International Journal of Knowledge and Language Processing
Volume 9, Number 2, **2018**

www.ijklp.org
KLP International ©2018 ISSN 2191-2734
pp.1—19

An Overview of Plagiarism Recognition Techniques

Xie Ruoyun
School of Software & Microelectronics
Peking University
No.5 Yiheyuan Road Haidian District
Beijing, 100871, China
xieruoyun@pku.edu.cn

Received June 2018; revised July 2018

ABSTRACT. *Academic integrity is an extremely important issue emphasized by academic community, especially universities and research institutions. Plagiarism, as a kind of academic dishonorable behavior, will damage academic integrity to a large extent. To fight against plagiarism, different plagiarism detection systems are developed, some of which can be easily accessed online. By extensively investigating plagiarism detection approaches both at home and abroad, this review provides an overview of studies on plagiarism detection techniques. First it summarizes the basic types and development of copy detection techniques in the early days and then performs in-depth comparisons on the state-of-the-art natural language-based copy detection techniques, especially in the PAN competition (Plagiarism analysis, Authorship identification, Near-duplicate detection) from 2009-2014. Finally, with analysis of some existing plagiarism detection systems, this review gives an outlook of the development of plagiarism detection systems.*

Keywords: Plagiarism detection, Natural Language Processing, PAN

1. **Introduction.** With computer and network technology developing rapidly, newspapers, books, academic papers, etc. are no longer limited to the traditional version of paper, which is difficult to for large-scale search and reading. Nowadays, more and more journals or articles can be easily accessible and editable online. Users can make use of search engines and professional knowledge discovery platforms to find what they are interested in. In China, academic search platforms, such as CNKI Scholar [1], integrate the academic journals, dissertations, books and patents of various countries in the world through methods including copyright cooperation, to provide free retrieval of international bibliographic resources to readers. This trend is conducive to the cross-border sharing of academic

resources. Moreover, it can help researchers and college students to quickly grasp the development of their research field and understand the latest technology. However, it also exacerbates academic plagiarism to some extent.

In academic community, academic integrity is a critical issue being repeated emphasized, especially universities and research institutions. Plagiarism, as a kind of academic dishonorable behavior, will damage academic integrity and need to be resist determinedly. However, plagiarism is still very serious, not only in academia areas, but also in the industry. According to a report *Plagiarism* [2]*: Facts&Stats* published by Turnitin in 2017, "The Josephson Institute Center for Youth Ethics surveyed 43,000 high school students in public and private schools and found that one out of three high school students admitted that they used the Internet to plagiarize an assignment." While in college and graduate school, "A survey of over 63,700 US undergraduate and 9,250 graduate students over the course of three years (2002-2005) --conducted by Donald McCabe, Rutgers University--revealed that 36% of undergraduates admit to paraphrasing or copying few sentences from Internet source without footnoting it; 24% of graduate self-report doing the same." To fight against plagiarism, different copy detection systems are developed, some of which can be easily accessed online.

The research of plagiarism recognition is based on the analysis and processing of digital documents and texts to a large extent [3]. There are mainly two ways to resist plagiarism. One occurs before people want to plagiarize. By setting the document itself to ensure that the protected text is difficult to be copied directly, plagiarism is blocked from the source. However, with the development of technologies such as optical character recognition (OCR), the copying and editing of electronic text has become easier and easier; the second is the plagiarism detection method mentioned in this paper. Different from the first one, this method is always used when plagiarism behavior is conducted. It mainly compares the similarity of the document and sets a certain threshold to identify and confirm the plagiarism document. Plagiarism detection can also be called copy detection or duplicate detection. Like information retrieval, and sentiment analysis, the main task of plagiarism detection also focuses on to calculate the similar information shared between two documents, which is a key research topic in the field of natural language processing (NLP).

