12th IMEKO TC1 & TC7 Joint Symposium on
Man Science & Measurement
September, 3 – 5, 2008, Annecy, France

Perception of Plagiarism and Fraud in Education

Kees van Noortwijk [1], *Richard V. De Mulder* [2]

[1] Erasmus University, Rotterdam, The Netherlands, vannoortwijk@law.eur.nl
[2] Erasmus University, Rotterdam, The Netherlands, demulder@law.eur.nl

Abstract: Plagiarism and fraud by students have increased considerably in the past decade. Information technology plays an important role here, because it provides techniques that make it possible to copy text, images and other materials almost effortlessly. Educational institutions are aware of this, and try to halt it. The problem of monitoring and dealing with fraud and plagiarism is a matter of perception in two respects:

1. it is often difficult to identify fraudulent student papers because of the large scale and intelligent techniques for cheating, and

2. students who are blamed of plagiarism often claim a different perception or qualification of their acts.

Information technology can play a role in identifying plagiarism and fraud. There are a number of computer programs in existence that will check large numbers of documents to do exactly that. Conventional measures as well as effective anti-fraud policies remain necessary as students often do not perceive their actions as illegal or unethical, or use their own "perception" as an excuse.

Keywords: plagiarism, similarity, perception of fraud

1. INTRODUCTION

Educational institutions expect their students to work hard and to produce original work, suitable for assessing the progress they make during the curriculum. Students, of course, want to graduate, but are not always convinced of the necessity to do all the work themselves. Why spend effort in doing something that somebody else has already done before you? Copying strategic parts of existing work can seem much more appealing then. And this is even more the case if the subject is difficult or does not interest you all that much.

Several terms are used to describe this undesirable – at least to the educators – student behaviour. 'Plagiarism' is defined in the Oxford Dictionary of English as "the practise of taking someone else's work or ideas and passing them off as one's own" [1]. It is a broad term, indicating any copying of the work of another author without giving proper credit and /or specifying the source, therefore making it appear as if he who has copied the work is the author himself. Sometimes this happens unintentionally. For example, an author simply forgets to acknowledge one of his sources.

When the term 'fraud' is used, however, the intention to copy without the readers noticing this dominates. 'Fraud' includes any form of plagiarism in situations where this is expressly forbidden, for instance during examinations or when completing an individual assignment. Certain forms of (unintentional) plagiarism can be overcome by providing proper instructions on how students should refer to sources. Identifying and preventing fraud, however, should be a top priority for every educational institution. Legal options for this can for instance be found in the official university regulations regarding education and examinations. It could even be contended that fraud in examinations should be seen as forgery, as the fraudulent actions serve the intention to obtain an official diploma (which can be used as proof of abilities).

2. TYPES OF PLAGIARISM

Several forms of plagiarism can be distinguished [2].

- Plagiarism of ideas, claiming credit for someone else's thoughts, ideas of inventions. An example of this would be if a student writes a thesis, but copies a line of thought or an important insight from a book he or she has read, without mentioning that book. This form of plagiarism is sometimes difficult to avoid, as people are not always consciously aware of the source of an idea. Here, we put emphasis on forms of plagiarism where this consciousness *does* exist.
- Word for word plagiarism, literally copying parts of someone else's work without indicating this. This happens when a student reproduces phrases from someone else's work without using quotation marks. If the original author is not mentioned either, this can also be called plagiarism of ideas.
- Plagiarism of sources, copying citations from another author without mentioning that the citations were brought together by him. A more serious form is when the references are simply copied, while the publications were in fact not read by the new author at all. This form is often found in conjunction with the previous two.
- Plagiarism of authorship, which involves claiming to be the author of a whole work that was in fact written (at least for a substantial part) by someone else. This happens when a student copies (important parts of) a thesis, written by a fellow student (possibly in another university). But it is also possible that a student pays someone else to write the work for him, which falls into the same category.

