A Review on Plagiarism Detection Tools

Ramesh R. Naik
Ph.D. Student
Dept. of CS & IT
Dr. B.A.M.U.,Aurangabad

Maheshkumar B. Landge
Ph.D. Student
Dept. of CS & IT
Dr. B.A.M.U.,Aurangabad

C. Namrata Mahender
Assistant Professor
Dept. of CS & IT
Dr. B.A.M.U.,Aurangabad

ABSTRACT
Plagiarism has become an increasingly serious problem in the academic world. It is aggravated by the easy access to and the ease of cutting and pasting from a wide range of materials available on the internet. It constitutes academic theft - the offender has 'stolen' the work of others and presented the stolen work as if it were his or her own. It goes to the integrity and honesty of a person. It stifles creativity and originality, and defeats the purpose of education The plagiarism is a widespread and growing problem in the academic process. The traditional manual detection of plagiarism by human is difficult, not accurate, and time consuming process as it is difficult for any person to verify with the existing data. The main purpose of this paper is to present existing tools about in regards with plagiarism detection. Plagiarism detection tools are useful to the academic community to detect plagiarism of others and avoid such unlawful activity. This paper describes some of the plagiarism detection tools available for plagiarism checking and types of plagiarism.

Keywords
Plagiarism detection, types of plagiarism, plagiarism tools, plagiarism detection methods.

1. INTRODUCTION
Now a day's theft of information as widely increased in the form of computer data. This also comes in the academic or education era this parts known as plagiarism which is specifically defined as a form of research misconduct, "Misconduct means construction, distortion, copy or any other practice that seriously deviates from practices commonly accepted in the discipline or in the educational and research communities generally in proposing, performing, reviewing, or reporting research and inventive activities".

Plagiarism is the act of stealing someone else's work and attempting to "pass it off" as your own. This can apply to all the terms like papers, photographs, songs, even ideas, thoughts etc...

According to the Merriam-Webster Online Dictionary, to "plagiarize" means to steal and pass off the ideas or words of another person as created by own self.

- To use (another's creation) without crediting the source.
- To commit literary theft.
- From an existing source deriving an idea or product and present it as new [1].

Types of Plagiarism

There are different types of plagiarism shows in below figure1.

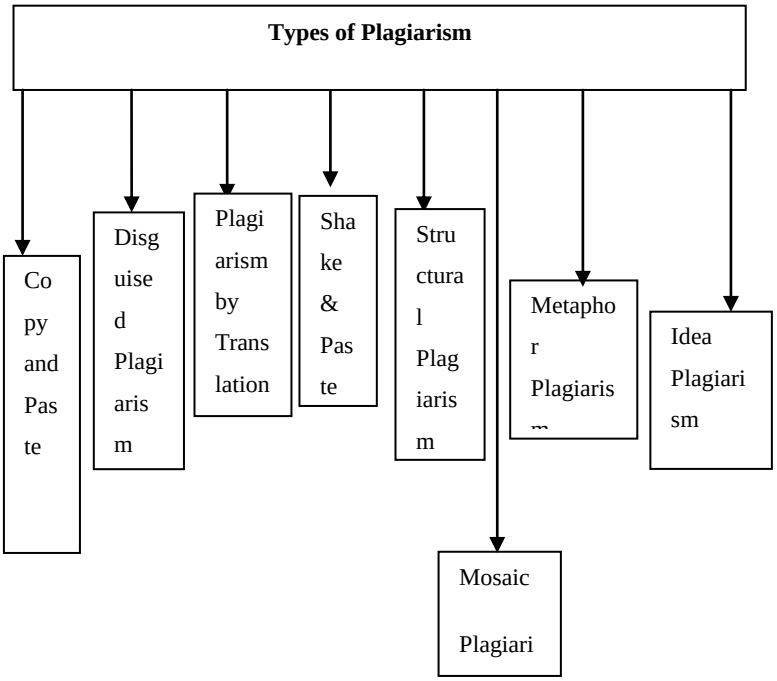

Figure 1: Types of plagiarism

1.1 Copy & Paste
This is more or less the only kind of plagiarism that is quickly recognizable and generally granted on to be plagiarism. The plagiarist finds a useful source and copies a portion of that, perhaps with a few minor changes, into the text that is to be changing the name of the author [2].

