Computer-Based Plagiarism Detection Methods and Tools: An Overview

Romans Lukashenko, Vita Graudina, Janis Grundspenkis

Abstract: *The paper is dedicated to plagiarism problem. The ways how to reduce plagiarism: both: plagiarism prevention and plagiarism detection are discussed. Widely used plagiarism detection methods are described. The most known plagiarism detection tools are analysed.*

Key words: *Plagiarism, plagiarism prevention, plagiarism detection, similarity measures*

INTRODUCTION

Lancaster and Culwin in [13] state that "plagiarism as „theft of intellectual property" has been around as long as human has produced work of art and research". Plagiarism can be defined as turning of someone else's work as your own without reference to original source. Commonly in practice there are different plagiarism methods. Some of them include [13]:

- copy – paste plagiarism (copying word to word textual information);
- paraphrasing (restating same content in different words);
- translated plagiarism (content translation and use without reference to original work);
- artistic plagiarism (presenting same work using different media: text, images etc.);
- idea plagiarism (using similar ideas which are not common knowledge);
- code plagiarism (using program codes without permission or reference);
- no proper use of quotation marks (failing to identify exact parts of borrowed content);
- misinformation of references (adding reference to incorrect or non existing source).

It is hard to disagree that plagiarism's problem becoming more and more actual – society's knowledge about plagiarism is constantly increasing and plagiarism's problem seriously starts to attract society's attention. More and more people begin to realize that plagiarism is amoral phenomena that can't exist in society with high ethical standards.

But why does plagiarism's problem become particularly actual just nowadays? We are living in the age of information technologies that despite of making our life easier also creates a set of problems. Availability of digital documents (for instance, easy access to the Web) and telecommunications in general open good chances for plagiarism prosperity turning cheating into extremely easy and engaging process. In [13] it is stated that "nowadays plagiarism has turned into a serious problem for publishers, researchers and educators".

The remainder of this paper is organized as follows. The next section gives some ideas about plagiarism reduction. Then different methods for plagiarism detection are described. After that analysis of already developed tools are presented. Finally, some conclusions are given.

WAYS HOW TO REDUCE PLAGIARISM

Nowadays many methods to fight against plagiarism are developed and used. These methods can be divided into two classes: (1) methods for plagiarism prevention, and (2) methods for plagiarism detection.

If we consider plagiarism as a kind of social illness then we can say that methods of the first class are precautionary measures which aim are to preclude rise of illness, but methods of the second class are cures which are aimed to avert existing illness.

Some examples of methods in each class are as follows: plagiarism prevention – honesty policies and/or punishment systems, and plagiarism detection – software tools to reveal plagiarism automatically.

Each method has a set of attributes that determine its application. Two main attributes which are common to all methods are (see, Table 1):

1) work – intensity of method's implementation;
2) duration of method's efficiency.

Work – intensity of method's implementation means amount of resources (mainly time) which is needed to develop this method and bring into usage. Plagiarism prevention methods are usually time – consuming in their realization, while plagiarism detection methods require less time.

Duration of method's efficiency means the period of time in which positive effect of method's realization exists. Implementation of prevention methods gives a long-term positive effect. In contrast, implementation of detection methods gives short – term positive effect. Methods have different duration of positive effect, because of antipodal approaches which methods use to fight against plagiarism – detection methods based on society's intimidation, while prevention methods more rely on the society's change of attitude against plagiarism.

Table 1: Attributes of plagiarism detection and prevention methods

Method	**Attributes of method**	
	Implementation work – intensity	Duration of positive effect
Plagiarism prevention methods	Require more time to implement	Positive effect isn't momentary, but it is long – term
Plagiarism detection methods	Require less time to implement	Positive effect is momentary, but it is short – term

Despite of differences in prevention and detection methods all these methods are used to achieve one common goal – to fight against plagiarism. To make this fight efficient, system approach to plagiarism problem solving is needed, i.e. it is needed to combine plagiarism prevention and detection methods. To achieve momentary, short – term positive results plagiarism detection methods must be applied at problem's initial stages, but to achieve positive results in long – time period plagiarism prevention methods must be put into action. Plagiarism detection methods can only minimize plagiarism, but plagiarism prevention methods can fully eliminate plagiarism phenomena or at least to a great extent decrease it. That is why plagiarism prevention methods without doubt are more significant measures to fight against plagiarism. Unfortunately, plagiarism prevention is a problem for society as a whole, i.e., it is at least national wide problem which can not be solved by efforts of one university or its department. That is why only plagiarism detection methods and tool are discussed in this paper.

