ISSN: 0067-2904

Plagiarism Detection Methods and Tools: An Overview

Farah Khaled, Mohammed Sabbih H. Al-Tamimi
Department of Computer Since, Collage of Since, University of Baghdad, Baghdad, Iraq

Received: 22/8/2020 Accepted: 1/11/2020

Abstract

Plagiarism Detection Systems play an important role in revealing instances of a plagiarism act, especially in the educational sector with scientific documents and papers. The idea of plagiarism is that when any content is copied without permission or citation from the author. To detect such activities, it is necessary to have extensive information about plagiarism forms and classes. Thanks to the developed tools and methods it is possible to reveal many types of plagiarism. The development of the Information and Communication Technologies (ICT) and the availability of the online scientific documents lead to the ease of access to these documents. With the availability of many software text editors, plagiarism detections becomes a critical issue. A large number of scientific papers have already investigated in plagiarism detection, and common types of plagiarism detection datasets are being used for recognition systems, WordNet and PAN Datasets have been used since 2009. The researchers have defined the operation of verbatim plagiarism detection as a simple type of copy and paste. Then they have shed the lights on intelligent plagiarism where this process became more difficult to reveal because it may include manipulation of original text, adoption of other researchers' ideas, and translation to other languages, which will be more challenging to handle. Other researchers have expressed that the ways of plagiarism may overshadow the scientific text by replacing, removing, or inserting words, along with shuffling or modifying the original papers. This paper gives an overall definition of plagiarism and works through different papers for the most known types of plagiarism methods and tools.

Keywords: Information and communication technologies, intelligent plagiarism, plagiarism detection, scientific papers, verbatim plagiarism

دراسة بحثية: طرق وأدوات كشف الاستلال العلمي

فرح خالد ، محمد صبيح *
قسم علوم الحاسوب ، كلية العلوم ، جامعة بغداد ، بغداد ، العراق

الخلاصة:

تم تطوير أنظمة الكشف عن الانتحال لتحديد حالات الانتحال في الأوراق العلمية من خلال الكشف على نطاق واسع عن أشكال الانتحال. تحدث فكرة اكتشاف الانتحال عندما يتم نسخ محتوى أي مستند دون إذن أو اقتباس. للكشف عن السرقة الأدبية من أي نوع ، من الضروري أن يكون هنالك معرفة واسعة بأشكالها وفئاتها المحتملة. بفضل وجود أدوات وطرق مختلفة ، من الممكن الكشف عن العديد من أنواع الانتحال. ان تطور تقنيات المعلومات والاتصالات (ICT) وتوافر الوثائق العلمية عبر الإنترنت ادى إلى سهولة الوصول إلى هذه

*Email: m_altamimi75@yahoo.com

الوثائق ، مع توفر العديد من محرري النصوص البرمجية ، تصبح عمليات اكتشاف الانتحال قضية مهمة. هناك عدد كبير من الأوراق العلمية التي تم التحقيق فيها بالفعل في الكشف عن السرقة الأدبية ، والأنواع الشائعة من مجموعات بيانات الكشف عن السرقة الأدبية المستخدمة لأنظمة التعرف ، وقد تم استخدام عدة بيانات لتدريب انظمة كشف السرقة العلمية مثل WordNet و PAN Dataset منذ عام 2009. وقد عرّف الباحثون عملية الكشف الحرفي للسرقة الأدبية كنوع بسيط من نسخ ولصق النصوص ، ثم قاموا بإلقاء الضوء على الانتحال الذكي حيث أصبح الكشف عن هذه العملية أكثر صعوبة لأنه قد يتضمن التلاعب بالنص الأصلي. أعرب باحثون آخرون عن أن طرق الانتحال قد تلقي بظلالها على النص العلمي من خلال استبدال الكلمات أو إزالة أو إدخال أو خلط الأوراق الأصلية أو تعديلها. يقدم هذا البحث تعريفًا شاملاً للسرقة الأدبية من خلال عرض أكثر أنواع أساليب وأدوات الانتحال شهرة.

1. Introduction

Due to the rapid advancement of the computer and network technologies, such as the Internet that enables anyone to access online contents anytime and from anywhere, academic integrity in the academic community is becoming a highly sensitive issue, especially among universities and research institutions. Plagiarism, on the other hand, is defined as a kind of academic dishonest behavior that will damage academic integrity [1]. Thus, it is needed to be resisted determinedly. However, plagiarism is not only an academic issue, but it extends to almost all industries. Occasionally, plagiarism occurs accidentally but most of the time it is the outcome of a conscious process [2]. The best definition of plagiarism might be that it is "the unacknowledged copying of documents or programs" [3]. To overcome the problem of plagiarism, large number researchers have worked on detecting plagiarism since the past decades through software detection methods [4, 5]. Plagiarism was originally detected manually (by hand) or by resembling previously consulted content. Today, the great number of the available online documents make it harder to detect plagiarism manually. Therefore, there is an urgent need to produce automatic plagiarism detectors [5]. There are two main types of plagiarism, namely the verbatim/literal and the intelligent plagiarism. Plagiarism detection methods are also classified into the internal detection method, where the document is analyzed for plagiarism alone, and the external detection method, where detection is made among a collection of documents. Verbatim/literal plagiarism describes the plagiarized content as the exact copying of the source content without altering or modifying the original content. While, in intelligent plagiarism, the main content is altered/modified by different ways. Intelligent plagiarism is more difficult to reveal and includes adoption of the ideas, translation to another language, and manipulations [6, 7].