1.2 Disguised Plagiarism
Disguised plagiarism when text from a source is copied and then some effort is exerted in order to disguise the copy. Words may be removed or added, word order is changed, or even an attempt at paraphrase may be undertaken. However, source is not given, or only given for a part of the text taken, this is still considered to be plagiarism [3].

1.3 Plagiarism by Translation
When a text is taken from one language and translated, either manually or with the help of an automatic translation system, and used without the source being named, then we speak of plagiarism by translation [2].

1.4 Shake & Paste
Among students a variation of copy & paste can often be seen whereby paragraphs are taken from a number of different sources and collected, often without a functional order. Each paragraph will be well written in and of itself, but there is no clear change from one paragraph to the next. When this is done on the level of snippets, that is parts of sentences

"joined" together, we sometimes speak of mosaic plagiarism [2].

1.5 Structural Plagiarism

Taking the idea of any person, their sequence of arguments, their selection of quotations from other people, or even the footnotes that they use in the same order without giving credit is considered to be structural plagiarism. This type of plagiarism is fairly difficult to control, as one must read both texts very closely to see what has been taken [3].

1.6 Mosaic plagiarism

Patchwork paraphrasing refers to obtaining content from a various sources catering to the same topic of interest and rephrasing the sentences, switching words, using synonyms and improvising on the grammar styles to finally producing one's own research paper without citing the sources [2][3].

1.7 Metaphor plagiarism

"Metaphors are used either to make an idea clearer or give the reader an analogy that touches the senses or emotions better than a plain description of the object or process. Metaphors, then, are an important part of an author's creative style" [4][5].

1.8 Idea plagiarism

If one copies an innovative idea or a solution provided by another author in a source document, whilst one cannot provide a solution or an idea of his own, the idea plagiarism is said to have occurred. The research paper authors have a hard time distinguishing the ideas and/or solutions provided by the author of the source paper from public domain information. Public domain information is any idea or solution about which people in the field accept as general knowledge [6].

1.9 Self-plagiarism

Here the author of the research paper reuses his own previous work to produce a new work [7].

2. PLAGIARISM DETECTION METHODS

There are two main plagiarism detection methods and its general techniques which are classified as shown below figure2:

