# TOWARDS EFFICIENT PLAGIARISM DETECTION

Shin M.V.
Krasnova T.I. (research adviser)
National Research Tomsk Polytechnic University
marishapicke@gmail.com

## Introduction

Nowadays plagiarism has become a global problem in academic background. Students use different methods and approaches to create plagiarized assignments which can be very disappointing and even demotivating for their lecturers. Internet development gave a striking rise of plagiarism as it expanded the possibilities of finding information and quite often students can't resist the temptation of citing it without referencing the author. Simple reminding that plagiarism is a way of cheating and violation of rules and ethical principles doesn't work. The only solution of overcoming this inadmissible practice is plagiarism detection. Special software is created to make barriers for this type of academic dishonesty. The main objective of this software is assisting people in the task of detecting plagiarism **(**Barron-Cedeno et al., 2013). A growing number of tools for automated plagiarism detection are now in use at universities around the world. In Russia the most popular system counteracting this phenomenon is called Antiplagiat. Such systems have a number of drawbacks as well as advantages. Pecorari (2010) believes that usually the problems are associated with the following:

- plagiarism detection software can only identify electronic sources but not printed ones;
- password-protected databases can be excluded;
- this software doesn't compare the submitted document with the full text of the stored data, as it usually makes a 'digital fingerprint' for each document to be compared therefore some copying from sources may escape detection.

Plagiarism detection is based on different checking approaches and procedures. In this article we seek for efficient plagiarism prevention measures by analyzing the operation principles of automated plagiarism detection systems.

## Plagiarism Methods

The most wide-spread plagiarism methods are full-borrowed plagiarism (which is known as copy & paste plagiarism), paraphrase, translation and idea plagiarism. In the *Figure 1* these plagiarism methods appear according to the difficulty of their detection (from left to right). However nowadays plagiarism can be easily detected with the Internet and network search systems. This procedure is pretty fast and not costly. Today people have a lot of special search systems that are made for plagiarism detection. Those are called "antiplagiarism systems".

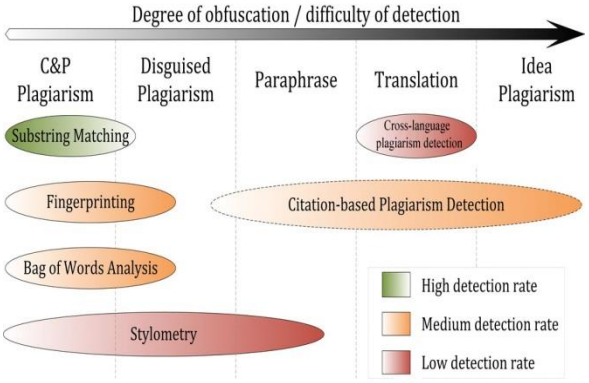


Figure 1. Detection complexity

Launched antiplagiarism service collects data from possible sources, after that it uses special methods for plagiarism detection. This service checks formatted document and then includes it to its own existing database with the texts that are already checked by source. Services check articles one by one and a user is informed what place in line his or her document has. At the end of document check process the system informs the user about originality (percentage), underlines borrowed fragments (phrases or whole texts). Some sources can give the user links that contain the same phrases as in his or her document.

## Plagiarism Detection Principles of Operation

All systems that are used for plagiarism detection can be divided into three big classes:

1) *Internet-search systems*. Using this system people can search borrowed texts manually. In addition to this, Google search (the project Google Scholar) identifies some science works and citations in them. 2) *Metasearch systems and antiplagiarism systems* that do not have their own document database. These systems work by forming the requests to the popular search machines on the Internet and then show the results.

3) *Special antiplagiarism systems* with their own search algorithms of matching and document database. The way this systems works:

1. it converts unchecked document into a text;
2. it indexes this text. This operation may include:
   - simple text fragment extraction and its reduction (e.g. elimination of short words and words which do not exist in the vocabulary) and then bringing words to their basic form.
   - text indexing;
3. while searching it finds matched fragments and sorts the results.