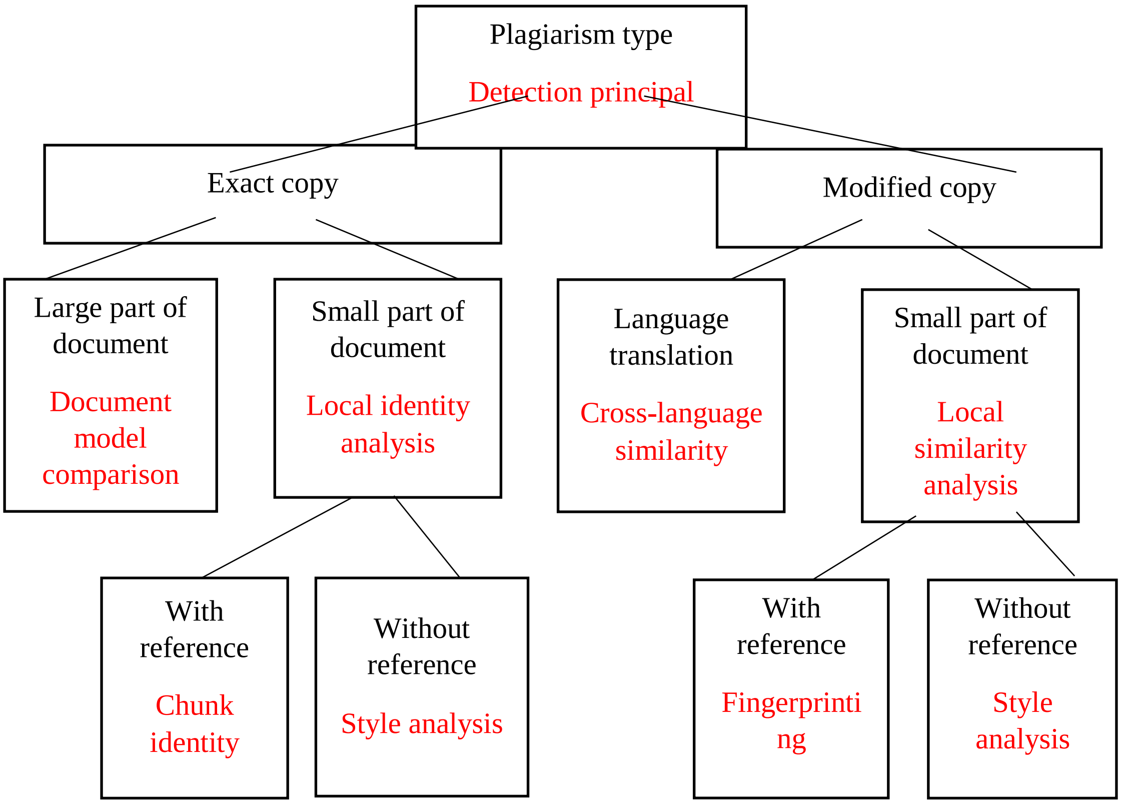

Figure 1- Plagiarism types with some related detection principles [6,8]

This overview paper sheds the light on the description of plagiarism. In section two, plagiarism process will be reviewed, in section three, plagiarism classification and methods will be explained in details, in section four, plagiarism tools will be reviewed, in section five, the types of datasets used in plagiarism detection will be illustrated, in section six, a discussion about the reviewed works will be summarized, and finally in section seven, a conclusion will summarize the topic of plagiarism.

2. Plagiarism Process

To design and produce robust and no error Plagiarism Detection Process, four main stages are required. These stages described below [8, 9]:

1. Collecting the content: It is the first stage, where the plagiarism detector collects the required content from the users through a search engine which acts as the interface between the users and the detector.

2. Analyzing the similarities: After collecting the content (scientific papers, assignments, and other softcopies), the detector runs an analyzing method to search for the similarities among the documents and reveal the original copy.

3. Confirming the copy: After the analyzing stage, a function for plagiarism conformation is required to reveal the plagiarized text from an original one. Sometimes, a degree of the plagiarized text is confirmed with this process.

4. Investigation: It is the final stage, which depends highly on the interference made by the user whenever a plagiarism is confirmed. It also relies on the expert of the user to distinguish between the really plagiarized documents and the cited ones.

3. Plagiarism Classification

Plagiarism can be divided into two basic categories, which are the monolingual and the cross-lingual. Monolingual plagiarism works with most detectors. It is about homogeneous languages, as in the case of English language setting-English language setting. Cross-Lingual plagiarism works with heterogeneous languages; for example, English language setting-Chinese language setting, and this type is quiet rare [5] [10] [11]. In the next section plagiarism, types will be discussed in details.