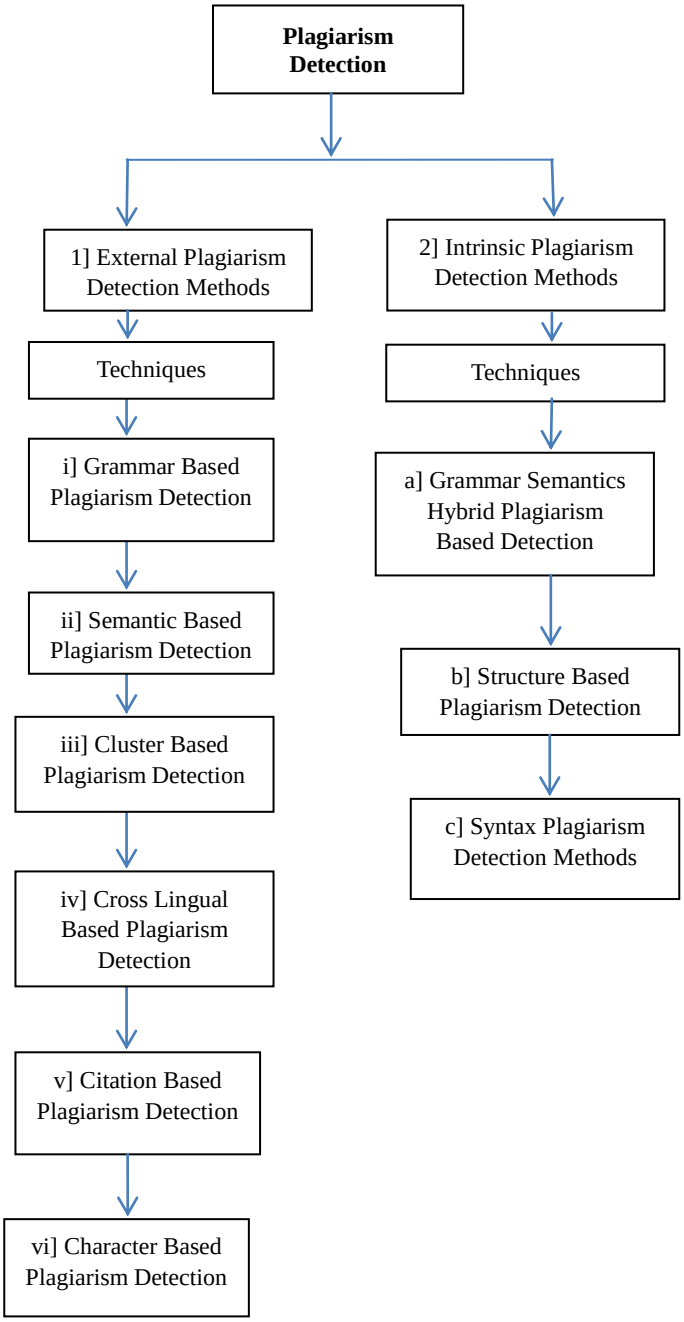

Figure 2: Classification of Plagiarism Detection Methods

2.1 External Plagiarism Detection Methods

Plagiarism is detected by comparing the contents of the submitted research paper with the contents of the already published and publicly available in various databases. It requires a reference corpus.

There are six general techniques in the External Plagiarism Detection methods which are as follows:

1. Grammar Based Plagiarism Detection

This technique uses a string-based matching approach to detect and to measure similarity between the documents available within a database under consideration. The grammar-based technique is suitable for detecting clone documents and fails to detect plagiarism in paraphrased documents [8].

2. Semantics Based Plagiarism Detection:
This technique focuses on determining similarities in the use of words between documents stored in the given database using a vector space model. It is also capable of calculating the redundancy count of the words used in the document under review. It does not give accurate results for partially paraphrased documents as it cannot actually locate the plagiarized section in the submitted research paper [9].

3. **Clustering Based Plagiarism Detection**
A cluster-Based Plagiarism Detection method, use the grammar-based technique largely, by dividing it into three steps: first step called pre-selecting, so as to narrow the scope of detection using the successive same fingerprint; the second, called locating, is to find and merge all fragments between two documents using cluster method; the third step, called post-processing which deals with some merging errors. There are two traditional clustering algorithms implemented with document representation based on winnowing fingerprints, by adapting the similarity measures for working with multi-sets and designed a new way of centroid computation [10].

4. **Cross Lingual Plagiarism Detection**
This technique is used for detecting suspected documents plagiarized from other language sources. In this method, the similarity between a suspected and an original document is evaluated using statistical models to establish the probability that the suspected document is related to the original document regardless of the order in which the terms appear in suspected and the original documents. This approach necessitates the construction of the cross-lingual corpus [11].

5. **Citation Based Plagiarism Detection**
This technique is used for identifying academic documents that were read and used without referred to those documents. It actually belongs to semantic plagiarism detection techniques because it focuses on the detection of semantic content in the citations used in a text academic document. It intends to identify similar patterns in the citation sequences of academic works for similarity computation [12].

6. **Character Based Plagiarism Detection**
Character Based Plagiarism Detection has two subtypes namely, Fingerprinting and String Matching. In the fingerprinting technique, the pre-processing step involves creating representative digests of documents by selecting a set of multiple substrings using n-grams from them. These digests are referred to as fingerprints.