PLAGIARISM DETECTION METHODS

Plagiarism detection usually is based on comparison of two or more documents. In order to compare two or more documents and to reason about degree of similarity between them, it is needed to assign numeric value, so called, similarity score to each document. This score can be based on different metrics. There are many parameters and aspects in the document which can be used as metrics. In this paper we don't pay attention to specific metrics used for plagiarism detection in the source code, like Halstead metrics [6]. The most widely used general purpose metrics are described in this section.

Lancaster and Culwin in their work [12] have tried to classify metrics used for plagiarism detection. They have proposed two ways how to classify metrics. First classification is based on the number of documents involved in the metrics calculation process and second one is based on computational complexity of the methods employed to find similarities. In the first classification metrics can be classified as singular or paired metrics and as corpal or multi-dimensional metrics, depending on how many documents

are proceeded, and depending on the set of documents involved in proceeding, respectively. A corpal metric operates on an entire corpus of documents. A multi-dimensional metric operates on a chosen number of documents. In the second classification metrics can be classified as superficial metrics and structural metrics. A superficial metric is a measure of similarity that can be gauged simply by looking at one or more documents. In this case knowledge of the linguistic features of natural language is not necessary. A structural metric is a measure of similarity that requires knowledge of the structure of one or more documents.

Another way how to classify metrics is according to main principle build-in them, i.e., documents' contents analysis is based on semantical or statistical methods. In statistical methods there are no needs to understand the meaning of the document. A common statistical approach is the construction of document vectors based on values describing the document, like, the frequencies of words, compression metrics [16], Lancaster word pairs [11] and other metrics. Statistical metrics can be language-independent or language-sensitive [5]. Purely statistical method is N-gram approach where text is characterized with sequences of N consecutive characters [3]. Based on statistical measures each document can be described with so called fingerprints, where n-grams are hashed and then selected some to be fingerprints [17; 18]. There can be also measures which contain probabilities. These measures are information theoretical measure [1], BM25 [15], and language model measure [22].

In many cases similarity score between two documents is calculated as Euclidean distance between document vectors. The similarity of identical documents is zero [9]. Similarity also can be calculated as scalar product of document vectors divided by their lengths. This is equivalent to the cosine of the angle between two document vectors seen from the origin [20]. In many cases document vectors are composed from word frequency and word weight which are automatically calculated for each document. Word frequency is taken into account in proportion function [2]. Also cosine formula can have variation (see, Eq. 1), where also word weights are taken into account [2]:

$$S_{\cos}(A,B) = \frac{\sum_{i=1}^{n}[\alpha_i^2 \times F_i(A) \times F_i(B)]}{\sqrt{\sum_{i=1}^{n}[\alpha_i^2 \times F_i^2(A)] \times \sum_{i=1}^{n}[\alpha_i^2 \times F_i^2(B)]}} \quad (1)$$

where α_i – word weight vector; $F_i(A)$, $F_i(B)$ – frequency of the ith word in documents A and B, respectively.

Cosine function, proportion function, as well as dot production, Jaccard measure, Dice measure, overlap measure [21] are symmetric similarity measures [2]. Symmetric or asymmetric similarity measures are one more classification. Asymmetric similarity measures are heavy frequency vector and heavy inclusion proportion model, which are derived from cosine function and proportion function by combining asymmetric similarity concept with heavy frequency vector [2]. Asymmetric similarity measures can be used for searching subset coping.

Usually in different tools statistical methods are implemented due to their simplicity.

TOOLS FOR DETECTING PLAGIARISM

In [13] authors conclude that "looking at the extend of the problem, it is quite obvious that academia requires tools to automate and enhance plagiarism detection". In accordance with [12] "plagiarism detection tools are programs that compare document with possible sources in order to identify similarity and so discover submissions that might be plagiarized".

There is a number of tools available to detect plagiarism in documents. The most known plagiarism detection tools are Turnitin, Eve2, CopyCatchGold, WordCheck, Glatt,

Moss, JPlag [4; 10; 12; 13; 14; 19]. According to analytical information available on the Web leader between detection tools is Turnitin [7; 8], due to it's functionality. Each tool has a set of attributes that determine its application. Two main attributes which are common to all tools are:

1) type of text tool operates on;
2) type of corpus tool operates on.