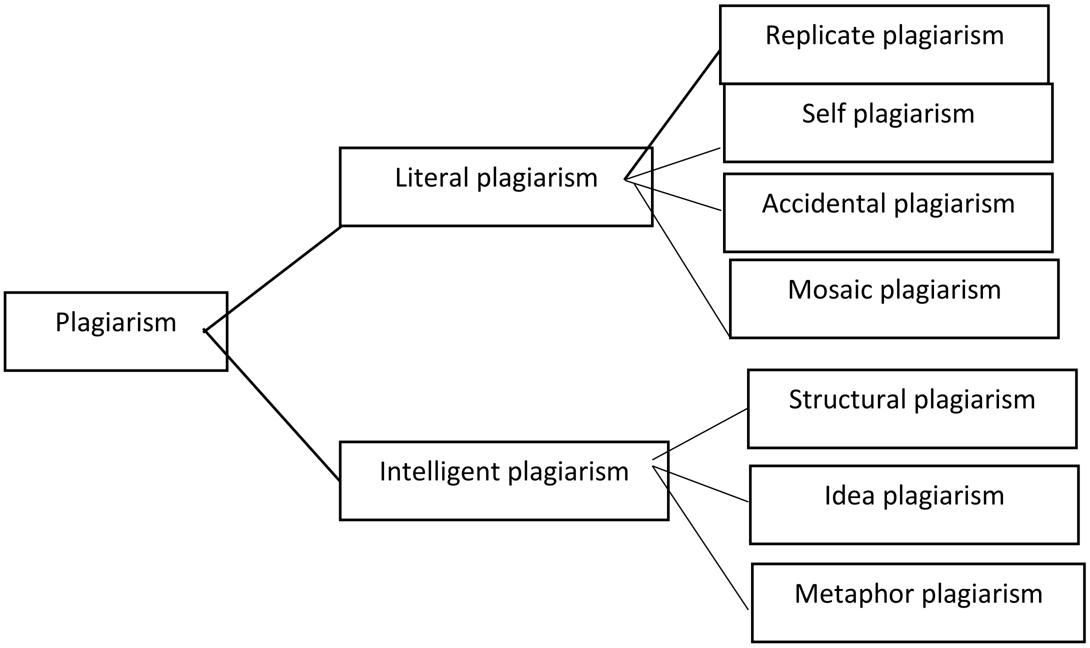

Figure 2-Taxonomy of plagiarism.[6,8]

3.1 Plagiarism Types

Plagiarism types appear in different works, documents, scientific papers, and research article. It can be classified as in the following ways [12]: (i) pretension of others work as your work, (ii) copying others' work without mentioning the credit or citation, (iii) whether citation was mentioned or not, calming someone’s contribution as your own, (iv) refereeing to others work as yours by reconstructing their work, and (v) adding a misleading acknowledgments of others as your work. Textual plagiarism and Source Code plagiarism are the two main types of plagiarism and they will be reviewed in the following [13, 14]

3.1.1 Textual Plagiarism

In researches and scientific fields, this type of plagiarism is the most common one, where the entire text or document is taken without referring to the author or mentioning a quotation. This type of plagiarism can be further divided into seven sub-classes, as in the following [13, 14]:

1. Copy-paste plagiarism: This process refers to copying the original text without any acknowledgment about the authors or the original paper as if it was your work.
2. Paraphrasing plagiarism: It is classified into two categories: (i) simple paraphrasing, where the original text is presented into different way be replacing the words into similar ones with the same meaning and, (ii) Mosaic/Hybrid/Patchwork paraphrasing, where the text is a result of combining different contributions from different papers and presented differently without referring to the original citations of the works.
3. Metaphor plagiarism: presenting other ideas in better ways.
4. Idea plagiarism: the entire solution and ideas are stolen from others and claiming that it is an original research paper.
5. Recycled plagiarism: The authors here use their previous/old works and papers for a new publication.
6. 404 Error / Illegitimate Source plagiarism: when the citation of the works is invalid.
7. Re-tweet plagiarism: In this type, the citation is referred to but it is no difference between the original work and the author's work from the point of structure, grammar, and words.

3.1.2 Source-Code Plagiarism

This type appears typically in educational fields, where the programming code of a specific program written originally by someone is copied, adjusted, or reused by others partially or completely. It has the following four sub-classes [13, 14]:

1. Manipulation plagiarism: where the source code is altered or modified by other developers by either deleting or inserting sub-codes to an original one without referring to the citation or acknowledgment.
2. Reordering structure plagiarism: where the syntax of source code is modified by functions or statements recording without referring to the original work.
3. No-change plagiarism: where the developers do not change anything in the code but add/remove spaces or comments as it was their work.
4. Language switching plagiarism: where the source code language is rewritten by other languages and declared as original code.

3.2 Plagiarism Detection

Many papers have searched for highly accurate plagiarism detection methods using different tools, but it was always challenging to find the perfect one, due to the rapid development of the technologies, software, and data mining tools. This development has become a double-edged weapon, as the methods of plagiarism have evolved; on the other hand, the methods of detecting this theft have developed in a response to the curbing of illegal methods of copying the original work of researchers [10, 11]

Plagiarism detection can be performed manually or by using an automated process. The automated process is very similar to natural language processing, visual identification, and bio-metric processes. All of these have a foundation of pattern recognition. Automated process does not give 100% accuracy. Thus, the manual checking is still needed.

Internal Plagiarism Detection

Thus type involves finding plagiarized passages within a document without access to the potential original text, also called intrinsic plagiarism detection.

External Plagiarism Detection

External plagiarism detection involves comparing suspicious plagiarized documents against potential original documents.