Of these, word for word plagiarism is usually the easiest to identify. Information technology can be of assistance here, as will be discussed in the next paragraph. Unfortunately, students are increasingly aware of that and in response attempt to mask their copying activities, for instance by substituting synonyms for certain terms or by rearranging the words. They are often surprised when they learn that even the copying of ideas can already be plagiarism.

Plagiarism of authorship has been in the news recently (see for instance [3]) because it is exploited commercially nowadays. On internet sites such as www.ukessays.com, tailor-made essays and dissertations can be ordered by everyone willing to pay for them. These sites officially do not advocate plagiarism (in fact they guarantee the originality of the essays they sell!), they state that the texts can be used as examples, or to improve the contents of the work the student has written himself. But if students feel the same about this remains to be seen. That they would be willing to pay £ 400,- or more for something they only use 'as an example' seems unlikely.

This example clearly illustrates that it is absolutely necessary to monitor student activities these days, such as the production of essays and other assignments. IT provides students with new tools and options, many of which can be used in fraudulent actions. As educational institutions have an important responsibility to make sure that students who graduate indeed possess the required knowledge and skills, they must take action.

This responsibility of educators has a wider significance than just the reliability of the diploma and the reputation of the institution providing it. Unqualified professionals could easily inflict major damage. Therefore, certifying that future engineers, doctors and lawyers are indeed qualified is in the interest of society as a whole and has a clear relationship with public safety and security as well.

3. INFORMATION TECHNOLOGY AS SOURCE OF PROBLEMS

Information technology (IT) plays an important role when dealing with plagiarism and fraud nowadays. To students, it provides the tools to copy and paste large amounts of text or other data almost effortlessly. The results of this are noticeable at all different educational levels. For instance in first-year education, where in a recent take-home assignment at the School of Law, Erasmus University Rotterdam, more than 10% of all students were found to have copied each others work (word for word plagiarism). But even at the graduate level, a number of Master theses were identified as complete copies of someone else's work in the past few years (plagiarism of authorship).

4. USING INFORMATION TECHNOLOGY TO PERCEIVE PLAGIARISM

IT can play a role in the perception of plagiarism and fraud as well, however. Tools exist to identify fraudulent work in several different ways. Three main categories can be distinguished.

- Checking student work 'externally', i.e. comparing it with every other available piece of work.
- Checking student work 'internally', i.e. comparing it with the work of fellow students who did the same assignment or examination.
- Checking the 'consistency' of the written work by a certain student.

4.1. External checking

To check student work externally, a basic requirement is that all documents (both the student work and the external work) are available in electronic form (a computer file). For assignments and 'take home exams' this is usually not a problem nowadays. Most students use word processing software for the production of any substantial piece of text anyway. For most written examinations, taken by groups of students in examination rooms, however, the use of computers to type the answers is still rather uncommon.

External checking typically involves a comparison of the student work with either all documents that are available on the internet (an 'unlimited check') or with a particular subset of these (a 'limited check'). The simplest way to achieve an unlimited check is by using an internet search engine such as Google or Yahoo. An advantage of this approach is that no special software or license is necessary. The precision[1] of a general internet search operation is usually low, however, which means that a considerable number of documents must be opened and inspected manually. Furthermore, a disadvantage is that only generally accessible or 'open' sources can be searched in this way. Copying from commercial databases, to which students often have access because their institutions hold a license, therefore remains unnoticed when using general search engines.

Commercial plagiarism detection services, a number of which have emerged in the past few years, often do not have that drawback. These services tend to operate on a subscription basis; if an institution wants to make use of it, it has to pay an annual fee and sometimes also a fee that is dependant upon the amount of requests. In return, the services offer some advantages, such as:

- the possibility of searching in certain 'closed' data collections, such as commercial databases;
- the possibility to include certain 'private' document collections (such as sets of student assignments) in the search operation;
- the option to operate interactively (for a single file) or to process batches (for a whole set of assignments, to be checked overnight).

