ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 9, September 2018

A Review on Various Plagiarism Detection Systems Based on Exterior and Interior Method

Nosheena Khan[1]*, Chetan Agrawal[2]*, Tehreem Nishat Ansari[3]

Department of Computer Science and Engineering, RITS, Bhopal, M.P[1,2,3]

Abstract: Plagiarism is an intellectual theft. It means representation of author's output as self - contribution without mentioning reference or attribution. Plagiarism is considered as a fraudulent act which simply means forgery of someone's fresh content by not paying any tribute. It is proliferated due to increasing access to multiple resources and simply using copy paste method. With the immense growth of internet resources new technologies and original ideas or innovative thoughts are available effortlessly due to which protection over such intellectual property is challenging, but important. Therefore, in this paper we delineate about plagiarism and then classification of plagiarism. We also discuss two main approaches- Exterior and Interior methods for plagiarism detection with their characteristics methods based on their grammar styles, lexical, syntax or Stylometric features. Further, literature analysis of various algorithms is described to find out their issues and challenges improvised scope of work in future.

Keywords: Plagiarism, Classification of Plagiarism, Textual Plagiarism Detection, External Plagiarism detection, Internal Plagiarism detection

I. INTRODUCTION

Plagiarism is considered as an act of gaining someone's academic knowledge illegally without providing credit or acknowledgement [1, 2]. With the incremental growth of various internet resources, availability of large number data is easily possible due to which plagiarism and copyright infringement occurs. Hence, it's a need to give attention on detection of plagiarism by comparing the test documents with the registered documents. It is considered as forgery or piracy of violating laws for copyright document. It is based on two attributes (1) Stealing of another person's words, texts without any acknowledgement or (2) Adapting someone's ideas, restructuring or changing its grammar style [3]. In this scholarly time students, professors or researchers perform some common form of plagiarism activities such as using someone's work without quoting the sources. Digitalization enables easy availability of text on web interrelated to several academic areas. Due to this problem several authors rewrite the data in their documents from original sources.

There are mainly two types of plagiarism [4]:

1. Source code Plagiarism
2. Natural languages processing

Classification of Plagiarism

Plagiarism can be seen in two varieties which depend on the basis of work done on any document or program

(a) **Textual Plagiarism** and (b) **Source code** [4]. Scamming of text can be seen where someone uses words, ideas and simply presenting in other way such as using synonyms or improving its grammar structure. Here we will chiefly concentrate on **Textual Plagiarism** which is easily performed in research and education.. Some common types of textual plagiarism are listed below:

a) **Self-Plagiarism:** It is the practice or act of recycling your own previous works by manipulating its text or its writing pattern in order to hide its forgery means recycling of old research or publishing new research paper without citation [3].

b) **Accidental Plagiarism:** When a plagiarist does not know academic requirement he/ she unknowingly translates a portion of a data by acknowledging related words, groups of text, or sentence structuring are used without any attribution or due to lack of understanding user is unaware of mentioning the resource[5].

c) **Replicate/Clone Plagiarism:** It is known as substitution in a piece of text of an author's work or extended portion of a source without citation or duplicate of data, copying or ripping whole document or a section of someone else's demonstration [5, 6].

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 9, September 2018

d) Mosaic-Plagiarism: It occurs when generally someone reuse series of sentence without using quotation marks along with substitution of synonyms, omission of words, or improvising its grammatical structures and often retention of the same sentence structure as the original source [7].

e) Idea Plagiarism: Taking up or claiming of ideas from someone such as translating their concept or attempt to steal someone's findings, conclusion or result without giving credit or taking permission [8].

f) Metaphor Plagiarism: In this type of plagiarism author uses an imaginative way of giving someone's knowledge or act of hiding the original words which maybe similar in a particular way without giving credit to the foremost source [7, 8].

g) Structural Plagiarism: It involves change in sentence agreements, translating someone's work by changing its grammar structure. Likewise, simulating the original sentences or piece of work or organization without appropriate citation [5].

