Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 8 (2017) pp. 2359-2365
© Research India Publications
http://www.ripublication.com

A Survey on Plagiarism Detection

Bhardwaj Akanksha[1], Arya Anukruti[1], Vyas Tarjni[1], Shivani Desai[1] and Anuja Nair[1]

[1]*Institute of Technology, Nirma University, 382481-Ahmedabad, India.*

Abstract

Plagiarism detection has been observed as most practiced activity in academia wherein students indulge in practice of copying in-formation or ideas of others which they find of useful without giving credit. Not just in academic institutions but in other fields like business, research this practice is common. Plagiarism is not all about copying text from others genuine work but it also includes copying ideas, source codes, patterns or style of writing. Several ways of plagiarism have been detected so far and various tools to identify such plagiarised material has been developed. This paper provides a brief idea about the types of plagiarism, how it is practised and ways to detect them.

Keywords: Plagiarism, Plagiarism Detection Methods and Tools, Plagiarism Detection Framework

INTRODUCTION

Plagiarism is a deliberate act of acquisition of someone else's work for ones own benefit without giving credit to the original owner(s). The derived word plagiarism was introduced into English language around 1620. Its original word in Latin Plagiarius means someone guilty of literary theft & was introduced into English in 1601 by dramatist Ben Jonson. Now a day it is a common practice in academic institutions where students tend to use research work or ideas of other rightful owners, in business world competitors steal ideas of the opponents to achieve their success goals. This problem is growing due to the unavailability of intelligent and efficient detecting tools. The existing tools usually do not detect intelligent plagiarism and image plagiarism. They only detect the similar text patterns or similar text written. It maps the text of the plagiarized document with the repository of relevant available source documents and returns the text, pattern or part of the documents which matches with the original source documents.

In this paper, we discuss the taxonomy of plagiarism which dictates various types and methods of plagiarism. The paper also covers the detection techniques through which we get a brief idea about text plagiarism detection.

Types of Plagiarism

"Immature poets imitate; mature poets steal" - TS Eliot

Plagiarism can be categorized into two types namely (i) Literal Plagiarism and (ii) Intelligent Plagiarism.

Both the types are different in way that how the plagiarist does plagiarism.

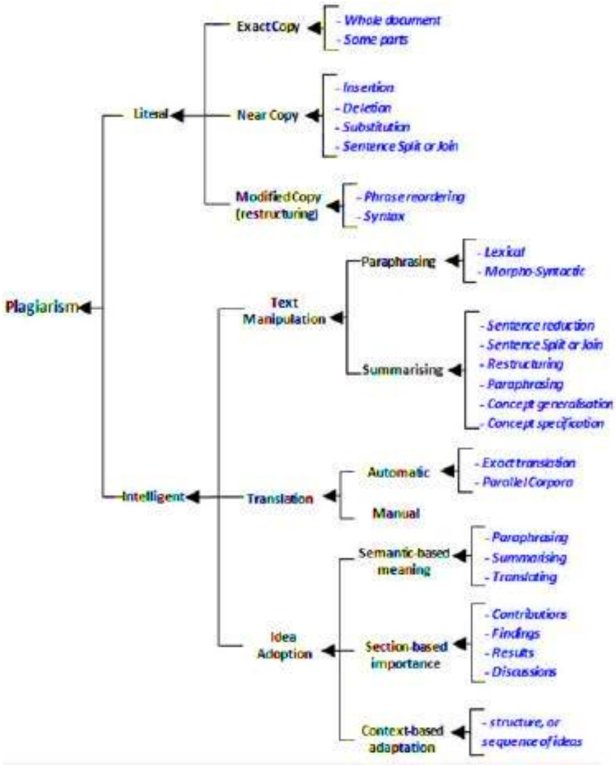
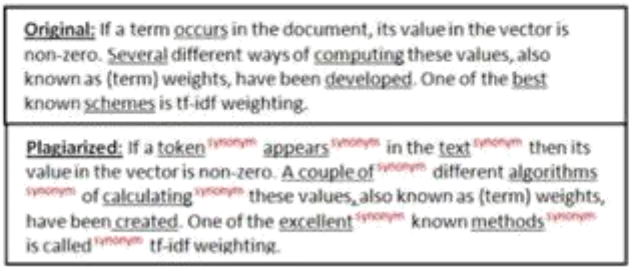


Figure 1. Types of Plagiarism [1]

The Literal Plagiarism is often practiced institute level where just copying of text is done from various sources with a little alteration like changing the tense or conversion of active voice to passive or adding extra spaces or replacing with synonyms in the original source. This type of plagiarism can be detected by any modern detection tools like plagiarism checker, grammarly.com, et al.