At present, plagiarism detection techniques have been widely used in the process of examining academic papers, which presents to be very practical. In foreign countries, online copy detection systems like Turnitin, Dupli Checker, etc. are welcomed, while in China, the check-in system launched by the companies or institutions of some academic databases such as CNKI (China National Knowledge Infrastructure) and VIP (China Science and Technology Journal Database) has occupied the mainstream, which can offset academic misconduct effectively and further realize intellectual property protection.

By extensively investigating plagiarism detection approaches both at home and abroad, this review provides an overview of studies on plagiarism detection techniques. First it summarizes the basic types and development of copy detection techniques in the early days and then performs in-depth comparisons on the state-of-the-art natural language-based copy detection techniques, especially in the PAN competition (Plagiarism analysis, Authorship

identification, Near-duplicate detection) from 2009-2014. Finally, with analysis of some existing plagiarism detection systems, this review gives an outlook of the development of plagiarism detection systems.

2. Basic Types and Development Routes of Plagiarism Recognition Techniques.

2.1. Basic Types of Plagiarism Recognition Techniques.

Judging from the text type of plagiarized detection object, plagiarism detection can be divided into the detection of formal language text and the detection of natural language text [4]. Formal language text usually contains data files, computer program code, etc., with a standardized syntax and a clear chapter organization in most cases. Plagiarism detection for natural language texts includes novels, essays, student assignments, etc. Plagiarism of formal language text is easier to be detected.

From the perspective of plagiarism type, "Copy" not only includes verbatim plagiarism, statement rewriting by simple means such as deletion, word order adjustment, etc., but also includes intelligent plagiarism, such as opinion plagiarism. Intelligent plagiarists try to hide, and change the original work in different clever methods, contain text treatment, conversion, and thought acceptance [5]. Considering the difficulty levels in copy detection, direct or word-by-word copy is the easiest to identify by string matching methods. Generally, direct or verbatim copy is the easiest to identify by string matching. However, the detection of opinion plagiarism often involves semantic analysis. It has always been a difficult point for plagiarism detection tasks to overcome.

From the perspective of the comparison resources used by copy detection system, the technology can generally be divided into two categories [6], namely extrinsic plagiarism detection methods and intrinsic plagiarism detection methods. The extrinsic method compares a suspicious document with a specific set of documents and sets a certain threshold. Any document whose similarity exceeds the threshold is determined to plagiarize, which is conducting similarity computation. Intrinsic plagiarism detection does not compare the suspicious document with the external document set. It only uses the internal comparison of the document to determine whether the document is suspected of plagiarism. For example, whether the author's writing style is consistent in the same article. In general, the study which is not explicitly indicate the type of replication detection refers to extrinsic plagiarism detection methods, and this article also focuses on extrinsic plagiarism detection methods.

2.2. Development Route of Plagiarism Recognition Techniques in the early Days.

The plagiarism detection technique was first applied in recognizing plagiarized programs. As early as 1976, Ottenstein [7] proposed a method to detect software plagiarism. However, plagiarism detection techniques based on natural language documents have lagged for more than two decades. The main reason is that the structured language represented by programming language has formal grammar, clear semantic expression, and more standardized text organization structure. On the contrary, the structure of natural language documents is not clear enough, and natural language itself is also vague and difficult to recognize by the machine. As a result, natural language-based plagiarism detection is also

the difficulty in the development of copy detection technology.

In 1987, Rabin and Karp [8] proposed the string matching algorithm based on pattern matching, also named the Rabin-Karp algorithm. This algorithm used overlapping k-gram method and window sliding-based method of string matching, which laid a good foundation for natural language-based copy detection techniques. Almost fifteen years later, the Winnowing algorithm proposed by Schleimer [9] in 2003 draws on the basic idea of Rabin-Karp, and on this basis, the author adds noise removal and filtering methods. The core of the Winnowing algorithm is to extract the fingerprint with the smallest value in each window, obtain the similarity of the document by calculating the sampling fingerprint matching rate, which shows strong anti-interference ability for text block rearrangement and statement rearrangement. Through reasonable parameter setting, the Winnowing algorithm can effectively reduce the influence of noise words. From this perspective, the Winnowing algorithm can be seen as a lightweight, efficient and highly flexible similarity detection method [10].