Examples of commercial services are SafeAssignment (www.mydropbox.com), Turnitin (www.turnitin.com), Urkund (www.urkund.com) and Ephorus (www.ephorus.nl). Although all of these services are relatively user-friendly, the procedure to check one or more documents varies considerably. Some services depend almost completely on e-mail for file uploading and for the reporting of results,

[1] The precision of a search operation is in this case defined as the ratio of the number of useful documents (documents from which parts were copied) divided by the total number of documents that were retrieved [4].

which can be problematic when larger numbers of files are involved. Other services can only be accessed through a 'Network Learning Environment' (NLE) such as Blackboard (www.blackboard.com). At the Erasmus University, where the authors of this paper do their work as teachers and researchers, this is the case with SafeAssignment. To use this plagiarism detection system, the Blackboard user interface must be used, which still a lot of teachers are not familiar with. As it seems, there is not a single system that is fit for every purpose. A teacher should make his own choices, based on the intended use.

A characteristic these services share with software that performs 'internal' checking is that in fact only *similarities* (see for example [5], [6]) between documents are identified and reported. It is always the teacher who has to decide whether these common phrases (or perhaps paragraphs, or even pages) constitute plagiarism or not. Basically, similarity between documents (available in electronic form) can be calculated completely automatically from for instance the *word use* in the documents. How this can be achieved is explained in the next section, parts of which are based on [6].

4.2. Calculating Similarity

The simplest method to calculate the similarity of two documents utilizes just the presence or absence of word types[2] . This method will be described here. A more sophisticated approach could also take into account the frequency of a word type within each document. With the simple method, only the number of documents in which a word type appears plays a role. This characteristic, the 'document frequency', has a strong relation to the dispersal of word types over the documents.

When we determine the similarity of two documents by means of the word types present in these documents, two situations seem to be possible at first sight:

- a word type is present in both documents; because this means that the documents have a common characteristic, it should increase similarity. For this situation the term 'hit' has been introduced.
- a word type is present in one document, but not in the other; at this point the documents differ from each other and therefore similarity should decrease. This situation is called a 'miss'.

The 'misses' in fact come in two different types. With two documents X and Y there could be

- a 'type 1 miss' (in short, 'miss1') if a word type is present in document X, but not in Y; and there could be
- a 'type 2 miss' (in short, 'miss2') if a word type is present in document Y, but not in X.

With these three characteristics, the number of hits, miss1's and miss2's, the relationship between two documents can be effectively established. However, when the documents do not stand on their own but are part of a database containing many other documents, there is a fourth characteristic. The other documents will probably contain quite a number of word types which are neither present in document X nor in Y. The absence of such a word type in both documents can even be considered a point of *resemblance*, which should increase the similarity of the documents. Therefore, this is also a kind of 'hit', just like in the situation where a word type is present in both documents. This means that next to two types of misses, two types of hits are also possible:

- a 'type 1 hit' (in short, 'hit1') if a word type is present in both documents; and
- a 'type 2 hit' (in short, 'hit2') if a word type is absent in both documents, but is used in other parts of the database (see for example [7]).

The number of documents in which a word type is present (the 'document frequency' of a word type[3]) differs from the number of documents in which other word types appear, this can have an influence on similarity. For the probability that a word type with a high (document) frequency is present in a certain pair of documents, and therefore is responsible for a 'hit1', is much higher than the probability that this happens with a low frequency word. Conversely, the probability of a 'hit2' is higher with word types of a low frequency. For the two types of misses an analogue conclusion can be drawn.