The Intelligent Plagiarism is a difficult one to detect. It is not just copying the text but modifying it in such a way that the meaning of the text remains same and it appears as a new idea. Intelligent Plagiarism tends to by-pass and obfuscate the detection tool in order to achieve plagiarism without being detected. It can be further classified into three types:

Text Manipulation: This includes modifying the text, using synonyms, para-phrasing or summarizing, sentence reduction, combination or reconstruction.



Original: If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed. One of the best known schemes is tf-idf weighting.

Plagiarized: If a token$^{\text{synonym}}$ appears$^{\text{synonym}}$ in the text$^{\text{synonym}}$ then its value in the vector is non-zero. A couple of$^{\text{synonym}}$ different algorithms$^{\text{synonym}}$ of calculating$^{\text{synonym}}$ these values, also known as (term) weights, have been created. One of the excellent$^{\text{synonym}}$ known methods$^{\text{synonym}}$ is called$^{\text{synonym}}$ tf-idf weighting.

Figure 2. Literal plagiarism pattern extracted from the collection of plagiarized short answers[1]

Translation: Translating a sentence into another language and changing it back to the former language will change the way the original sentence was written thus by-passing the detection in an intelligent manner.

Idea Adoption: It is similar to the way of representing an idea in a different manner so that it looks like a new one. This is mostly practiced in business world or research where competitor's idea is stolen to gain success.

"Borrowing a few words, but no original ideas, to improve the quality of the English, especially by non-natives, should not be considered plagiarism."

CLUSTERING IN WSN

Plagiarism detection is the procedure of finding the instances of plagiarism in a document, such as that of Research etc.[8]. The distributed use of computers and little help from the web has made it easier to plagiarize the work of others. But detecting the same with such variety of resources has now become possible. One popular example of such resource is TurnItIn.

The g3 is a black box testing design for Plagiarism detection System. It has three main components, one input being a set of query / questionable set of document dq that are to be tested. Another input is a set of or collection of Documents D which will help to detect plagiarism, such as the Web. The output of this system is a suspicious set of sections that contains plagiarized text. Plagiarism detection can be divided into two different tasks, namely Extrinsic and Intrinsic.

Extrinsic Plagiarism Detection: In this approach, the suspicious document dq is compared or what is considered to be a set of genuine documents called source/ reference documents D. Fig 4 shows the basic process of Extrinsic Plagiarism detection.

There are three steps involved in the process. Firstly, a set of likely sources of Plagiarism is Candidate documents Dx. Depending upon the model selected for the process and some pre-defined similarity criteria, all the documents that may contain text that is similar to a level above a threshold chosen are retrieved. Second, a pair-wise, feature based, exhaustive testing is done to match d_q to its corresponding

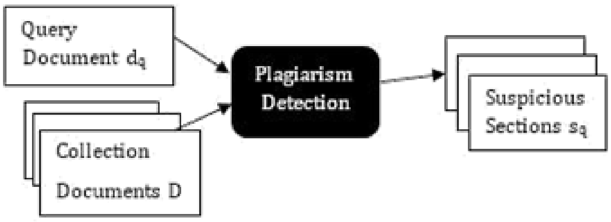


Figure 3. Black-Box design for Plagiarism Detection System[1]

D_x by taking n words (n-gram) or sentences per comparison, called as detected units. Third, a knowledge-based post-processing step is per-formed, that merges small detected units into passages or sections and to present the result to a human, who may decide whether or not a plagiarism offense is given in[1]. Thus, the final output is pair of sections (s_q; s_x); where s_q2 d_q; s_x2 d_x2 D_x:s_q is part that is plagiarised from s_x.

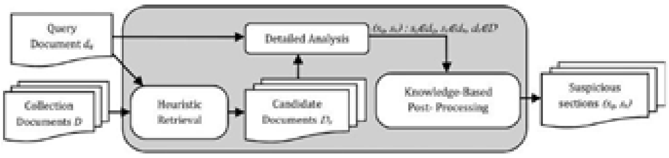


Figure 4. White box design for extrinsic plagiarism detection system[1]

Intrinsic Plagiarism Detection: It mainly focuses on analyzing the text of the suspicious document dq without comparing it to a set of source or candidate documents. This method mainly aims upon identifying changes in the unique writing style of an author, which acts as an indicator for potential plagiarism.