A suspicious document's passages are compared to the reference corpus based on their computed fingerprints. Fingerprint matching with those of other documents indicate shared text segments and suggest potential plagiarism if they exceed certain similarity threshold. Duplicate and near duplicate passages are assumed to have similar fingerprints [13].

6.1 Intrinsic Plagiarism Detection Methods

Plagiarism is detected without using any reference corpus. There are three general techniques in the Intrinsic Plagiarism Detection methods which are as follows:

a) Grammar Semantics Hybrid Plagiarism Detection

The base of this technique is Natural Language Processing (NLP) and thus makes it a good choice for intrinsic plagiarism detection. It can determine Paraphrasing and Mosaic types of plagiarisms in research papers. By calculating similarity measures between the words written, it can locate the plagiarized sections in the document [14].

b) Structure Based Plagiarism Detection
This technique focuses on structure features of the text in the document such as headers, sections, paragraphs, and references [15].

c) Syntax Similarity Based Detection
This technique is successful in the research field. Syntactical features are manifested in part of speech (POS) of phrases and words in different statements. Basic POS tags include verbs, nouns, pronouns, adjectives, adverbs, prepositions, conjunctions and interjections [16].

7. PLAGIARISM DETECTION TOOLS

The existing online tools and desktop tools that are currently available all detect plagiarism in textual documents and source code. Following Table1 shows that the exiting tools that are available to check plagiarism in text documents and source code.

Table 1: Different plagiarism detection tools

Text plagiarism detection tools **Online Paying**		
Name of Tools/ References	**Uses**	**Languages Supported**
Ephorus [17] http://www.ephorus.com/hhom	Ephorus is composed of three services: Ephorus Internet compares ith documents on the Internet, Ephorus Group with documents of parallel student groups, and Ephorus Database with documents handed in before or at other educational institutes with an Ephorus account	English, Spanish, Portuguese, German, Finnish, Swedish, Norwegian, Danish, Dutch, French, Italian, Polish, Russian, Turkish Greek, Croatian, Serbian Bosnian, Czech, Arabic
Plagiarism Scanner [18] http://www.plagiarismscanner.com/	Plagiarism Scanner is a commercial online plagiarism detecting application which runs against Internet resources, that is websites, digital databases and online libraries such as Questia or ProQuest.	
Safe Assign [19] http://safeassign.com	Safe Assign is a plagiarism prevention service which is not independent, but offered at no additional cost as a part of Blackboard products (Blackboard sells solutions in virtual learning environments).	English, Arabic , Chinese ,Dutch French , German, Japanese, Spanish.

