available from the Internet rather that getting material from e-books. Note that in Table 1 there can be an overlap in that one essay contain material from another essay which in turn contain material from the Internet. The overlapping cases were not separated.

FREE TOOLS

Besides the specialized plagiarism detection tools, the diligent tutor can of course manually submit suspicious phrases to any of the free Internet search engines, such as Google. A sentence or a paragraph that appears stylistically different from the rest of the article or from expectation based on previous experience with a particular student’s style is a likely candidate for submission. Bear in mind that Google impose a limit of not more than ten words for each phrase. Thus, the suspicious sentences must be cut to the appropriate length and submitted. In tests, all passages highlighted by MyDropBox as similar to Internet sources, when submitted manually to Google yielded most of the likely sources.

The readily available Microsoft Search is another useful tool for searching for the occurrence of a named folder, files or phrase. For this study the option to search for all occurrences of a particular phrase in all files was chosen. A suspicious phrase was submitted to Microsoft search and upon completion of the search a list of all the files that contain this phrase was returned. In the context of plagiarism detection the number of files containing similar phrases can give an indication of peer-to-peer copying.

Plagiarism in computer programming is a sub-specialization of plagiarism as a whole. A computer program to carry out a certain task will most likely have many similarities when written by various programmers, even if written independently. Clough (2000) discusses these issues in great detail in his excellent internal report. Computer programs can differ because of “extraneous” remarks or comments and use of different labels for variable names and other identifiers. Before analyzing for similarities computer programs are first stripped of all non-essential details such as comments or remarks, white space and followed by replacement of variable and identifier names with tokens. These “cleaned-out codes” are then compared for extent of similarities either among peers or with Internet sources. Plague, http://www.csse.monash.edu.au/projects/plague/ and JPlag, http://wwwipd.ira.uka.de:2222/ are examples of detectors that treat computer programs in this way. MOSS (Measure of Software Similarity), http://www.cs.berkeley.edu/~aiken/moss.html pre-clean differently but it uses similar comparison techniques to normal plagiarism detectors.

LESSONS FROM THIS STUDY

As can be seen in Table 1 about 23% of essays appear to include material found on the Internet. Also, about 13% share (knowingly or otherwise) essay material. At first glance these figures would appear to be quite high. However, this situation is not peculiar to the OUM only as plagiarism is now universally acknowledged to be ubiquitous (Lindsey S. Hamlin and William T. Ryan, 2006). This is also not a new phenomenon. In a major study, Newstead, Franklyn-Stokes and Armsadle (1996) showed that about 54% of undergraduates admitted paraphrasing from another source.

Full-blown matching of essays against internal databases and all accessible online resources takes a long time. MyDropBox would take anywhere between 8 hours to 15 hours to produce a report. In contrast matching against free Internet sources was very fast, usually less than one minute for a ten-page essay. In the context of the OUM, with 50,000 students and each submitting 3 essays the turn-around time may be very much longer than this if all essays were submitted for plagiarism detection at once. This type of submission, that is, all 50,000 essays submitted for checking at once is not the norm. Typically, submission to most plagiarism detections systems are made in small numbers at a time, usually on a per-class basis or upon...
suspicion only. It is the duty of the tutor to read carefully each student’s essay and submit any suspicious essay for plagiarism detection. In the Open Distance Learning context tutors are usually part-timers and some (albeit a minority) may not be fully committed to performing tasks as diligently as expected. This would give rise to non-uniformity in treating students’ essays which in turn can mean some students being penalized for plagiarism because their essays were submitted to some detection system while others who also plagiarized may be getting high marks.

In this study it was found that the freely available content on the Internet is the major source of plagiarized material. Most cases of plagiarism involve taking material verbatim from another source without any attempt at rewriting or citation. The one case that shows some rewriting only changes the arrangements of the same words. It appears that students prefer the cut-and-paste method to reading and comprehension and rewriting.

RECOMMENDATIONS FOR THE OUM

Plagiarism definitely occurs at the OUM and unless measures are taken to deter this practice the quality of education may be seriously compromised. While manual checking against the Internet can definitely help determine sources of suspicious passages in a submitted essay diligence and sufficient time is required. Automatic Plagiarism Detection Systems can speed up this discovery process. Of the two systems tested, MyDropBox was found to be the more useful. It is automatic whereas Pl@giarism requires the tutor to actively identify and submit phrases to a search engine. However, Pl@giarism is free but MyDropBox charges an annual fee that depends on the number of full-time equivalent students. If a tutor is conscientious and willing to do the extra work of identifying suspicious phrases in a student’s essay then Pl@giarism is good enough. In fact the free Microsoft Search (advanced search for text) and any Internet search engine (such as Google) would be adequate.

To enable detection of copying among peers or similarity to submissions of previous cohorts all essays must be submitted in soft copy to a central database by a set deadline. After this deadline the plagiarism detection process will be carried out. There must be a policy on the treatment of essays submitted after the deadline. Also, for fairness, plagiarism implementation must be a case of “all or nothing”. Either all essays are submitted automatically to a plagiarism detection system or none at all.

At the OUM the time from submission of essays by students to the time when marks must be submitted by the tutors is only about 2 weeks. The tutors would therefore be hard-pressed to read essays and select suspicious sentences for submission to an Internet search engine. Thus, an automatic system is a must for the OUM and other institutions with a large number of essays to analyze. The tutor reads and marks an essay after obtaining a report from the plagiarism detection system.

From this study it was found that at minimum the plagiarism detection system must be able to perform comparison between essays in a database as well as automatically extract and submit phrases from each essay to an Internet search engine. A report of similarities found and the location of the Internet sources must be provided. Perhaps a plagiarism detection system with this restricted scope can be built in-house. There are several advantages to own-built system with the most important being the certainty of fit into the existing workflow. Also, extras such as customized reports can be easily added-on. This was the approach adopted by Monash University with their Damocles (http://viper.csse.monash.edu.au/damocles/about/) even though most universities in Australia used Turnitin.

CONCLUSIONS

From the small study it was found that plagiarism from Internet resources is widespread. There is also a high incidence of peer-to-peer plagiarizing. Automatic Plagiarism Detection Systems greatly speed up the process of identifying essays with similarity to Internet sources. It cannot, however, rule on whether plagiarism has taken place. All Plagiarism Detection System can act as a deterrent and can contribute greatly to improving the quality of education. In order to ensure minimal disruption to the current workflow any Plagiarism Detection System must have a good fit. An own-built system that can be incorporated into the OUM Learning Management System would be desirable.
REFERENCES