According to attribute "type of text tool operates on" tools can be divided into two groups: tools that operate on non-structured (free) text and tools that operate on structured (source code) text. In fact, detection tools are not limited to operate on free text or source code. It may be used to find similarity in spreadsheets, diagrams, scientific experiments, music or any other non-textual corpora [12].

According to attribute "type of corpus tool operates on" tools can be divided in three groups: tools that operate only intra-corpally (where the source and copy documents are both within a corpus), tools that operate only extra-corpally (where the copy is inside the corpus and the source outside) and tools that operate both – intra- and extra-corpally [12].

Table 2 shows detail attributes of plagiarism detection tools. All tools in table are grouped into specific tools, which were specially developed to detect plagiarism in submissions, and Internet search engines – alternative tools to detect suspected plagiarism [19]. It is worth to point out that alternative tools haven't appropriate set of instruments to analyze suspected submissions qualitatively that is why these tools can't be viewed as serious plagiarism detection tools.

Table 2: Attributes of plagiarism detection tools [based on 4]

Attributes	**Detection tools**							
	Specific tools							***Alternative tools***
	Turnitin	Eve2	CopyCathGold	WordCheck	Glatt	Moss	Jplag	Google Yahoo AltaVista
Type of text tool operates on								
Checks source code?	-	-	-	-	-	**Y**	**Y**	-
Checks free text?	**Y**	**Y**	**Y**	**Y**	**Y**	-	-	**Y**
Type of corpus tool operates on								
Operates intra-corpally?	**Y**	-	**Y**	**Y**	-	**Y**	**Y**	-
Operates extra-corpally?	**Y**	**Y**	-	-	-	-	-	**Y**
Other attributes								
Designed for use by students?	**Y**	-	-	-	-	-	-	**Y**
Designed for use by teachers?	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
Instant response?	-	**Y**	-	**Y**	-	-	-	**Y**
Free?	-	-	-	-	-	**Y**	**Y**	**Y**

Operation of plagiarism detection tools is based on statistical or semantical methods or both to get better results. Information about methods and algorithms which are applied in each particular tool is a kind of business secret that is not manifested. From available descriptions of some detection tools it may be concluded that the great part of tools uses statistical methods to detect plagiarism, because these methods are well – understood and they are easier to implement in software.

In [19] it is stressed that "although plagiarism detection tools provide excellent service in detecting matching text between documents, care needs to be taken in their use". Plagiarism detection tools inability to distinguish correctly cited text from plagiarised text is one of the serious drawbacks of these tools [4; 19]. That is why human interposition is necessary before a paper is declare plagiarised – manual checking and human judgment are still needed [4].

CONCLUSIONS

In the age of information technologies plagiarism has become more actual and turned into a serious problem. In the paper ways how to reduce plagiarism are discussed.

Plagiarism prevention methods which are based on society's change of attitude against plagiarism without any doubt are the most significant means to fight against plagiarism, but implementation of these methods is a challenge for society as a whole. Education institutions need to focus on plagiarism detection methods.

Analysis of widely used plagiarism detection methods shows that usually different statistical metrics are used due to their simplicity and easiness to be implemented in tools.

Analysis of the known plagiarism detection tools shows that although these tools provide excellent service in detecting matching text between documents, even advanced plagiarism detection software can't detect plagiarism so good as human does. They have several drawbacks and, so manual checking and human judgment is still needed. Human brain is universal plagiarism detection tool, which is able to analyze document using statistical and semantical methods, is able to operate with textual and non-textual information. At the present such abilities are not available for plagiarism detection software tools. In accordance with [19] "…at least for now – nothing can completely replace the watchful eye of human beings". But nevertheless computer – based plagiarism detection tools can considerably help to find plagiarised documents.

ACKNOWLEDGEMENT

This work has been partly supported by the European Social Fund within the National Programme "Support for the carrying out doctoral study program's and post-doctoral researches" project "Support for the development of doctoral studies at Riga Technical University".

The main results are outcomes of the research project ZP-2006/06 "Development of the intelligent system's prototype for plagiarism detection in students' works".

REFERENCES

[1] Aslam, J.A., M. Frost. An information-theoretic measure for document similarity. Proceedings of the 26th international ACM/SIGIR conference on research and development in information retrieval, pp. 449–450, 2003.

[2] Bao, J.P., J.Y. Shen, H.Y. Liu, X.D. Liu. A fast document copy detection model. Soft Computing - A Fusion of Foundations, Methodologies and Applications. Vol. 10 (1), pp. 41 – 46, 2006.