On the basis of their characteristics, we can further categorize into two types ***Literal*** and ***Intelligent plagiarism***. The former is comprised of Replicate, Self-plagiarism, Accidental and Mosaic plagiarism and the later one is consist of idea, Metaphor and Structural plagiarism. It is briefly shown in fig.1

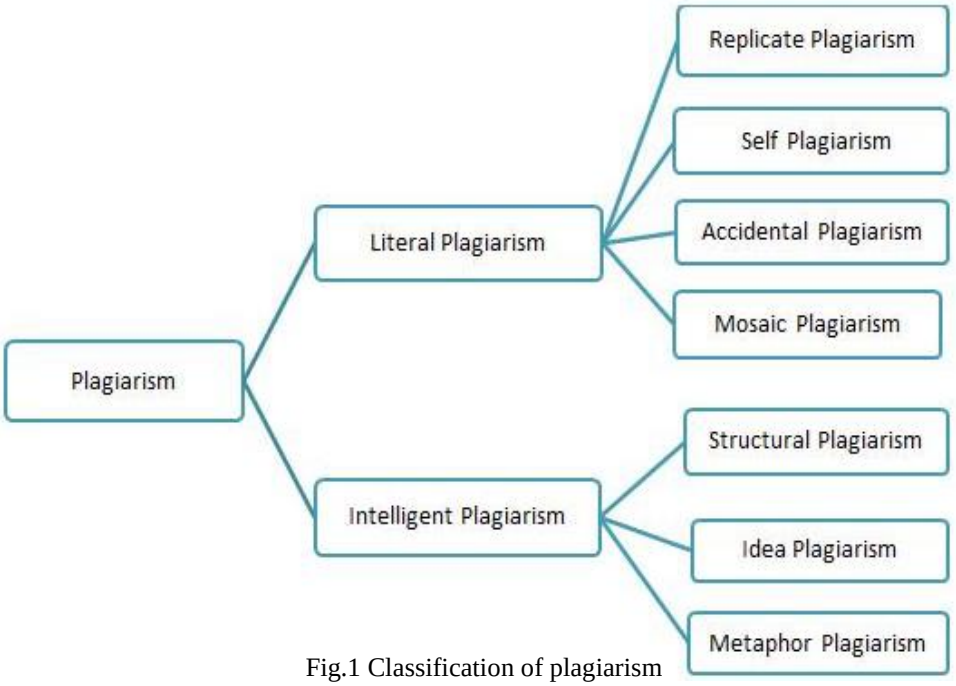

Fig.1 Classification of plagiarism

Therefore, the organisation of this paper is given as follows: The II section describes two main detection methods- external plagiarism detection and internal plagiarism detection method along with their general techniques are delineated to distinguish between kinds of plagiarism based on their textual features, Stylometric, semantical and syntactical features for the comparison between suspicious document and the original document. In III section a brief survey analysis of related works is given to find the challenges and issues in the previous algorithm. Then in IV section explains the objective of the paper and further in V section give some problem statement which describes the difficulties came in to previous algorithm. Then, in VI section a brief discussion of future scope enhancement is given to detect plagiarism and then in VII section we conclude about the core discussion of plagiarism and their detection methods.

II. PLAGIARISM DETECTION METHODS

The two main methods for textual plagiarism detections are: Exterior and Interior plagiarism detection and their general techniques are shown based on their different types of textual features for comparing document to measure similarity [10].