Turnitin **[20]** **http://turnitin.com/static/index.php**	Turnitin is commercial anti plagiarism most popularly used system. Turnitin stores and computes unique fingerprint for a given document. It computes detailed document similarities for a selected set of documents with similar fingerprint. Internal document storage is composed of archived student papers, journals, periodicals and books. The document storage is being enlarged by automatic web page crawling.	Turnitin supports 19 languages: English, Arabic, Chinese (Traditional and Simplified), Dutch, Finnish, French, German, Italian, Japanese, Korean, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, Turkish and Vietnamese.
Urkund **[21]** **http://www.urkund.com/int/en/**	Ukund is an automated online plagiarism detection system. Its system is easier to use than previous ones, since the entire process is automated by email sending (no need to access to a site or login),and therefore it only requires that you know how to send and read emails	
Noplagiat.com **[22]** **http://www.noplagiat.com/**	Noplagiat.com is a French online detection tool with a very simple interface. It can be interesting if you are looking for something very simple to use, or for an occasional use, with no subscription. The user sends his documents to the site through a form, then he launches the analysis and the engine checks for similarities with contents found on the internet or also in an internal database.	
compilatio.net **[23]** **http://www.compilatio.net/en/**	Compilatio.net is an interesting ,Antiplagiarism solution, since it offers a different point of view from other tools.	
Pompotron.com **[24]** **http://www.pompotron.com/**	The user sends his document and the plagiarism detection is launched. Many formats are supported for the detection. Once the analysis done, the user gives his mail address and is directed to the payment step.	
Paying Desktop		
Plagiarism Detect **[25]** **http://www.plagiarismdetect.com/**	The fact that Plagiarism Detect emphasizes this plugin reflects that the concept of saving time by doing the plagiarism detecting task directly in an editor is as interesting as an online solution for some people. However, it limits the format of documents to MS Word, which is not very practical	
Plagiarism Detector **[26]** **http://plagiarism detector.com/**	Plagiarism Detector is a standalone computer desktop application for plagiarism detection, which runs only on Windows.	
EVE2 **[27]** **http://www.canexus.com/**	EVE2 is another commercial anti plagiarism system. For an input document it returns links to web pages from which an author could plagiarized.EVE2 uses" advanced searching tools" to locate suspect sites. It compares both the given and the found document and highlights" plagiarism" in red.	These systems were developed only for English, while other programs were adapted to deal with French, German and Chinese languages
Copy Catch **[28]** **http://www.cflsoftware.com/**	A UK system which concentrates on comparison within a group of students. The software compares text from work collected by email or on disk using a similarity threshold that will detect essays which are very similar or dissimilar to other class essays by communality of words and phrases	
Free Online		
Plagium **[29]** **http://www.plagium.com/**	Plagium is a very simple online plagiarism detection tool. You just have to paste your original text, And Plagium will search for redundancies over the web. There are many free,	

	online tools, but most of them look like Plagium, meaning they are very simple, with just a copy- paste system.	
Plagiarism Checker [30] http://www.plagiarismchecker.com/	SeeSources.com is also an online plagiarism detection tool. It resembles to Plagium and many other free online tools, but here you can also load documents in MS Word, HTML and Text format.	
See Sources [31] http://www.plagscan.com/seesources/	SeeSources.com is also an online plagiarism detection tool. It resembles to Plagium and many other free online tools, but here you can also load documents in MS Word, HTML and Text format.	
Copy scape [32] http://www.copyscape.com/	Copy scape is another variant of free online tool; nevertheless its particularity is that it is designed for checking plagiarism of web pages only.	
Plagiserve [33] http://www.plagiserve.com/	The service is based in Ukraine.	
Dupli Checker [34] http://www.duplichecker.com/	Dupli Checker just automates a process that the user could do himself.	
Dupli Checker [34] http://www.duplichecker.com/	Dupli Checker just automates a process that the user could do himself.	
Free Desktop		
Viper [35] http://www.scanmyessay.com/	Viper is free plagiarism detection software, exclusively for Windows and in English. According to them, it scans over 10 billion online sources including websites, online journals, and news sources.	
Free, Open Source		
WCopyfind [36] http://plagiarism.phys.virginia.edu/Wsoftware.html	One advantage of WCopyfind is that it can compare several documents at the same time. It can then indicate if one file is a copy of another file, or if they are both copies of a third document.	
Copy Tracker [37] http://copytracker.ec-lille.fr/	Copy Tracker is plagiarism detection software developed by a team at the Ecole Centrale de Lilles, and distributed under a General Public License.	
Source code Plagiarism detection tools. Paying		
Code Match [38] http://www.safe-corp.biz/	Code Match has also some additional functionalities, which allow finding open source code within proprietary code, determining common authorship of two different programs, or discovering common, standard algorithms within different programs.	BASIC, C, C++, C#, Delphi, Flash ActionScript, Java, JavaScript, MASM, Pascal, Perl, PHP, PowerBuilder, Ruby, SQL, Verilog, VHDL
Marble [39] foswiki.cs.uu.nl	Marble uses a structure based approach to compaire the submissions .	Java, perl,php,Xslt
Source code Free Online		
SID [40] http://genome.math.uwaterloo.ca/SID/	SID is an online application, which detects similarity between programs by computing the shared information between them. It was originally an algorithm developed for comparing how similar or Dissimilar genomes are. It was then realized that this algorithm could be extended to many other applications including finding chain letter history and detecting plagiarism	
Moss [41] http://moss.stanford.edu/general/scripts/mossnet	Moss is an automatic system for determining the similarity of programs and detecting plagiarism in programming classes. Moss is a free service but the users must create an account	C, C++, Java, C#, Python, Visual Basic, Javascript, FORTRAN, ML, Haskell, Lisp, Scheme,