In 2003, Chinese researchers Junpeng Bao, Junyiyi Shen, Xiaodong Liu [11] and others followed the time sequence of the document copy detection system based on natural language, and briefly enumerated the nine most representative copy detection systems in 2001, which is useful to the research of early document copy detection techniques. In figure 1, the feature extraction method, the similarity calculation method and the selection granularity of the text block enumerated by the author are worthy of further study.

System (Method)	Developer (s)	Year	Feature extracting method	Algorithm of similarity	Text chunk
Sif	U. Manber	1993	String matching	Number of common fingerprints	50 bytes after anchor
COPS	S. Brin, H. Garcia-Molina, *et al.*	1995	String matching	Matching ratio of fingerprints	Sentence
SCAM	N. Shivakumar, H. Garcia-Molina	1995	Word frequency	RFM	Word
YAP3	M.J. Wise	1996	RKR-GST, the longest matching string	Matching ratio	
KOALA	N. Heintze	1996	String matching	Matching ratio of fingerprints	20 characters
CHECK	A. Si, H.V. Leong, *et al.*	1997	SC and key words	Cosine function and matching ratio of section	Word and variable granularity
Shingling	A.Z. Broder, *et al.*	1997	String matching	Matching ratio of fingerprints	10 words
MDR	K. Monostori, *et al.*	2000	Suffix tree, the longest matching string	Matching ratio	60 characters
CDSDG	Song Qin-Bao, *et al.*	2001	SC and key words	Overlap of semantic and overlap of structure	Word and variable granularity

FIGURE 1. SUMMARY OF COPY DETECTION SYSTEMS BASED ON NATURAL LANGUAGE FROM 1993 TO 2001.

According to the table above, between 1993 and 2001, the natural language-based copy detection system mainly uses string matching methods such as Sif, COPS, KOALA, etc. when performing feature selection. The method based on word frequency or key words has also been applied. In the early days, these matching algorithms mainly focus on the lexical structure of the text. Among those methods, characters matching, string matching, fingerprints and VSM are widely used.

Since the method of string matching makes use of the grammatical structure of the document, some scholars also name it grammar-based copy detection method, which performs better in identifying verbatim plagiarism; the method based on word frequency

focuses more on the semantic features of the document and so to be called semantic-based methods, though it does not touch deep semantics. The combination of these two methods has been a third development direction of copy detection technology.

The most common method of text copy detection based on the method of string matching is fingerprinting. This method draws on the idea of the traditional hash algorithm and maps the text to a fingerprint by some fingerprint feature extraction algorithm. Similar texts are mapped to similar fingerprints, and the similarity between fingerprints is calculated to achieve the purpose of copy detection. The main points that must be taken into consideration when creating a digital fingerprint include the selection of fingerprint granularity, the choice of Hash Function, the strategy for fingerprint selection, and the resolution of the fingerprint [12]. The fingerprint granularity can be divided into large chunks (coarse granularity) and Small chunks (fine granularity). The selection of fingerprint granularity is related to the accuracy of fingerprints. The chapters and paragraphs can be considered as large chunks, but in practice, such large fingerprints are rarely selected. The choice of Hash Function which is suitable can minimize the conflicts when mapping different chunks to the same hash; the fingerprint resolution means the number of fingerprints used to reflect the characteristics of the paper; in the fingerprint selection strategy, different chunk selection strategies also produce different detection effects. These four main aspects are usually the first to be considered when improving a copy detection system based on fingerprinting.