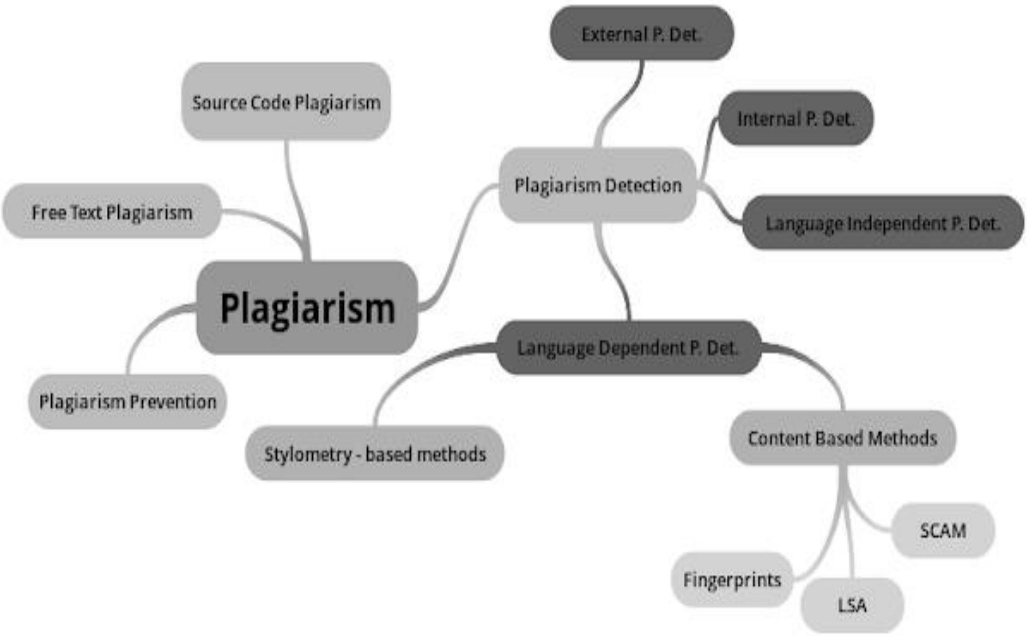

Figure 3- Mind map for Plagiarism [10,15]

In the next sections, different methods and tools for plagiarism detection will be illustrated.

3.3 Plagiarism Detection Methods

Many methods have been implemented by researchers to overcome plagiarism as it has grown to form a serious issue among the academic community; researchers have used different methods to overcome these activities [5]. Therefore, a comparative study depending on the attached sources about plagiarism detection, as viewed by researchers, is illustrated in Table-1.

Table 1- A Comparative Study [15]

Author	Approach	Illustration	Limitation
Rajkumar Kundu, Kartik. K [16]	Latent Semantic Analysis LSA	Used to find out the semantic SVD and reduction to capture all similar text	A distributed model that is not suitable for nonlinear equations
Alireza Talebpour et al [17]	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 that requires processing of content which is not an efficient solution for large number of files
S.N. Autade et.al [18]	Evolutionary multi-agent system	System synonym recognition and word -generalization is used	A large word dictionary update is required
Jingling Zhao et. Al [19]	An AST-based Code Plagiarism Detection Algorithm	They proposed AST- CC algorithm to generate and compare hash values	Less efficiency for the storage of data structure
Mayank Agrawal et al [20]	A State of Art on Source Code Plagiarism	Various methods like NLP and machine learning	It is hard to locate the plagiarism among different source codes of different languages
Agung Sediyono et al [21]	Longest common consecutive word	A numerical based comparison algorithm that outsources suffix tree algorithm	The drawback of this proposed algorithm is loading time.
Michal Ďuračíka et. al [22]	Detection of clones and methods for determining similarity	Provides anti- plagiarism system which is to handle large size of dataset	Needs to process multiple documents with similar identity which exhibits high execution

It is important to mention that almost all the types, described in details below, try to find a similarity between an existing document and a query document [5]:

1. Character-Based Methods: Based on either n-gram or word n-gram methods, it works on the string-matching techniqueI. It is the most used and well-known method used by many researchers to reveal the degree of matching/mismatching among different strings [23, 24].
2. Vector-Based Method: It implements lexical and syntax features as tokens rather than strings [25].
3. Syntax-Based Method: This method implements parts of speech (POS), such as verbs, nouns, pronouns, adjectives, adverbs, prepositions, conjunctions, and interjections, of a set of phrases in a text document to detect plagiarism action [26].
4. Semantic-Based Method: The semantic similarity is detected in this method by comparing the similarity between two words in a text, such as transforming from active voice to passive voice with the same semantics of two different sentences [27, 28] .
5. Fuzzy-Based Method: Machine learning is implemented in this type. The sentences are presented as a numerical or character values to detect plagiarism. The output is either 0 or 1, where zero means that the documents are completely different (no plagiarism) and one indicates that there is a matching between the documents (plagiarism is found) [29, 30, 31].
6. Structure-Based Method: The search for how the words are written in a specific block of text in a document [32] [33].

7. Stylometric-Based Method: To detect plagiarism, this type aims for finding the writing styles of the author. The similarities are detected between two blocks of sentences based on the stylometric features of the writer [34, 35].

8. Methods for Cross-Lingual Plagiarism Detection: It is quite challenging to reveal a plagiarism act because it has many types of plagiarism, like cross-lingual syntax-based methods and/or dictionary-based methods. It requires extensive in-depth knowledge for multiple languages in more than one document [36, 37].

9. Grammar Semantics Hybrid Plagiarism Detection Method: A very effective and extensively implanted approach in the field of natural language processing (NLP) . It is very accurate in revealing copy-paste plagiarism or paraphrasing plagiarism. It provides a remedy for the limitations of the semantic-based method [38].

10.Classification and Cluster-Based Methods: These are greatly helpful methods to retrieve the information during the process of searching in any plagiarized document. Also, the comparison time is reduced during the detection process when comparing these methods with other ones [39].

11.Citation-Based Method: This is a novel method; it mainly belongs to semantic plagiarism detection methods for the usage of the semantics in the cited document. It looks for identical pair of documents based on the citation, because these techniques use semantics contained in the citation [40, 41].