*Plagiarism Detection Principles of Operation: Morphology.* In search systems the words are usually used not in their standard forms but converted into basic forms.

*Plagiarism Detection Principles of Operation: Fingerprinting.* The main purpose of this search is to find inaccurately matched words or texts fragments. Mostly for this search N-gram method is used (N is a number of consecutive symbols from text in some fragment) or its variation. The comparison can be made, for instance, by a number of matched bigrams.

*Plagiarism Detection Principles of Operation: strings and patterns*. A text matching word-by-word search is extremely resource-intensive operation so it can be simplified by searching not for words but for specified fragments (e.g. sentence searching). Its efficiency is very poor when sentences are divided into several parts or combined together. That is why a sequence of words extracted from the text is usually used. When it is used a sequential extraction these divided parts are called **strings** (special feature is $L$ – it is a length of a string given in words). When it is used an inconsequential extraction (e.g. a search is done with another principle) these extracted parts are called **patterns**.

For example, we have such phrase: “by the way, oranges, apples and pears are fruits”. The strings (for $L = 2$) for this phrase are as following: “by the”, “the way”, “way oranges”, “oranges apples”, “apples and”, etc. These words within each string can be sorted for their own normalization (e.g. sorting according to the alphabet). Patterns for the same phrase are as following (according to the principle of separation punctuation): “by the way”, “oranges”, “apples and pears are fruits”.

On the one hand, pattern extraction method is more preferable as string extraction method because patterns have a bigger number of words than strings have, that is why the amount of patterns is less than the amount of stings. This increases the process speed significantly. On the other hand, patterns can be subjected to changes more than strings. The main problems of both strings and patterns are: speech tokens, proper names, etc.

## The ways of Deceiving Plagiarism Detection Systems

The ways of deceiving plagiarism detection systems can be divided into two main approaches: technical and nontechnical.

*Technical* methods include:

- letters change (one letter is changed to the letter from another alphabet that has the similar way of writing, e.g. changing English “a” to Russian “a”);
- single letters, dots, spaces (or other symbols repainting to background) color;
- invisible text insertion;
- orthographical mistakes addition;
- Synonymizer usage (Synonymizer is a programme for automatic or semi-automatic words replacement with their synonyms);
- Antiplagiarism systems vulnerability usage (the possibility to make the required originality percentage).

There are software products, such as AntiPlagiatKiller v2, which analyze text and show text edition recommendation (e.g. remove old word and add a new one, “something must be changed”, etc.). The advantages of technical methods when deceiving plagiarism detection systems are: large-scale usage, availability and high operating speed.

*Nontechnical* methods consist of text paraphrasing. Nowadays simple text transformation, such as sentence splitting or joining, words inversion, words replacement to their synonyms, explanation of abbreviations or some fragments rewriting, does not have a significant impact on the detection process. Speaking about paraphrasing, one should mention Search Engine Optimization. This is a comprehensive set of form edition and content (text, website) measures with the aim of increasing its position in search results. In addition to this, it contains methods that make text unique and fill it with keywords. Rewriting is a method of changing text narration form and saving its original meaning. Copywriting is professional text writing, mostly advertising.

## Conclusion

The increasing availability of Internet sources caused an increase in plagiarism and it made academic dishonesty much easier and faster. The concern of academic community over the scope of plagiarism in higher education is very high. IT technologies help in plagiarism detection and different systems are actively used by universities. But due to some vulnerabilities and imperfections of these systems there are still ways of deceiving them. Therefore academic community is still expecting further improvements in plagiarism detection systems.

## References

1. Barrón-Cedeño, A., Vila, M., Martí, A., Rosso, P. (2013). Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics* 39 (4), 917-947.
2. Pecorari, D. (2010). Academic Writing and Plagiarism: a Linguistic analysis. London. Continuum International Publishing Group.
3. Plagiarism Detection. Retrieved 18 October, 2014 from http://en.wikipedia.org/wiki/Plagiarism_detection
4. Williams, H. (2008). Plagiarism. Farmington Hills: Greenhaven Press.