The text chunk also has great influence on the copy detection systems. In most cases, copy detection methods do not select chunk of text or paragraph size, because the selection of text blocks is too large, which makes it difficult to identify part of plagiarism in comparison, and in fact complete plagiarism is relatively rare. Choosing a too small text chunk may also cause problems, because it can easily lead to misjudgment and increase the amount of computation. Each copy detection system listed differs from others in choosing text chunk. The COPS tool chooses sentence-level chunk while SIF selects 50 bytes after anchor. Other systems also select strings of fixed length or variable length as text chunk. For example, MDR selects text chunk with fixed length of 60 characters, and the authors of KOALA believe 20 characters are the best.

2.2.1. **Grammar-based Copy Detection Methods.** SIF proposed by Manber [13] marks the beginning of copy detection technology for natural language documents. But the main purpose of SIF is not to perform text copy detection, but to find similar documents in a large file system. Manber's main contribution is to propose the concept of approximate fingerprints. He believes that fingerprints can represent documents, and that similar documents must have same fingerprints. The method using fingerprints draws on the idea of the traditional Hash algorithm and maps the text into a fingerprint by some fingerprint feature extraction algorithms. Similar texts are mapped to similar fingerprints, and the similarity between fingerprints is calculated to achieve the purpose of copy detection. In 1995, Brin [14] and his team designed the COPS (Copy Protection System) system, which divides text into sentence sequences and replicates the number of sentences by comparing the same fingerprints between texts. From the perspective of system architecture, it is an

example for the later copy detection systems design. In addition, the prototype systems using fingerprints are KOALA, Shingling algorithm and Winnowing algorithm. In 2000, Monostori et al. [15] used a suffix tree-based string matching for text copy detection and proposed the MDR (Match Detect Reveal) system model, using suffix trees to search among strings to locate the largest substring. The suffix vector was then used to store the suffix tree, thereby improving recognition efficiency. This is another method based on the string matching.

2.2.2. **Semantic-based Copy Detection Methods.** The semantic-based copy detection technology mainly uses the concept of vector space, such as VSM, which uses the word frequency in the document to obtain the feature vector and performs the similarity computation of the document. In this method, the most basic method of similarity calculation is to use Dot Product, Cosine, Dice and Jaccard Coefficients. In practical applications, most of the methods have improved these basic calculation methods when performing similarity calculations.

The classical copy detection system prototypes using word frequency characterization are SCAM (Stanford Copy Analysis Method), CHECK model and HFM (High Frequency Model). In 1995, Shivakumar [16] proposed the SCAM (Stanford Copy Analysis Method) prototype system, which marks the beginning of the word frequency method. The SCAM aims to improve the detection scheme COPS which is based on sentence overlap in the same year. The SCAM system draws on VSM (Vector Space model) which is commonly used in the field of information retrieval for text representation and uses a more fine-grained word chunk instead of the sentence chunk used by COPS. Experiments show that SCAM can solve the problem of unclearness of the boundary boundaries in the COPS, and basically achieve better detection results than the COPS in the contrast experiment. Later, Molina et al. [17] upgraded SCAM to the DSACM system, expanding the scope of detection and applying it to web page text copy detection for the first time.

In 1997, Si, Leong [18] and others from the Hong Kong Polytechnic University established a CHECK prototype system based on keywords. The system decomposes the papers in Latex format, and then uses Vector Dot Product to compare the similarities, but CHECK changes the calculation method of vector Dot Products. Although CHECK can only identify papers in the format of Latex with small application scope, the structure of the text is introduced into the copy detection for the first time, which lays a foundation for the structure-based copy detection techniques in future study. In future study, VSM is also frequently used combining with various natural language processing (NLP) techniques to form a variety of detection methods.

2.2.3. **Semantic-grammar Based Copy Detection Methods.** In 2003, Hoad and Zobel [19] used word frequency and fingerprints to solve the problem of document recognition. It became the earliest detection system based on semantic and grammatical hybrid methods. In 2007, to solve the problem of Chinese copy detection, Bo Jin et al. [20] proposed a copy detection algorithm based on the similarity of document structure. Based on the analysis of the text structure of the paper, this method also comprehensively utilized fingerprints and word frequency, but only applies to those academic papers with accurate and canonical

structure.

