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 new 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

16
“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 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:

![Plagiarism Detection](image)

**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, uses 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 Table 1 shows that the exiting tools that are available to check plagiarism in text documents and source code.

<table>
<thead>
<tr>
<th>Table 1: Different plagiarism detection tools</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Online Paying</strong></td>
</tr>
<tr>
<td><strong>Text plagiarism detection tools</strong></td>
</tr>
<tr>
<td><strong>Name of Tools/ References</strong></td>
</tr>
<tr>
<td><strong>Uses</strong></td>
</tr>
<tr>
<td><strong>Languages Supported</strong></td>
</tr>
<tr>
<td>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</td>
</tr>
<tr>
<td>English, Spanish, Portuguese, German, Finnish, Swedish, Norwegian, Danish, Dutch, French, Italian, Polish, Russian, Turkish Greek, Croatian, Serbian Bosnian, Czech, Arabic</td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>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).</td>
</tr>
<tr>
<td>English, Arabic , Chinese ,Dutch French , German, Japanese, Spanish.</td>
</tr>
<tr>
<td>Tool</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>Turnitin</td>
</tr>
<tr>
<td>Ukund</td>
</tr>
<tr>
<td>Noplagiat.com</td>
</tr>
<tr>
<td>compilatio.net</td>
</tr>
<tr>
<td>Pompotron.com</td>
</tr>
<tr>
<td>Plying Desktop</td>
</tr>
<tr>
<td>EVE2</td>
</tr>
<tr>
<td>Copy Catch</td>
</tr>
<tr>
<td>Free Online</td>
</tr>
<tr>
<td>Tool Name</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>Plagiarism Checker</td>
</tr>
<tr>
<td>See Sources</td>
</tr>
<tr>
<td>Copy Scape</td>
</tr>
<tr>
<td>Plagiserve</td>
</tr>
<tr>
<td>Dupli Checker</td>
</tr>
<tr>
<td>Free Desktop</td>
</tr>
<tr>
<td>Viper</td>
</tr>
<tr>
<td>Free, Open Source</td>
</tr>
<tr>
<td>WCopyfind</td>
</tr>
<tr>
<td>Copy Tracker</td>
</tr>
<tr>
<td>Source code</td>
</tr>
<tr>
<td>Code Match</td>
</tr>
<tr>
<td>Copy Tracker</td>
</tr>
<tr>
<td>Source code</td>
</tr>
<tr>
<td>Code Match</td>
</tr>
<tr>
<td>Source code</td>
</tr>
<tr>
<td>SID</td>
</tr>
<tr>
<td>Moss</td>
</tr>
<tr>
<td>Free, Open Source</td>
</tr>
<tr>
<td>WCopyfind</td>
</tr>
</tbody>
</table>
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
[5] Liles, Jeffrey A. and Michael E. Rozalski., “It's a Matter of Style: A Style Manual Workshops for Preventing...


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


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


[36] foswiki.cs.uu.nl

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

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


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

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