Table 2- Plagiarism detection Methods Based on features of type, class, and mode [5]

#	Method	Type		Class		Mode	
		Internal	External	Mono-Lingual	Cross-Lingual	Literal	Intelligent
1	Character-Based	☒	✓	✓	☒	✓	☒
2	Vector-Based	☒	✓	✓	☒	✓	☒
3	Syntax-Based	☒	✓	✓	☒	✓	☒
4	Semantic-Based	☒	✓	✓	☒	✓	✓
5	Fuzzy-Based	☒	✓	✓	☒	✓	✓
6	Structural-Based	☒	✓	✓	☒	✓	☒
7	Stylometric-Based	✓	☒	✓	☒	✓	☒
8	Cross-Lingual	☒	✓	☒	✓		✓
9	Grammar-Based	☒	✓	✓		✓	☒
10	Classification & Cluster-Based	☒	✓	☒	☒	✓	☒
11	Citation-Based	☒	✓	✓	☒	✓	✓

4. Plagiarism Detection Tools

A large number of tools have been developed and utilized to detect plagiarism [5]. Table 3 shows plagiarism detection tools according to their pros and cons, covering a period of 22 years from 1994 to 2020 [5].

Table 3- The Most Known Plagiarism Tools [5].

Tool	Year	Characteristics
MOSS [42]	1994	MOSS (Measure of Software Similarity) detects source-code plagiarism; it takes parts of the code at a time as an input and produces HTML pages as an output to analyze the similarities between a pair of documents.
Ithenticate [43]	1996	It is a text-document based plagiarism detection tool that is presented as a web page. It compares a number of documents with the original one without the need for installation on the end-user computer, but it is limited to 25,000 words per time.
JPlag[32]	1997	Similar to the previous ones, this type is an online source-code plagiarism tool. It takes a number of programming codes and selects the identical lines among them. It works with C, C++, and Java programming languages, with less than one minute to detect hundreds of

Tool	Year	Characteristics
		code lines.
GPSP - Glatt Plagiarism Screening Program [44]	1999	Unlike the previous tools, it works off-line. It mixes different approaches and finds the similarities among the writing styles of differed authors. It reveals plagiarism by making the author goes through a fill-in-the-blank test. Then, it counts the correctly filled blanks and the time taken to finish the test. Finally, according to the results, it takes a decision about an act of plagiarism.
Turnitin [42][37]	2000	It is provided by iParadigms as a web based tool. The user is required to upload his/her required document online, then the document will be saved to the system's database. After that, the tool checks for plagiarism by creating a document fingerprint. It accepts nearly 15,000 institutions around the world, with more than 30 million users, for its flexibility and robustness. Therefore, it is considered as the best tool.
Plagiarism Checker [45]	2006	This is a free and online tool, using search engine services to detect for students' plagiarism by checking if their documents have a similar copy of another online document.
Plagiarism Scanner [46]	2008	It is an effective tool that detects throughout almost all online resources, like libraries (Questia and ProQuest), online databases, websites, and search engines. When plagiarism is detected, it produces a full report including the rate, originality, and percentage of plagiarized materials.
PlagTracker [47]	2011	It is a well-known tool for all kinds of users (teachers, websites owners, and students)that accommodates a large number of academic resources in its database and produces a detailed report whenever a plagiarism is detected.
PlagScan [48]	2015	This tool provides multiple services to companies, universities, and schools, but it is not free and the users must have a paid account to register to this tool.
Exactus Like [49]	2016	This tool is a web-based online tool that works with different formats, like HTML, Microsoft Word, and Adobe PDFs. It detects moderately disguised borrowing (word/phrase reordering, substitution of some words with synonyms) by a deep parsing function.
Grammerly [50]	2016	This is a website and a mobile application service that offers a great opportunity to the individuals to correct their documents within a real-time manner and a friendly user interface. It works online; therefore it requires an internet connection.
Grammerly [50]	2018	It is an evolved version of the previous one, representing the premium type. It targets business industries, such as teams and companies. Users reported that Grammarly helps them more professionally.
DupliChecker [51,52]	2020	This is an absolutely perfect method, available 24/7, and ready whenever the user needs it. It is one of the most effective and free plagiarism tools on the internet. The user only requires a search engine and a connection to the world wide web to access this tool. It enables the user to either copy-paste or upload the document to check for a plagiarism.

5. Datasets Used in Plagiarism Detection Systems

Two main datasets are used for plagiarism detection and are illustrated below.

5.1 WordNet Dataset

WordNet is a freely and publicly available dataset that contains large lexical English language words, such as nouns, verbs, adverbs, and adjectives. It contains over 155,287 words organized in 117,659 synsets. All these words are classified into groups of cognitive synonyms. WordNet not only links word forms (strings) but also specifies their meanings. As a result, words that are next to each other in the dataset are semantically disambiguated. Also, WordNet marks the semantic relationships among groups of words in the thesaurus that do not follow any clear pattern, except for similar meanings. Lexical-words that are represented by this dataset contain synonymy between words, like the words "large" and "wide", both having a relatively similar meaning. The phrase of a noun contains substitution definitions that contain one or more substitutions. Therefore, each formal meaning pair in WordNet is unique. The current version of this dataset contains not only English language words but also different languages, such as Italian and Spanish [53].

5.2 Plagiarism analysis, Authorship identification, and Near-duplicate detection (PAN)

Another well-known plagiarism detection dataset is PAN. It refers to plagiarism analysis, authorship identification, and near-duplicate detection of different types of plagiarism. It is a series of scientific events and shared tasks on digital text forensics and stylometry. Every year, an international conference and competition called PAN@CLEF are held to connect the most advanced publications about plagiarism detection techniques [54].

6. Discussion

In this part, an extensive review of the most serious and frequently used plagiarism techniques around the world is illustrated. Also, the most common challenges that are facing the development of effective and robust plagiarism detection systems are reviewed.