Of course, early classification methods could not cover all document copy detection techniques. The emergence of new techniques for text copy detection is also constantly complementing existing researches. In addition to improvements to the above three methods, copy detection techniques have also generated based on structured-based methods, syntax-based methods, cluster-based methods, and cross-language detection.

3. **New Progress in Plagiarism Recognition Techniques.** PAN [21] (Plagiarism analysis, Authorship identification, Near-duplicate detection) is dedicated to the evaluation of plagiarism detection algorithms. Each year, PAN@CLEF international conference and competition will be held, bringing together the world's most advanced copy detection methods and techniques. At the same time, the competition adopts a unified standard test corpus PAN, which enables the comparisons between different copy detection techniques. The PAN corpus released each year is different except the training sets PAN-2013 and PAN-2014. The existing PAN corpus versions include PAN-2009, PAN-2010, PAN-2011, PAN-2012, PAN-2013, PAN-2014.

In 2010, Martin Potthast el al. [22] presents an evaluation framework for plagiarism detection which is further used by PAN. The evaluation indicators are recall, precision, granularity and plagdet_score. Among them, plagdet_score is calculated by the combination of recall, precision and granularity, which can make an overall evaluation of system performance. Therefore, this paper directly uses plagdet_score to compare the performance between systems. For, those do not have plagdet_score, recall and precision will be taken into consideration.

3.1. **Participating Systems of PAN**.

TABLE 1. THE BEST-PERFORMING PLAGIARISM DETECTION ALGORITHM EACH YEAR

Techniques or Methodologies	**PAN**	**Plagdet**
Character 16-grams, VSM, Cosine metrics	2009	0.69
Word 5-grams, Hashing	2010	0.8
Stemming with word-level matches	2011	0.56
Stop word removal, Stemming, VSM with overlapping measures	2012	0.74
TF-ISF weighting schemes with cosine and dice coefficients	2014	0.88

TABLE 2. Top 5 copy detection algorithms in pan competition from 2009 to 2014

Rank	**Techniques or Methodologies**	**PAN**	**Plagdet**
1	TF-ISF weighting schemes with cosine and dice coefficients	2014	0.88
2	Contexual N-grams, surround context N-grams, TF-IDF sentence level	2014	0.87
3	Stop word removal, stemming, VSM with overlapping measures	2012	0.74
4	Fingerprints and overlapping N-grams	2010	0.71
5	Character 16-grams, VSM approach with cosine metrics	2009	0.69
6	Stop word N-grams, N-grams with at least one named entity, and all words N-grams	2013	0.69

Throughout the performance of the plagiarism detection system in 2009-2014, it can be found that the best performing plagiarism detection system each year has achieved a relatively high plagdet_score (In TABLE 1). Among them, Sanchez-Perez et al. [23] win the highest score with TF-ISF weighting schemes reaching 0.88. The method is based on a TF-ISF method, which is similar to TF-IDF, with Cosine and Dice coefficient to measure the similarity of sentences. Then further recursive algorithm is used to extend the sentence to the maximum length, forming the largest similar text segment between two documents, and further calculating the similarity based on Cosine Similarity. Finally, the method improves the accuracy of the test by filtering out plagiarized fragments that are not up to standard, such as overlapping fragments and fragments that are too short.

Among all the methods involved in the PAN assessment in 2009-2014, methods based on VSM and N-grams were the most widely used. The top five methods in the 2009-2014 PAN competition can be seen in TABLE 2, the VSM and N-grams methods can also achieve good results.