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 9, September 2018

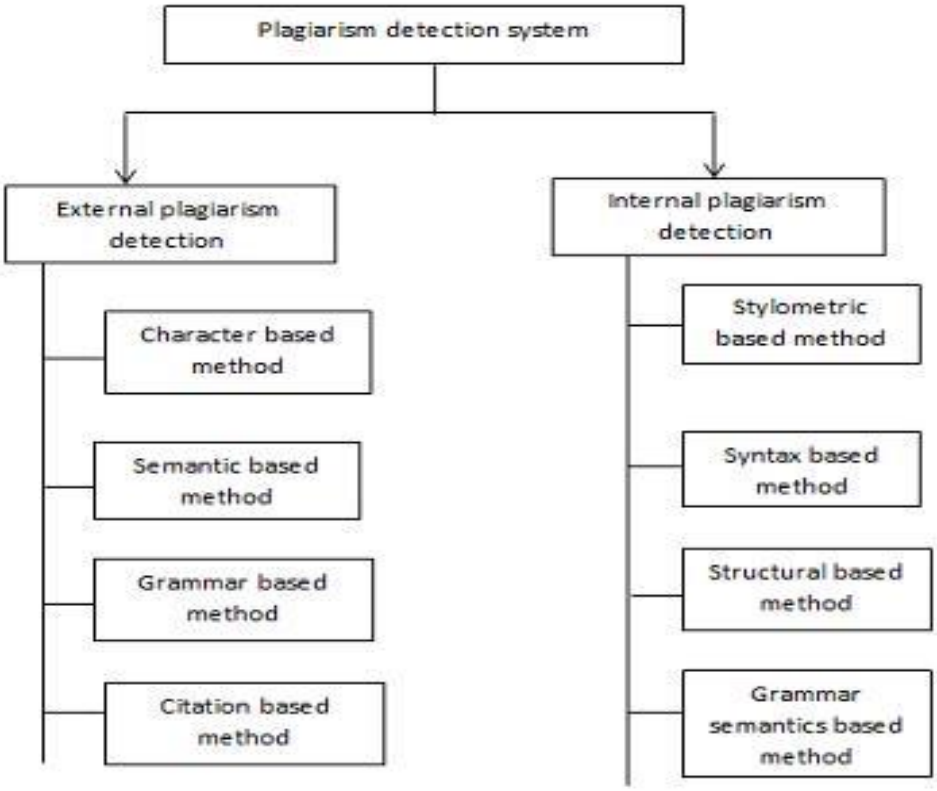

Fig.2: Types of Plagiarism Detection General Techniques

External Plagiarism Detection Methods

In this type of plagiarism detection, we detect similarity between suspicious document and the collection of documents which are already published their scholarly articles available to all publicly. It will need reference dataset collection to find similarity [11, 12].

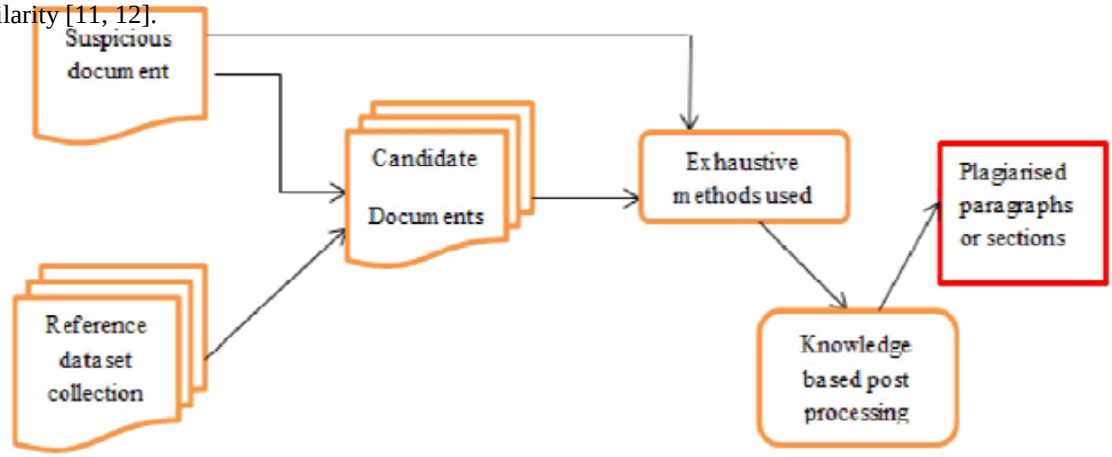

Fig.3 External plagiarism detection

There are various features on which plagiarism detection is carried out such as lexical features, Stylometric and semantical features. Fig.2 describes the general techniques of both detection systems.

a) Character based method

This method uses different based on characters, syntactical features to calculate similarity between suspicious document and original document. There are two types of scenario in case of calculating the correlation between documents either exact matching or approximate matching [14].