6.1 Comparison of Common Plagiarism Types

The detection systems are available in-hand and they are growing rapidly. Therefore, a comparison among the nowadays plagiarism techniques is required. This comparison will focus on two studies that show the most serious plagiarism techniques in the universities, schools, and higher education sector. This comparison is provided by conducting several scientific areas, such as those of medicine , natural sciences, engineering, and social sciences, from 40 different countries around the world, covering the period from 2013 through 2015. Tables 4 and 5 list the most commonly practiced plagiarism types [55] [56].

Table 4-The Ranks of Most Common Plagiarism Types [55]

Type	Rank of use
Paraphrasing	75%
Repetitive research	71%
Secondary source	69%
Duplication	63%
Verbatim	59%

Table 5-Plagiarism Types with their Rank and Degree of Seriousness [56]

Type	Rank of use	Problematic
Clone	9.5	9.5
Remix	5.6	0.5
Recycle	5.5	2.8
Retweet	4.4	0.5
404 Error	0.6	1.3
Find-Replace	3.9	1.2
Hybrid	0.5	1.1
Aggregator	2.8	2.9

6.2 Challenging Factors Facing Plagiarism Detection

Among the previously published papers that have been focusing on plagiarism, many works made a dense search on plagiarism types and techniques. Most of the plagiarism detectors available today can do the following: (i) distinguish between plagiarism in source code and/or in-text documents, with or without citation, (ii) feature extraction of semantic and/or salient syntactic, and (iii) plagiarism detection for both cross-lingual and monolingual documents [5]. Despite the availability and efficiency of these types, they are not effective enough to reveal the unattended research challenges. With the technology age that we are living in, new algorithms are produced to solve many problems, such as plagiarism, which is a problematic issue that needs to be extensively solved, especially in the scientific community. Computer science approach can address these challenges relying on the ICT advancement, some of these challenges are highlighted in the following: (i) A proof for correctness and completeness of the scientific works, i.e. whether they are ate in text documents or written as source code, is not available yet, (ii) a highly accurate framework for plagiarism detection that can reveal text segment(s), for both intrinsic and extrinsic plagiarism detection, is missing, (iii) The development of pilgrim checking systems without the need for external references and with high accuracy is considered a very challenging task, and (iv) Providing a full system for scientific works repository that can combine the works of all authors and the references to their works in one place is a difficult manner [5].

7. Conclusions

In this paper, an extensive literature survey about plagiarism types, methods, classification, and tools was conducted. Text plagiarism, with its seven sub-types, and source-code plagiarism, with its four sub-types were highlighted. Then, plagiarism detection methods were illustrated and summarized over nearly more than twenty years. The newly developed tools are more advanced, most of which are working online using an internet connection and a web page, some of them are delivered freely and others require subscription payment. Next, the most known datasets implemented in plagiarism detection were reviewed and a table was prepared to discuss the methods to be adopted in plagiarism . We notice that each method has its strengths and weaknesses that depend on how it is described to support two important factors: time and accuracy. We also notice that there are two mechanisms of action, namely the parallel and the series mechanisms. The parallel mechanism provides higher accuracy and less time because it performs a scanning for all the contents of a dataset. For this reason, we can say that a good algorithm is the one that covers the required conditions in terms of time, accuracy, or both.

Finally, a discussion about the most frequent plagiarism types was extensively provided and the most challenging steps during the implementation of plagiarism detection were investigated.

References

1. K. Vani and D. Gupta. **2016**. "Study on Extrinsic Text Plagiarism Detection Techniques And Tools," *J. Eng. Sci. Technol. Rev.*, **9**(5): 9–23, 2016, doi: 10.25103/jestr.095.02.

2. A. Ekbal, S. Saha, and G. Choudhary. **2012**."Plagiarism Detection in Text Using Vector Space Model," *Proc. 2012 12th Int. Conf. Hybrid Intell. Syst. HIS 2012*, pp. 366–371, 2012, doi: 10.1109/HIS.2012.6421362.

3. A. Sharma, V. Walia, and M. Gahlawat. **2016**. "Review : Plagiarism an Act of Unethics," *PharmaTutor Mag.*, **3**(2): 20–23.

4. X. Ruoyun. **2018**. "An Overview of Plagiarism Recognition Techniques," *Int. J. Knowl. www.ijklp.org Lang. Process. KLP Int.*, **9**(2): 1–19, 2018, [Online]. Available: www.ijklp.org.

5. H. A. Chowdhury and D. K. Bhattacharyya. **2018**. "Plagiarism: Taxonomy, tools and detection techniques," *arXiv*, **9**: 1–15.

6. A. Hamza Osman, N. Salim, and A. Abuobieda. **2012**. "Survey of Text Plagiarism Detection," *Comput. Eng. Appl. J.*, **1**(1): 37–45. doi: 10.18495/comengapp.v1i1.5.

7. R. Lukashenko, V. Graudina, and J. Grundspenkis. **2014**. "Computer-Based Plagiarism Detection Methods and Tools: An overview," *ACM Int. Conf. Proceeding Ser.*, **285**: 1–5, 2007, doi: 10.1145/1330598.1330642.

8. A. Hamza Osman, N. Salim, and A. Abuobieda. **2012**. "Survey of Text Plagiarism Detection," *Comput. Eng. Appl. J.*, **1**(1): 37–45. doi: 10.18495/comengapp.v1i1.5.

9. F. Culwin and T. Lancaster. **2001**. "Plagiarism Issues for Higher Education," *Vine*, **31**(2): 36–41. doi: 10.1108/03055720010804005.