That means that not every hit or miss can be considered to be of equal significance. When a word type with a frequency of only 2 (when the number of documents is high, say 20000) is found in a pair of documents, this gives us much more information than when a high frequency word type (for instance 'the', 'it', etc.) is found, and the similarity of the documents should therefore increase more in the first situation than in the second. This means that we have to take into account the probability that the hit or miss occurs in a certain database. The probability to *encounter* a word type is equal to the (document) frequency of the word type divided by the number of documents in the corpus. The probability to *miss* a word type is equal to the difference between the number of documents and the (document) frequency of the word type, divided by the number of documents in the corpus. The weight (indicating the significance) of a hit or miss of a certain word type is the complement of this probability (1-P(i)). Using these weights, the similarity between two documents could be calculated by adding the weights of the word types that constitute a hit1 or a hit2 in this particular document pair and subtracting from that the weights of the word types that constitute a miss1 or miss2. As not all documents are of equal size, however, these added weights of hits and misses should be made relative to the *maximum* weights that could have been achieved with that particular document, i.e. to the total weight of all words present or absent in it, respectively. A relatively simple similarity score, taking into account just the hits,[4] could then be calculated in the following way:

[2] Word types are the different words used in a certain document, also referred to as the vocabulary in that document. The term word token, on the other hand, is used to indicate one occurrence of a certain word type.

[3] As this 'document frequency' is the only frequency considered here, from now on we will refer to it simply as 'frequency'.

[4] In fact, misses *are* taken into account here, as they influence the relative values; see [6], p. 8.

$$S = \frac{\sum_{i=1}^{m}(1-P(i_{hit1})) + \sum_{i=1}^{n}(1-P(i_{hit2}))}{Hit1_{\max} + Hit2_{\max}}$$

where *m* stands for the number of hit1s, *n* for the number of hit2s, *P* for the probability that a particular word *i* constitutes a hit1 or a hit2, respectively. *Hit1*max is the maximum total hit1 weight for a particular document (= the total weight of all word types present in it) whereas *Hit2*max is the maximum total hit2 weight (= total weight of word types absent in it). For more information on this, see [6].

For a set of documents, this can lead to a series of similarity scores for every possible combination of two documents (i.e. every document pair) from the set. The highest ranking pairs, or pairs of which the score exceeds a certain threshold value, are candidates for closer inspection by the teacher. To support this inspection, some plagiarism detection services provide reports that list the common characteristics or highlight these in the original documents. This can speed up the process of assessing similar documents considerably.

Even with sophisticated report generation, however, assessing the originality of a set of, say, 400 documents can be quite a task. In such a set, it is not uncommon that 20 to 40 document pairs are reported as containing suspicious similarities. This means that the teacher must (re)read and compare 40 to 80 student assignments. If plagiarism is indeed confirmed, follow up must be given to this (for instance in the form of messages to students or to the examination board), which again could take a considerable amount of time and has other drawbacks as well, as will be discussed in section 5 of this paper.

4.3. Internal checking

For the internal checking of documents, IT can again only play a role if all documents are available in electronic form. When this requirement is met, the documents can be compared using a 'plagiarism detection' or 'fraud finder' program installed on a local PC. An important advantage of such a program is that, when a license for it has been obtained, usually no subscription to any commercial service is necessary. The software can be used for an unlimited number of checks, until the licence expires.

Programs that perform internal checking are usually intended to be used on a 'closed' group of documents, for instance all completed student assignments from a particular course or part of a course. Actually, this can be an advantage for the detection of similarities. This is because in a closed set, it is possible to take into account *omissions* of certain words (i.e. a word is not used in the two documents that are compared, but is present in other documents). As is explained in [6], p.4., such a word that is omitted in two documents can be seen as a point of resemblance, which should increase the calculated similarity score. Including this characteristic when calculating a similarity score usually improves results considerably; documents with identical parts get relatively higher scores which makes it easier to distinguish them from the rest. This is especially true if all documents are of more or less equal size (as is often the case with for instance student assignments). When document size differs a lot, however, calculating similarity from just the common (present) words could yield better results. [6], p. 8. Having the possibility to choose one of these (and possibly also other) options while determining similarity and to observe which one works best can be a considerable advantage.