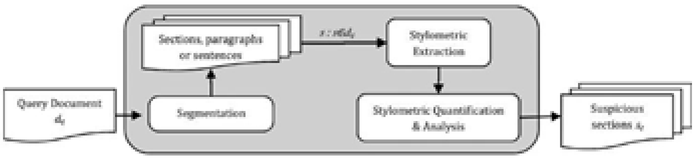


Figure 5. White-box for intrinsic plagiarism detection system[1]

"Intrinsic plagiarism aims at identifying potential plagiarism by analyzing a document with respect to undeclared changes in writing style. Authorship verification aims at determining whether or not a text with doubtful authorship is from an author A, given some writing examples of A, while authorship attribution aims at attributing a document d of unknown authorship, given a set D of candidate authors with writing examples."[4].

The procedure for Intrinsic Plagiarism[7] detection is as follows. The input is only a query document d_q. Three steps (as depicted in the rectangles) are as follows. Firstly, a d_q is divided into smaller parts, such as words, sections, fragments, sentences or paragraphs. Second, stylometric features are extracted from different units. Third, stylometric-based measurements and quantification functions are employed to analyze the variance of different style features. Parts with stylometric parameters which are usually uneven are marked possibly plagiarized and sent to humans for further investigation, i.e., the final output is fragments/sections s_q : s_q 2 d_q such that sq has quantified writing style features different from other sections as in d_q [1].

Monolingual Plagiarism Detection: This method of plagiarism detection deals with detection of plagiarism in homogeneous language, e.g.: English-English. Most of the plagiarism detection systems have been developed for monolingual detection, which is divided into two former tasks, extrinsic and intrinsic, as discussed earlier.

Cross-Lingual Plagiarism Detection: This method of plagiarism detection deals with detection of plagiarism in heterogeneous language, e.g.: English-Hindi, English-Chinese etc.

Plagiarism Detection Methods

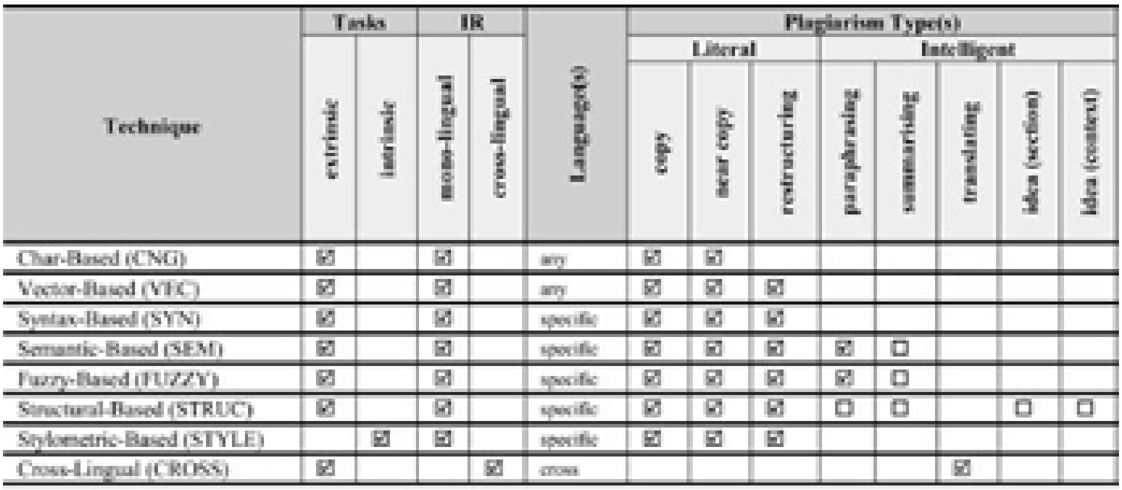
Various plagiarism methods have been proposed in order detect the type of plagiarisms. The tools that are available usually detect the Literal Plagiarism but are still unable to detect Intelligent Plagiarism like para-phrasing or idea-adoption. Basically, a detection method involves matching of strings or patterns of the plagiarized document with the available corpus collection of relevant documents based on query document through different IR techniques and models. For e.g. when submit a document to a plagiarism detection tool it performs pre-processing tasks like tokenization, stemming, pos-tagging and indexing and then it maps the words or sequence of words to the available relevant source documents to find the similarity. If matching exceeds threshold value then it returns plagiarized text highlighted with the source if available.

Various IR models like vector model, n-gram model has been used to count the term frequency and map the similarity between documents for detection and method like indexing is used for retrieval of the documents. Indexing uses linked list or hash tables to maintain the data, other data structures like tree can also be used for efficient retrieval [9]. Uses tree data structure for efficient retrieval and proposes a new system with it.

As discussed plagiarism could be of text, image, idea or source code. Source Code Plagiarism has been defined in [2] [3]. PlaGate [4] a source code detection tool which can be integrated with existing detection tool has been proposed.