10. C. Kustanto and I. Liem. **2009**. "Automatic Source Code Plagiarism Detection," *10th ACIS Conf. Softw. Eng. Artif. Intell. Netw. Parallel/Distributed Comput. SNPD 2009, conjunction with IWEA*

2009 WEACR 2009, pp. 481–486, 2009, doi: 10.1109/SNPD.2009.62.

11. Y. Shen, S. C. Li, C. G. Tian, and M. Cheng. **2009**. “Research on Anti-plagiarism System and The Law of Plagiarism,” *Proc. 1st Int. Work. Educ. Technol. Comput. Sci. ETCS 2009,* **2**: 296–300. doi: 10.1109/ETCS.2009.327.

12. M. S. Anderson and N. H. Steneck. **2010**. “The Problem of Plagiarism,” *Urol. Oncol. Semin. Orig. Investig.,* **29**(1): 90–94. doi: 10.1016/j.urolonc.2010.09.013.

13. N. Charya, K. Doshi, S. Bawkar, and R. Shankarmani. **2015**.“Intrinsic Plagiarism Detection in Digital Data,” *Ijiere.Com,* **2**(3): 23–30. [Online]. Available: http://ijiere.com/FinalPaper /FinalPaper 2015320223857549.pdf. a

14. C. barnbaum. **2009**. plagiarism: a student's guide to recognizing it and avoiding it.[online].[cit. 2010-12-14] (2009).

15. N. Khan, C. Agrawal, and T. Nishat Ansari. **2018**.“A Review on Various Plagiarism Detection Systems Based on Exterior and Interior Method,” *Ijarcce,* **7**(9): 6–12. doi: 10.17148/ijarcce.2018.792.

16. E. S. Al-Shamery and H. Q. Gheni. **2016**.“Plagiarism detection using semantic analysis,” *Indian J. Sci. Technol.,* **9**(1): 1–8. doi: 10.17485/ijst/2016/v9i1/84235. alireza talebpour, mohammad shirzadi - plagiarism detection based on a novel trie-tree based data structure.

17. S.N. Autade et al. **2017**. emas framework for text plagiarism detection, *International Journal Of Applied Engineering Research,* **12**(8): 1584-1590

18. J. Zhao, K. Xia, Y. Fu, and B. Cui. **2015**. “An AST-based Code Plagiarism Detection Algorithm,” *Proc. - 2015 10th Int. Conf. Broadband Wirel. Comput. Commun. Appl. BWCCA 2015,* **3**: 178–182. doi: 10.1109/BWCCA.2015.52.

19. M. Agrawal and D. K. Sharma. **2020**. “A State of Art on Source Code Plagiarism Detection,” *Proc. 2016 2nd Int. Conf. Next Gener. Comput. Technol. NGCT 2016,* no. May 2020, pp. 236–241, 2017, doi: 10.1109/NGCT.2016.7877421.

20. A. Sediyono, K. Ruhana, and K. Mahamud. **2008**. “Algorithm of the longest commonly consecutive word for plagiarism detection in text based document,” *3rd Int. Conf. Digit. Inf. Manag. ICDIM 2008,* **3**: 253–259. doi: 10.1109/ICDIM.2008.4746827.

21. M. Ďuračík, E. Kršák, and P. Hrkút. **2017**. “Current Trends in Source Code Analysis, Plagiarism Detection and Issues of Analysis Big Datasets,” *Procedia Eng.,* **192**: 136–141. doi: 10.1016/j.proeng.2017.06.024. c. grozea, c. gehl, m. popescu, encoplot: pairwise sequence matching in linear time applied to plagiarism detection, in: 3rd pan workshop. uncovering plagiarism, authorship and social software misuse, 2009, p. 10.

22. C. Basile, D. Benedetto, E. Caglioti, G. Cristadoro, and M. D. Esposti. **2011**. “A plagiarism Detection Procedure In Three Steps: Selection, Matches And Squares,” *CEUR Workshop Proc.*, vol. 502, pp. 19–23, 2009. h. zhang, t. w. chow, a coarse-to- ne framework to efficiently thwart plagiarism, pattern recognition 44 (2) (2011) 471-487.

23. M. Elhadi and A. Al-Tobi. **2008**. “Use of Text Syntactical Structures In Detection Of Document Duplicates,” *3rd Int. Conf. Digit. Inf. Manag. ICDIM 2008*, **4**: 520–525. doi: 10.1109/ICDIM. 2008.4746719.

24. S. Torres and A. Gelbukh. **2009**. “Comparing Similarity Measures for Original WSD Lesk Algorithm,” *Adv. Comput. Sci. Appl.*, **43**: 155–166.

25. P. Resnik. **1999**. “Semantic Similarity In a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language,” *J. Artif. Intell. Res.,* **11**: 95–130. doi: 10.1613/jair.514.

26. C. Leacock, G. A. Miller, and M. Chodorow. **1998**. “Using Corpus Statistics and WordNet Relations For Sense Identification,” *Comput. Linguist.,* **24**(1): 146–165.

27. R. Yerra and Y. K. Ng. **2006**. “A Sentence-Based Copy Detection Approach For Web Documents,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics),* **3613** : 557–570, 2006, doi: 10.1007/11539506_70.

28. J. Koberstein and Y. K. Ng. **2006**. “Using word Clusters To Detect Similar Web Documents,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, **4092**: 215–228, 2006, doi: 10.1007/11811220_19.

29. S. M. Alzahrani and N. Salim. **2009**. “On The Use of Fuzzy Information Retrieval for Gauging Similarity of Arabic Documents,” *2nd Int. Conf. Appl. Digit. Inf. Web Technol. ICADIWT 2009,*

3: 539–544, 2009, doi: 10.1109/ICADIWT.2009.5273835. m. rahman, w. p. yang, t. w. chow, s. wu, a exible multi-layer self-organizing map for generic processing of tree-structured data, pattern recognition 40 (5) (2007) 1406--1424.