Several software packages are available to perform internal document checking. Examples are WCopyfind (www.copycatchgold.com), Pl@giarism (www.plagiarism.tk) and Codas Fraud Finder (www.andromatics.com). Some of these programs are offered free of charge (sometimes with certain limitations, or for an evaluation period). They usually have a graphical user interface and are easy to operate. Using this kind of software is therefore one of the easiest countermeasures against unauthorised copying within a group of students.

4.4. Consistency checking

One major drawback of both external and internal plagiarism checking is that the original work must be available for comparison. But what if a student copies substantial parts from a book that is unknown to the teacher and has never been published in electronic form? This type of plagiarism is difficult to detect using the techniques described in the previous sections.

There is another option, however. A student who copies substantial amounts of text from external sources (i.e. written by others) mixes his own style of writing, word use etc. with that of other authors. This means that a number of characteristics of the language in the new text will be different from those in other texts produced by the same student. This difference can be detected. Several techniques to accomplish this have been developed in the past decades, for instance in order to find out if a particular text could be attributed to a certain author. Several researchers (for example [8], [9]) have used *word frequency* data from texts that were already known to be written by a certain author to construct a unique 'fingerprint' of that particular author. The same can be done for other texts (for instance, texts of unknown origin). When the key characteristics match, the texts can be attributed to the respective author.

To apply this principle in education, databases containing a broad selection of earlier work of each individual student must be compiled. One way to establish this is to make use of electronic 'portfolios', in which every piece of written work of a particular student is stored from the moment he commences his studies until the moment he leaves the institution. This material is usually well suited to construct a 'fingerprint' from, which can then be compared to that of any new production. The more documents the portfolio contains, the more reliable the fingerprint will be. Of course, to most teachers this is not an entirely new technique. When they know their students well, they tend to 'feel' that something is wrong if a mediocre student hands in productions that considerably exceed his usual level. Using a computer to compare linguistic fingerprints, however, makes it possible to apply the method with more precision and in educational situations where the number of students is too high to know the individual work of each of them.

5. CONVENTIONAL MEASURES AGAINST PLAGIARISM

Even though technology brings us powerful new tools to perceive plagiarism, conventional measures against this undesirable phenomenon are still just as important. This is especially true in education, as discussed in this paper. It is vital that students are taught the importance of producing their own, original work and the need to handle sources correctly. As stated earlier, certain forms of plagiarism are far from obvious to many students. Therefore, they could apply these forms unintentionally, unless taught otherwise.

Therefore, plagiarism should be a subject that is dealt with explicitly, starting from the first year of education. The different forms of plagiarism and the ways to avoid them should be explained. Furthermore, students should be taught the guidelines for the proper acknowledgement of sources, using a generally accepted method. Using plagiarism detection software can then be a logical complement to this education. It can be used to check that all students have understood and implemented the guidelines correctly. If done in this form, students will probably accept it more easily, as they will recognize that everyone is treated equally in this respect and that only original work is accepted and rewarded.

6. INTEGRITY, LEADERSHIP AND ACADEMIC CULTURE

Another important reason to take plagiarism seriously and to counteract it is the fact that the institutional reputation depends upon its perceived integrity. If standards for student work are maintained, this reputation will be reinforced. Students who have become used to this will probably see this advantage as well, since this will be reflected in their own status as graduates of a highly ranked university. It will simply be "not done" to copy work.

This approach, which puts emphasis on alerting and prevention, is preferable to one that mainly focuses on repression, not only from the educational point of view, but for practical reasons as well. Being able to detect that students have cheated is one thing, being able to prove it beyond doubt and documenting it in such a way that it will hold up in appeal procedures – and even in court, if a student wants to take it that far – is much more difficult. Appeal procedures, especially external ones, can be costly and time consuming for both parties and could damage the institution's reputation. Furthermore, proving that a student has copied work of others can be difficult, even in seemingly clear cases, as there is always the possibility that the similarities are caused by discussions between students that are, in themselves, allowed. For a student, a lot can be at stake when he or she is accused of fraud, which makes it attractive not to acknowledge the copying, but to deny everything.