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 9, September 2018

Basic formula used for calculating distance using n-gram distance method.

$$d_n(x,y) = \frac{1}{|D_n(x)|+|D_n(y)|} \sum_{w \in D_n(x) \cup D_n(y)} \left(\frac{f_y(w) - f_x(w)}{f_y(w) + f_x(w)}\right)$$

b) Semantic based method

A sentence is collection of words assembled and arranged in specific order. There may be a case where two distinct sentences can be attributively same but their grammatical structure can be different by doing some changes in their expressions. For ex- A sentence can be rebuilt by simply converting it in active voice to passive voice or vice versa [15]. This type of methods is limited in use because it is difficult to find out semantic similarity measure for sentences.

c) Grammar based method

In this method similarity measure is computed for plagiarism detection by using string-matching approach between suspicious and original documents already available in the database. It can easily detect literal plagiarism such as clone documents but fails to detect paraphrased sentences and documents which contain intelligent plagiarism [16].

d) Citation based method

This method is used to detect plagiarism in which author has not addressed tribute and used other paragraph or section of text [17]. Basically, this method is used semantics to relate as it uses this approach to find the semantics present in the citation that are used in the document. After analysing similarity pattern of the citation sequences similarity score is measured.

Internal Plagiarism Detection Methods

In this type of plagiarism detection, we detect plagiarised document by analysing the document alone not from the collection of documents [13]. It does not need any reference dataset collection.

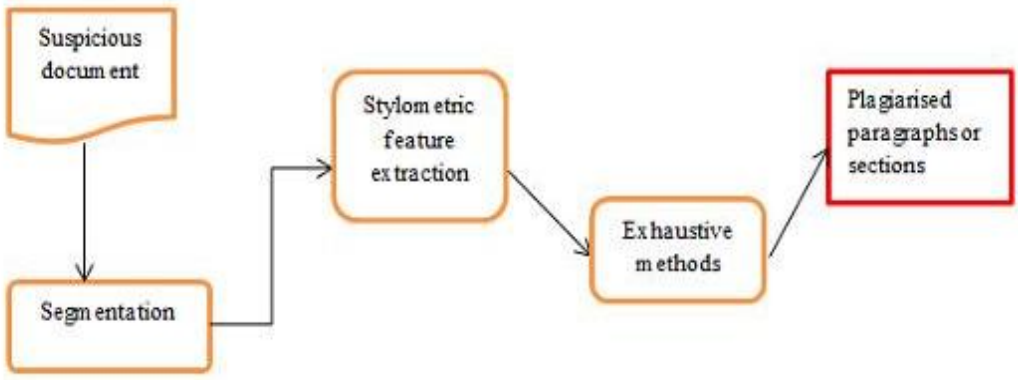

Fig.4 Internal plagiarism detection

a) Stylometric based method

This method uses Stylometric features to analogize two different documents based on their pattern or style of writing to detect plagiarism [18]. There are two formulas for determining the plagiarism based on its representation: Writer specific and Reader specific.

b) Syntax based method

As the name suggests this method uses syntactical features like parts of speech (POS) tags of sentences or phrases to find out plagiarism. It means documents contain same POS tagging is applied for comparison and analysis for the detection of plagiarism [19].

c) Grammar Semantics Hybrid based method

This method is the hybridisation of both grammatical and semantical approach which can override the drawback of semantic based method [20, 21]. It is effective to detect plagiarism in clone documents as well as restructured documents.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 9, September 2018

III. RELATED WORKS

Research work of [23] provides Latent Semantic Analysis (LSA) which is used to find correlation in a series of words in a text and records patterns of the word that are dormant hidden semantically. LSA is approximate model that find similarity by using SVD and map reduction. This method performs data pre-processing and transform it by removing special symbols and stemming procedure. The PAN-PC-11 is the dataset for the procedure, pre-processing is performed to eliminate unreliable information and to get trained dataset. This trained dataset is compared with the input file and cosine similarity measure is used.