30. T. W. S. Chow and M. K. M. Rahman. **2009**. "Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection," *IEEE Trans. Neural Networks*, **20**(9): 1385–1402, 2009, doi: 10.1109/TNN.2009.2023394.

31. S. Meyer, B. Stein, and M. Kulig. **1995**. "Plagiarism Detection Without Reference Collections Sven," *Springer, Berlin, Heidelb.*, **9**: 360–366.

32. X. Liu. **2013**. "A Survey of Modern Authorship Attribution Methods Efstathios," *J. Am. Soc. Inf. Sci. Technol.*, **64**(July): 1852–1863. doi: 10.1002/asi.

33. M. Potthast, A. Barrón-Cedeño, B. Stein, and P. Rosso. **2011**. "Cross-language plagiarism detection," *Lang. Resour. Eval.*, **45**(1): 45–62, 2011, doi: 10.1007/s10579-009-9114-z.

34. A. Hamza Osman, N. Salim, and A. Abuobieda. **2012**. "Survey of Text Plagiarism Detection," *Comput. Eng. Appl. J.*, **1**(1): 37–45, 2012, doi: 10.18495/comengapp.v1i1.5. j.-p. bao, j.-y.

35. N. Language and T. Copy. **2003**. "A survey on Natural Language Text Copy Detection," *sementic Sch.*, **14**(10): 1–8.

36. V. Mitra, C. J. Wang, and S. Banerjee. **2007**."Text classification: A least square support vector machine approach," *Appl. Soft Comput. J.*, **79**(3): 908–914, 2007, doi: 10.1016/j.asoc.2006.04.002.

37. B. Gipp and J. Beel. **2010**. "Citation based plagiarism detection - A new approach to identify plagiarized work language independently," *HT'10 - Proc. 21st ACM Conf. Hypertext Hypermedia*, **9**(January): 273–274, 2010, doi: 10.1145/1810617.1810671.

38 B. Gipp and N. Meuschke. **2011**. "Citation pattern matching algorithms for citation-based plagiarism detection: Greedy citation tiling, citation chunking and longest common citation sequence," *DocEng 2011 - Proc. 2011 ACM Symp. Doc. Eng.*, **3**(January): 249–258. doi: 10.1145/2034691.2034741.

39. R. A. Ahmed. **2002**. "Overview of Different Plagiarism Detection Tools," **2**(10): 2–4, 2015. l. prechelt, g. malpohl, m. philippsen, finding plagiarisms among a set of programs with jplag, j. ucs 8 (11) (2002) 1016.

40. H. A. Maurer, F. Kappe, and H. Maurer. **2017**. "Plagiarism - A Survey . Plagiarism - A Survey," **3**(March): 1–11, 2017.

41. B. Honig. **2012**. "The Fox in The Hen House : A Critical Examination of Plagiarism Among Members of," *Acad. Manag. Learn. Educ.*, **11**(1): 101–123, 2012. r. r. naik, m. b. landge, c. n. mahender, a review on plagiarism detection tools, international journal of computer applications 125 (11).

42. http://www.plagtracker.com/, plagtracker.

43. http://www.plagscan.com/, plagscan.

44. N. Khan, C. Agrawal, and T. Nishat Ansari. **2018**."A Review on Various Plagiarism Detection Systems Based on Exterior and Interior Method," *Ijarcce*, **7**(9): 6–12, 2018, doi: 10.17148/ijarcce.2018.792.

45. N. M. Mohan Kumar P. **2015**. Swapna Priya N1 Musalaiah SVVS, "Knowing and Avoiding Plagiarism During Scientific Writing," *African J. online*, **4**(3): 1–6, 2015.

46. https://www.duplichecker.com/

47. https://elearningindustry.com/top-10-free-plagiarism-detection-tools-for-teachers

48. K. Evanini and X. Wang. **2015**. "Automatic detection of plagiarized spoken responses," **5**: 22–27, 2015, doi: 10.3115/v1/w14-1803.

49. pan: http://pan.webis.de/

50. D. B. Dasari and V. G. R. A. O. K. **2014**. "Understanding Plagiarism For Contextual Features Abstract :," *Int. J. Softw. &Hardware Researvhe Eng.*, **3**(12): 24–27.

51. Phillips,V. **2015**. turnitin report. available at https://www.geteducated.com/elearning-educationblog/10-types-of-plagiarism-and-academic-cheating/.

52. L. Zahrotun. **2016**. "Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method," *Comput. Eng. Appl. J.*, **5**(1): 11–18, 2016, doi: 10.18495/comengapp.v5i1.160.

53. X. Ruoyun. **2018**. "An Overview of Plagiarism Recognition Techniques," *Int. J. Knowl. www.ijklp.org Lang. Process. KLP Int.*, **9**(2): 1–19 [Online]. Available: www.ijklp.org.

54. S. Meyer Zu Eissen and B. Stein. **2006**. "Intrinsic Plagiarism Detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics),* **3936** LNCS: 565–569, 2006, doi: 10.1007/11735106_66.

55. M. Roig and D. Ph. **2006**. "Avoiding Plagiarism , Self-Plagiarism , And Other Questionable Writing Practices : A guide To Ethical Writing," *Off. Res. Integr.*, pp. 1–63, 2006.

56. D. B. Dasari and V. G. R. A. O. K. **2014**. "Understanding Plagiarism For Contextual Features Abstract :," *Int. J. Softw. &Hardware Researvhe Eng.*, **3**(12): 24–27, 2014.

.