In former times, the university was the province of a privileged group. An academic education was less a matter of a necessary prerequisite for a well-paid, prestigious job and more a matter of personal development. Nowadays, the student population is more diverse. There are often large numbers of students studying any given course. The anonymity of students on these courses means that the social control to prevent 'free riding', mediocre results or plagiarism and fraud has decreased. Fraud is becoming a huge problem. Many students find it more important to support each other, *not* monitoring each other's behaviour, than to strive for systematic acquisition of knowledge for themselves and for others.

The lack of integrity of some of those who will be leaders in society has become a worldwide problem. Its dangers can be seen for instance in fraud in large companies such as Enron, the housing crisis, environmental pollution and the scarcity of energy and food. These leaders, both in the public and private sectors, have largely been educated at universities. Therefore these problems must be addressed here (although not only here). Companies, governments and other organizations will function better when academic students receive training in leadership and integrity.

7. CONCLUSION

Plagiarism is a serious problem in modern education. Information technology makes it easier for students to find and copy work of others. Many students are so used to looking up information on the internet, that they do not even realise that what they find should not just be copied, but should be properly acknowledged as work done by someone else. In High School many of these students were even praised for handing in extended project reports containing lots of copied material. For these students, making them aware of the problem and teaching them the right methods might suffice. But those who deliberately copy the work of others and present it as their own should definitively be identified and "helped" to end this behaviour.

Fortunately, IT can play a role to counteract against plagiarism as well. Several tools exist that compare student work with external sources (documents available on the internet) as well as with internal sources (essays or assignments from other students in the same group). The effectiveness of these tools differs, as it depends, among other things, on the number and the quality of the sources that are used for comparison. Internal checking can be very effective in a closed group, for instance to check the originality of take-home assignments.

Teaching students not to plagiarise and educating them to use sources properly should of course lay the foundation. But to make sure that they have understood the lesson and really do not cheat, technical means are indispensable. These make it possible to perceive (= identify) what is happening, and only then the students' 'perception' (=qualification) of their own actions can be changed. These tools should be embedded in institutional policies against plagiarism and fraud. Doing so is vital to the reputation of the institutions and to that of students.

As it now appears that academic integrity is no longer a given in the student population, it is time that universities specifically address the issues of leadership and integrity. This is not only to preserve the reputation of the academic institutes themselves, but also to promote the leadership qualities of their alumni.

REFERENCES

[1] The Oxford Dictionary of English, second edition, Oxford University Press 2003.

[2] Martin, B., Plagiarism: policy against cheating of policy for learning?, University of Wollongong 2004, p. 2-3.

[3] Taylor, M. and Butt, R., 'How do you make £1.6m a year and drive a Ferrari? Sell essays for £ 400', The Guardian, July 29 2006.

[4] Salton, G., *Automatic Text Processing, the Transformation, Analysis and Retrieval of Information by Computer*, Massachusetts: Addison-Wesley 1989, p. 248-249.

[5] Meadow, Ch. T., Boyce, B.R & Kraft, D.H, *Text Information Retrieval Systems*, San Diego (Ca): Academic Press 2000, p. 221-224

[6] Noortwijk, C. van & Mulder, R.V. De, 'The Similarities of Text Documents', in: *JILT – Journal of Information, Law and Technology*, Issue 2/1997, Coventry: University of Warwick 1997.

[7] Batagelj, V. & Bren, M. *Comparing Similarity Measures*. University of Ljubljana, Ljubljana 1993.

[8] Ellegard, A., *A Statistical Method for Determining Authorship: The Junius Letters 1769-1772*, Gothenburg Studies of English no. 13, Gothenburg: University Press 1962

[9] Kenny, A., 'A stylometric study of Aristotle's Ethics', in: Lusignan, S. and North, J.S. (eds.), *Computing in the Humanities*, Waterloo, Ontario: University Press 1977.