Author [24] proposed a novel technique to detect suspicious text documents over original documents. Their dataset is PersianPlagDet text documents then data preparation is done step by step such as tokenization, POS tagging, cleansing and finding nouns and synsets through Farsnet after it they used trie tree data structure which enhance quick searching and insertion. At last they used macro-averaged precision and granularity measure for evaluation.

In this study [25], author delineated Multi-agent system following an algorithmic procedure of map reduction as well as parallel processes can be done by applying multithreading procedure. Implementation methodology has five layers and each layer is with following functionalities like data cleaning, pre-processing to remove stop word and symbol, multithreading-procedure, use of automated web crawler to examine web documents and n-gram technique is implemented and then correlational similarity score is calculated.

In this paper [26], author proposed a proficient algorithm for detecting plagiarism which is based on abstract syntax tree (AST) and further computed hash values and compares them. It introduces AST-code comparison algorithm which reduces the storage format as the syntax tree is converted into array of linear linked list. After that it called hash comparison algorithm. In order to implement the algorithm more effectively, special measurement is taken to reduce the error rate when calculating the hash values of operations, especially the arithmetic operations like subtraction and division.

In this paper [27], author deals in the problem to find plagiarism in source code. Therefore, he discussed five ways by which we can detect plagiarism in software programs by string matching, tokenization, AST and program dependency and metrics. The tools they used for source code plagiarism detection can be used in various software industries and in academic areas such as colleges, school etc. are JPlag, SIM, MOSS and plaggie which has done tokenization and then compared them.

In this paper [28], the researcher proposed an algorithm known as longest common consecutive series for the discovery of plagiarism in the source document. In this algorithm the document is broken down into small paragraphs then each paragraph is divided into text of consecutive words and from this words triangle can be carved out by locating longest common consecutive word (LCCW) by conducting paragraph by paragraph. This algorithm is compared with suffix tree algorithm as the former one is evaluation of the later one by considering space and time complexity as precision metrics. LCCW algorithm is approximate matching.

In this paper [29], author efforts to analyse the source code for plagiarism. Find the important words from the program with the help of tokenization. They use Running-Karp-Rabin and Greedy-String-Telling (RKR-GST) algorithm over the winnowing algorithm. In GST is used to find the longest common subsequence between the two strings and RKR is to find short substrings and for matching. After getting token streams they create abstract syntax tree to successfully detect plagiarism in the source code. Many source codes can be detected with the help of Latent Dirichlet Allocation (LDA) which represent a given algorithm by means of topic model.

Table 1: Comparative Studies

Author	**Approach performed**	**Description**	**Limitation**	**Plagiarism detection methods**
Rajkumar Kundu, Kartik. K **[23]**	**Latent Semantic Analysis**	LSA used to find out the semantic it uses SVD and reduction to capture all similar text.	It is a distributed model and not fit for also not suitable for nonlinear equations.	**External**
Alireza Talebpour et al **[24]**	**Plagiarism Detection Based on Trie-tree based data structure**	Both character- based and knowledge-based approaches are used for comparing data at high speed.	A Comparison based technique required processing of content which is not an efficient solution for large number of files.	**External**
S.N. Autade et.al **[25]**	**Evolutionary multi-agent system**	Synonym recognition and word -generalization Is	A large word dictionary update is	**External**