		Pascal, Modula2, Perl, TCL, Matlab, VHDL, Verilog, Spice, MIPS Assembly 8086, HCL2.
Source code Desktop		
SIM **[42] http://www.cs.vu.nl/dick/sim.html**	It can be used to detect potentially duplicated code fragments in large software projects, in program text, in shell scripts and in documentation.	C, Java, Pascal and natural language
JPlag **[43] https://www.ipd.uni-karlsruhe.de/jplag/**	JPlag can compare two directories, or two files between them. It displays results in HTML Format, showing histograms of similarity values found for all pairs of programs, similar pairs and their similarity Values. Similar lines are matched with the same color.	Java, C#, C, C++, Scheme and natural language text.
Free, Open Source		
AC **[44] http://tangow.ii.uam.es/ac/**	AC is an anti-plagiarism system for programming assignments. It currently supports programs written in C, C++ or Java. AC incorporates multiple similarities detection algorithms found in the scientific literature, and allows their results to be visualized graphically. It is distributed under the General Public License.	C, Java, natural language
Sherlock **[45] http://sydney.edu.au/engineering/it/scilect/sherlock/**	Sherlock is a program which finds similarities between textual documents. It works On text files, as well as source code files. It uses digital signatures to find similar pieces of text. A digital Signature is a number which is formed by turning several words in the input into a series of bits and joining those bits into a number.	
Baldr **[46] http://labs.esiea.fr/2007/10/11/baldr/pl**	Baldr is source code plagiarism-detecting software. It has been programmed in Java and therefore it allows a multiplatform use. This software compares source	
angen	codes of a large number of files.	
Plaggie **[47] http://www.cs.hut.fi/Software/Plaggie/**	Plaggie is a stand-alone source code plagiarism detection engine purposed for Java programming exercises. It can compare two directories containing several files, or two files. It generates a report In HTML format, showing percentage of similarity values between projects.	Java 1.5
PMD **[48] https://en.wikipedia.org/wiki/PMD**	The PMD open source tool provides a Copy/Paste Detector (CPD) for finding duplicate code. CPD uses the Karp-Rabin string matching algorithm.	

8. CONCLUSION

In this paper the issues relevant to plagiarism detection are discussed as it is one of the most publicized forms of text reuse around us today. This paper covers the different types of plagiarism, different types of plagiarism detection methods and general techniques which are beneficial to the research scholars. The available plagiarism detection tools have been briefed. Now a days Turnitin and Viper are the mostly used plagiarism tools in universities and academic areas for detecting plagiarism. These tools are freely available online and more features included in that tools. Due to that features they are costly. Antiplagiarism tool will be developed for Marathi language using Marathi text corpus. In that tool extrinsic features will be extracted. On the basis of that features the antiplagiarism tool will be designed. A web based system will be developed. That tool will be helpful to all research scholars.

9. ACKNOWLEDGMENT

We are thankful to the Computational and Psycho-linguistic Research Lab, Department of Computer Science & Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) for providing the facility for carrying out the research.

10. REFERENCES

[1] http://www.jnu.ac.in/Library/RameshCGaur.htm

[2] WEBER WULFF, D. Copy, Shake, and Paste - A blog about plagiarism from a German professor, written In English. Online Source. Retrieved Nov. 28, 2010 from: http://copyshake-paste.blogspot.com, , Nov. 2010

[3] LANCASTER, T. Effective and Efficient Plagiarism Detection. PhD thesis, School of Computing, Information System and Mathematics South Bank University, 2003.