[3] Brin, S., J. Davis, H. Garcia Molina. Copy detection mechanisms for digital documents. Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, pp. 398–409,1995.

[4] Delvin, M. Plagiarism detection software: how effective is it? Assessing Learning in Australian Universities, 2002. Available at: http://www.cshe.unimelb.edu.au/ assessinglearning/docs/PlagSoftware.pdf

[5] Gruner, S., S. Naven. Tool support for plagiarism detection in text documents. Proceedings of the 2005 ACM Symposium on Applied Computing. pp. 776 – 781, 2005.

[6] Halstead, M. Elements of Software Science, Elsevier Publishers, New York. 1977.

[7] iParadigms. Plagiarism Prevention: Stop plagiarism now… with Turnitin®. Product datasheet, 2006. Available online at: http://turnitin.com/static/pdf/datasheet_plagiarism.pdf

[8] iParadigms, LLC. Turnitin. Plagiarism prevention engine. Available online at: http://www.turnitin.com

[9] Jones, E.W. Plagiarism monitoring and detection – towards and open discussion. In Proceedings of the twelfth annual CCSC South Central conference on The Journal of Computing in Small Colleges, pp. 229–236, 2001.

[10] Lancaster T., F. Culwin. A review of electronic services for plagiarism detection in student submissions. Paper presented at 8th Annual Conference on the Teaching of Computing, Edinburgh, 2000. Available at: http://www.ics.heacademy.ac.uk/events/presentations/317_Culwin.pdf

[11] Lancaster T., F. Culwin. A visual argument for plagiarism detection using word pairs. Paper presented at Plagiarism: Prevention, Practice and Policy Conference 2004.

[12] Lancaster, T., F. Culwin. Classifications of Plagiarism Detection Engines. ITALICS Vol. 4 (2), 2005.

[13] Maurer, H., F. Kappe, B. Zaka. Plagiarism – A Survey. Journal of Universal Computer Sciences, vol. 12, no. 8, pp. 1050 – 1084, 2006.

[14] Neill, C.J., G. Shanmuganthan. A Web – enabled plagiarism detection tool. IT Professional, vol. 6, issue 5, pp. 19 – 23, 2004.

[15] Robertson, S., S. Walker. Some simple effective approximations to the 2-poissonmodel for probabilistic weighted retrieval. Proceedings of the 17th international ACM/SIGIR conference on research and development in information retrieval, pp. 232–241, 1994.

[16] Saxon, S. Comparison of plagiarism detection techniques applied to student code: Computer science project (Pt. II). Cambridge: Trinity College, 2000.

[17] Stein, B., S.M. zu Eissen. Near Similarity Search and Plagiarism Analysis. Proceeding of 29th Annual Conference of the German Classification Society, pp. 430-437, 2006.

[18] Schleimer, S., D.S. Wilkerson, A. Aiken. Winnowing: Local Algorithm for Document Fingerprinting. Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 76-85, 2003.

[19] The University of Sydney Teaching and Learning Committee. Plagiarism detection software report. Draft One, 2003.

[20] Tan, C.L, W. Huang, S.Y. Sung, Z. Yu, Y. Xu. Text Retrieval from Document Images Based on Word Shape Analysis. Applied Intelligence 18, pp. 257–270, 2003

[21] Wan, X. Beyond topical similarity: a structural similarity measure for retrieving highly similar documents. Knowledge and Information Systems, 2006.

[22] Zhai, C., J. Lafferty. A study of smoothing methods for language models applied to ad-hoc information retrieval. Proceedings of the 24th annual international ACM/SIGIR conference on research and development in information retrieval, New Orleans, Louisiana, United States, pp. 334–342, 2001.

ABOUT THE AUTHORS

Assistant, Romans Lukashenko, M.sc.ing., Faculty of Computer Science and Information Technology, Riga Technical University. Phone: +371 (6)7 089 529, e-mail: lrexpress@inbox.lv

Assistant, Vita Graudina, M.sc.ing., Faculty of Computer Science and Information Technology, Riga Technical University. Phone: +371 (6)7 089 095, e-mail: Vita.Graudina@cs.rtu.lv

Professor, Janis Grundspenkis, Dr.habil.sc.ing., Faculty of Computer Science and Information Technology, Riga Technical University. Phone: +371 (6)7 089 581, e-mail: Janis.Grundspenkis@cs.rtu.lv