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 9, September 2018

		used.	Required.	
Jingling Zhao et. Al **[26]**	**An AST-Based Code Plagiarism Detection Algorithm**	They proposed AST- CC algorithm to generate hash values and compare them.	Less efficiency for the storage of data structure.	**Internal**
Mayank Agrawal et al **[27]**	**A State of Art on Source Code Plagiarism**	To find source code plagiarism various tools are used and various methods like using NLP and machine learning and compare.	It is difficult to find out plagiarism between different source codes of different languages.	**External**
Agung Sediyono et al **[28]**	**Longest common consecutive word algorithm**	It is numerical based comparison algorithm that outsource suffix tree algorithm.	The drawback of this proposed algorithm is loading time.	**External**
Michal Ďuračíka et. al **[29]**	**Detection of clones and methods for determining similarity.**	It provides anti- plagiarism system which is to handle large amount of dataset.	It is need to process the multiple documents with similar identity which exhibit high execution.	**External**

IV. OBJECTIVES

Web or article content with its originality always impact for the end user and original content impact high search usage, high visibility on the web platform. So, keeping it genuine for the originality point of view it is an important task to detect copy infringement of original document.

Therefore, keeping this in mind a proper plagiarism detection method is required. Our main objective towards this dissertation is pointed below:

1. Finding a proper approach which can deal with the multiple formats of documents available on the web.
2. Finding a proper solution, this can deal with the multiple references and http protocol related data extraction.
3. Finding a method and proper similarity measure analytics, this can deal in finding proper plagiarism in provided input document.
4. Comparison between the existing algorithms in the field of plagiarism detection and new proposed algorithm.

V. PROBLEM STATEMENT

As per the previous technique discussion is made, there are following system challenges which needed to monitor while implementing the document similarity matching algorithm system.

1. A large processing of document need an efficient system which can process such big data files.
2. Maintenance of large data dictionary and updatable content for effective similarity matching.
3. Finding document scenario, generating the permission context, making it accessible for detection purpose.
4. Using the multiple API system which helps in understanding different document formats.

Thus, these are the core challenges while building any of the system related with document similarity finding algorithm.

VI. FUTURE SCOPE

In order to overcome the limitations of the previous applied algorithms mentioned in the above comparative analysis of various research paper. Some enhancements are as follows:

1. To find plagiarism by considering all parameter and then some appropriate further enhancements are done in the previous approaches.
2. Compressive Data Access & Lexical Indexing based Semantic Approach is proposed in which three different phases are carried using the semantic based approach.
3. To find the amount of time taken to find similar contents and finally get the value of precision and recall.
4. To reduce the computation time and find the relevant document from the list of the documents and improve efficiency.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 9, September 2018

CONCLUSION

Plagiarism is a term which comes with the similar content publishing or making over the internet. Many researches outperformed in the past which need always a unique content of publishing. But the original research is also used by different researcher to use in personal content. On considering the influence of plagiarism occurring due to availability of various web articles or it is represented in number of ways. This survey paper has come out with a systematic classification of plagiarism and also encloses text plagiarism systems. This paper has given information about an exhaustive research on impact of plagiarism and scale out their general techniques. A large number of tools had come into existence but there are still some challenges and issues to be answered. Thus, the avoidance of such scenario is needed to be investigated. In this paper survey of previously given technique for plagiarism detection is discussed along with their comparative analysis to figure out its approach and limitations. Thus a further work is going to be carried out in working with the given proposed solution for plagiarism content detection over the document which works on textual data.