[4] Barnbaum, C., "Plagiarism: A Student's Guide to Recognizing It and Avoiding It.", ValdostaStateUniversity,http://www.valdosta.edu/cbarnbau/personal/teaching_MISC/plagiarism.htm (Accessed 23 January 2006).

[5] Liles, Jeffrey A. and Michael E. Rozalski., "It's a Matter of Style: A Style Manual Workshops for Preventing

Plagiarism.", College & Undergraduate Libraries, 11 (2) , p. 91-101, 2004.

[6] Maurer H., Kappe F. , and Zaka B. ,Plagiarism- A survey. Journal of Universal Computer Science 12,8, 1050-1084, Aug. 2006.

[7] Bretag T., and Mahmud S. , self-plagiarism or Appropriate Textual Re-use, Journal of Academics Ethics 7, 193-205, 2009.

[8] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, and Vaclav Snasel, "Overview and Comparison of Plagiarism".

[9] Clough, P. Old and new challenges in automatic plagiarism detection. Plagiarism Advisory Service, vol. 10, Department of Computer Science, University of Sheffield. 2003.

[10] http://www.ukessays.com/essays/information-technology/a-survey-of-plagiarism-detection-methods-information- technology-essay.php.

[11] Ahmed Hamza Osman, Naomie Salim and Albaraa Abuobieda, "Survey of Text Plagiarism Detection", Computer Engineering and Applications Journal 2012.

[12] Garfield e. ,Citation indexes for science: A new Dimension in documentation through assocation of ideas. science 122,3159,108-111, july 1955.

[13] http://en.wikipedia.org/wiki/Plagiarism_detection

[14] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, and Vaclav Snasel, "Overview and Comparison of Plagiarism Detection Tools", Department of Computer Science, VSB Technical University of Ostrava, 17. listopadu 15, Ostrava Poruba, Czech Republic, ceur-ws.org, Vol-706.

[15] T. W. S. Chow and M. K. M. Rahman, "Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection,", vol. 20, pp. 1385-1402, 2009.

[16] Salha Alzahrani, Naomie Salim, and Ajith Abraham, "Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods", SMIEEE.

[17] http://www.ephorus.com/home

[18] http://www.plagiarismscanner.com/

[19] http://safeassign.com/

[20] Turnitin web page: http://www.turnitin.com, (online 2011)

[21] http://www.urkund.com/int/en/

[22] http://www.noplagiat.com/

[23] http://www.compilatio.net/en/

[24] http://www.pompotron.com/

[25] http://www.plagiarismdetect.com/

[26] http://plagiarism detector.com/

[27] http://www.canexus.com/

[28] http://www.cflsoftware.com/

[29] http://www.plagium.com/

[30] http://www.plagiarismchecker.com/

[31] http://www.plagscan.com/seesources/

[32] http://www.copyscape.com/

[33] http://www.plagiserve.com/

[34] http://www.duplichecker.com/

[35] http://www.scanmyessay.com/

[36] http://plagiarism.phys.virginia.edu/Wsoftware.html

[37] http://copytracker.ec-lille.fr/

[38] http://www.safe-corp.biz/

[39] foswiki.cs.uu.nl

[40] http://genome.math.uwaterloo.ca/SID/

[41] http://moss.stanford.edu/general/scripts/mossnet

[42] http://www.cs.vu.nl/~dick/sim.html

[43] https://www.ipd.uni-karlsruhe.de/jplag/

[44] http://tangow.ii.uam.es/ac/

[45] http://sydney.edu.au/engineering/it/~scilect/sherlock/

[46] http://labs.esiea.fr/2007/10/11/baldr/plang=en

[47] http://www.cs.hut.fi/Software/Plaggie/

[48] https://en.wikipedia.org/wiki/PM