The Plagiarism detection can be done through different approaches based on the type of plagiarism done. The table below compares the efficiency of different techniques used by detection methods.

Table 1. Comparison between different detection methods and their efficiency [1]



Technique	Tasks		IR		Language(s)	Plagiarism Type(s)							
						Literal			Intelligent				
	extrinsic	intrinsic	mono-lingual	cross-lingual		copy	near copy	restructuring	paraphrasing	summarizing	translating	idea (section)	idea (context)
Char-Based (CNG)	☑		☑		any	☑	☑						
Vector-Based (VEC)	☑		☑		any	☑	☑	☑					
Syntax-Based (SYN)	☑		☑		specific	☑	☑	☑					
Semantic-Based (SEM)	☑		☑		specific	☑	☑	☑	☑	□			
Fuzzy-Based (FUZZY)	☑		☑		specific	☑	☑	☑	☑	□			
Structural-Based (STRUC)	☑		☑		specific	☑	☑	☑	□	□		□	□
Stylometric-Based (STYLE)		☑	☑		specific	☑	☑	☑					
Cross-Lingual (CROSS)	☑			☑	cross						☑		

Text Based Plagiarism: This approach aims at finding text similarities between the source document and query document. It uses vector space model to find the similarity. Though other model like Boolean model can also be used it is not efficient in finding the partial similarity. If the similarity is found to be above threshold it is considered to be a serious plagiarism. The source document could be web, magazines, books, research papers, journals etc. The detection process stages has been discussed in[5].

Fingerprinting: This is one of the most widely used method to detect plagiarism. This method uses Hash technique to calculate hash value of each document by choosing a set of multiple substrings (n-grams) from them. These sets represent fingerprints and the elements are called minutiae. The minutiae of suspicious documents are compared with the other document(s) minutiae. The approach can be further classified into (i) character-based fingerprints, (ii) phrase-based fingerprints and (iii) statement-based fingerprints. These methods have been briefly discussed in [6].

Citation-Based Plagiarism: This approach relies on the citations and is the only method that does not depend on textual features. The method takes into account the citation & references pattern and check for the similarity between documents.

Stylometry: This method considers the style of unique writing of the author. No two human can have exactly same style of writing.

Shape-Based Plagiarism: This approach is to detect the plagiarism of figures and flowcharts. Most of the tools neglect the figures while checking plagiarism. This approach uses shape-based image processing and multimedia retrieval [5].

Clustering: It uses specific terms to find the similar clusters between documents. It is an exhaustive type of approach [1].

CONCLUSIONS

As discussed Plagiarism is highly practiced in academic institutions and various detection tools has been developed and proposed based on different types of information retrieval models and techniques yet no tool has been introduced that can detect intelligent plagiarism specifically idea adoption. Turnitin is a detection tool used by major institutions but its focus is mainly on text detection. Various techniques like semantic analysis and fuzzy logic has been implemented to make detection tools detect intelligent plagiarism.

REFERENCES

[1] Salha M. Alzahrani, Naomie Salim, and Ajith Abraham: Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods, IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews, Vol. 42, No. 2, March 2012

[2] G. Cosma and M. Joy: Towards Definition of Source-Code Plagiarism, IEEE Trans. Education, vol. 51, no. 2, pp. 195-200, May 2008

[3] G. Cosma and M. Joy: Source-Code Plagiarism: A U.K Academic Perspective, Research Report, No. 422, Dept. of Computer Science, Univ. of Warwick, Coventry, 2006.

[4] Georgina Cosma and Mike Joy: ”An Approach to Source-Code Plagiarism Detection and Investigation Using Latent Semantic Analysis”, IEEE Transactions On Computers, Vol. 61, No. 3, March 2012.

[5] S.A.Hiremath and M.S.Otari: Plagiarism Detection-Different Methods and Their Analysis: Review, International Journal of Innovative Research in Advanced Engineering

[6] (IJIRAE) ISSN: 2349-2163, Volume 1, Issue 7 (August 2014).

[7] Salha Mohammed Alzahrani & Naomie Salim: Ch-3: Plagiarism Detection Techniques,

[8] http://www.academia.edu/1458324/Plagiarism Detection Techniques.

[9] B. Stein, N. Lipka, and P. Prettenhofer: Intrinsic plagiarism analysis , Language

[10] Resources and Evaluation, January 2010.

[11] https://en.wikipedia.org/wiki/Plagiarism detection.

[12] Tommy W. S. Chow and M. K. M. Rahman: ”Multilayer SOM With Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection”, IEEE Transactions on Neural Networks, VOL. 20, NO. 9, September 2009.