REFERENCES

[1]. C . Justicia de la Torre et al - Text Mining: Intermediate forms on knowledge representation in EUSFLAT-LFA 2005.
[2]. Lea M. R. and Street B. 2014 - Understanding textual practices in higher education. *Writing: Texts, processes and practices* (2014), 62.
[3]. M. S. Anderson, N. H. Ste neck - The problem of plagiarism in Seminars and Original Investigations, Vol. 29, Elsevier, 2011, pp. 90-94.
[4]. Hussam M. –Overview and Comparison of plagiarism Detection Tools, in Dateso 2011, pp. 161-172, ISBN 978-80-248-2391-1
[5]. WEBER WULFF et al – "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.
[6]. LANCASTER, T. et al - Effective and Efficient Plagiarism Detection in PhD thesis, School of Computing, Information System and Mathematics South Bank University in 2003.
[7]. Barn Baum, C. - Plagiarism: A Student's Guide to Recognizing It and Avoiding It, http://www.valdosta.edu/cbarnbau/personal/teaching_MISC/plagiarism.html (Accessed 23 January 2006).
[8]. Hermann Maurer, Frank Kappe, Bilal Zaka – Plagiarism - A survey in Journal of Universal Computer Science, vol. 12, no. 8, 1050-1084, Aug. 2006.
[9]. Tracey Bretag and Saadia Mahmud - Self-plagiarism or Appropriate Textual Re-use in Journal of Academics Ethics vol.7:193-205, 2009.
[10]. Benno Stein et al – "Intrinsic plagiarism detection using character n-gram profiles" in: Proceedings of 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse, pages (38–46), Donostia, Spain, 2009.
[11]. Sven Meyer zu Eissen and Benno Stein – Intrinsic plagiarism detection in n Information Retrieval Proceedings of the 28th European Conference on IR Research, ECIR 2006 London, ISBN 3-540-33347-9, pp. 565-569, c Springer 2006.
[12]. B Stein, M Koppel, and E Stamatatos - Plagiarism analysis, authorship identification, and near-duplicate detection in; Special Interest Group of Information Retrieval, 41(2):68–71, 2007.
[13]. Martin Potthast et al – Overview of the 1st International competition on plagiarism detection in Proceedings of 3rd PAN Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, pages 1– 9, Donostia, Spain, 2009.
[14]. C. Grozea et al: "Pairwise sequence matching in linear time applied to plagiarism detection", in: 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse, 2009, p. 10.
[15]. Asim M. El Tahir Ali et al - Overview and Comparison of Plagiarism.
[16]. Paul Clough - Old and new challenges in automatic plagiarism detection in Plagiarism Advisory Service, vol. 10, Department of Information studies, University of Sheffield, 2003.
[17]. Garfield e. Citation indexes for science: A new Dimension in documentation through association of ideas sciences 122, 3159, 108-111, July 1955.
[18]. S. M. Zu Essen, B. Stein, M. Kulig – "Plagiarism detection without reference collections", in: Advances in data analysis, Springer, 2007, pp. 359--366.
[19]. M. Elhadi, A. Al-Tobi - "Use of text syntactical structures in detection of document duplicates", in: Digital Information Management, ICDIM 2008. Third International Conference on, IEEE, 2008, pp. 520-525.
[20]. J.P. Bao et al - A survey on natural language text copy detection in Journal of software 14 (10) (2003) 1753-1760.
[21]. Joshi, M., & Khanna, K. (2013) - Plagiarism detection over the web: A review in International Journal of Computer Applications, 68(15).
[22]. Daniele Anzelmi et al -- Plagiarism Detection Based on SCAM Algorithm in IMECS 2011 IEEE.
[23]. Raj Kumar Kundu and Karthik. K – "Contextual plagiarism detection using latent semantic analysis", International Research Journal of Advanced Engineering and Science, Volume 2, Issue 1, pp. 214-217, 2017.
[24]. Alireza Talebpour, Mohammad Shirzadi - Plagiarism Detection Based on a Novel Trie-tree based data structure.
[25]. S.N. Autade et al--- EMAS Framework for Text Plagiarism Detection, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 8 (2017) pp. 1584-1590
[26]. Jingling Zhao, Kunfeng Xia, Yilun Fu, Baojiang Cui- "An AST-Based Code Plagiarism Detection Algorithm", 10th International Conference on Broadband and Wireless Computing, Communication and Applications,2015.
[27]. Mayank Agrawal, Dilip Kumar Sharma, 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016
[28]. Agung Sediyono et al- "Algorithm of the Longest Commonly Consecutive Word for Plagiarism Detection in Text Based Document", 978-1-4244-2917-2/08/$25.00 ©2008 IEEE
[29]. Michal Ďuračíka et al- Current trends in source code analysis, plagiarism detection and issues of analysis big datasets in TRANSCOM 2017: International scientific conference on sustainable, modern and safe transport Procedia Engineering 192 (2017) 136 – 141.