TABLE 3. METHODS USING VSM

Year	Techniques or Methodologies	PAN	Plagdet
2009	Character 16-grams, VSM approach with cosine metrics	2009	**0.69**
	Word 5-grams, VSM approach with Jaccard similarity		0.61
	Word 8-grams, VSM approach with an N-gram based distance metric		0.60
	VSM with cosine metrics		0.18
2010	VSM based IR ranking with cosine metrics	2010	0.52
2012	Stop word removal, stemming, VSM with overlapping measures	2012	**0.74**

As is shown in Table 3, in 2009, all of four approaches in PAN competition used VSM model.

VSM is more commonly used in text categorization and works well. This method usually represents each text in a text data set as a vector. Based on the word frequency (TF, Term Frequency) and the inverse text frequency (IDF, Inverse Document Frequency), the corresponding weights of each term are used, and then the similarity between the texts is calculated by the method of finding the cosine of the vector angle with the European space. The method proposed by Grozea [24] and his partners abandons word-based method by selecting Character 16-grams, and calculates similarity with cosine metrics, which achieves the highest plagedet score of 0.69. The other two systems select Word 5-grams and Word 8-grams, and further use Jaccard coefficients and N-gram based distance metrics to calculate similarity. The scores also reached a high level of 0.61 and 0.60. Muhr [25] and others simply used VSM and cosine metrics to achieve better results in internal plagiarism testing, but only got 0.18 in external plagiarism detection. In 2012, the Stop word removal, stemming and VSM models adopted by Kong [26] and his team members also achieved a high score of 0.74, which is high among PAN participants. We can draw a conclusion that VSM model with appropriate feature extraction method and similarity calculation method can achieve better results in the future copy detection.

N-grams have also been widely used each year, and gradually achieved good results. In 2010, Gupta et al. [27] used the N-grams with entity identification method, but the score is only 0.2. In the same year, N-grams method of Oberreuter [28] scored 0.61. In 2012,

Torrejon et al. [29] used the method of Surrounding Context N-grams with a performance score of 0.63. In 2013 and 2014, with the method N-grams continuously improved, the scores also appeared to be higher, and the scores of all methods were in the upper position. In the 2009-2014 PAN competition, the second-ranked method combines three N-grams methods to improve the previous N-grams (Contexual N-grams, surround context N-grams, TF-IDF sentence level).

TABLE 4. METHODS USING N-GRAMS

Year	Techniques or Methodologies	PAN	Plagdet
2010	Contextual N-grams	2010	0.59
	N-gram models		0.61
	N-grams with named entity recognitions		0.20
	Fingerprints and overlapping N-grams		0.71
2012	Surrounding context N-grams	2012	0.74
2014	Contextual N-grams, surrounding context N-grams, named entity based N-grams,	2014	0.87

Among methods in the PAN competition, the semantic-syntactic method and the linguistic method are rarely used, and their general performances are not as good as VSM and N-gram methods. However, in recent years, these two types of methods can be proved to achieve good results in other tests. It is worth noting that after 2010, natural language processing technology has been continuously applied in document processing and similarity computation, showing powerful improvement effects, such as part-of-speech tagging (POS), semantic role tagging (SRL), stemming, Lemmatization, etc.

According to the analysis above, copy detection methods tend to combine multiple algorithms and use NLP techniques to analyze the structure and syntax of the text to further improve their performance. At the same time, traditional methods like fingerprints are still in use through further improvement. In this way, the accuracy can be further improved.

3.2. **Existing New Copy Detection Approaches.** In addition to participating systems of PAN above, many of the copy detection systems proposed in recent years shows great performance, some of which still use PAN corpus as a standard evaluation. In this section, some prominent plagiarism detection methods will be discussion, which show a basic route of plagiarism detection.

In the earlies days, semantic-based copy detection systems tend to use VSM to compute semantic similarity. However, VSM cannot touch the deep semantics of words or sentences. Plagiarism types including rewording and paraphrasing can be hard to detect. To solve these problems, first many researchers try to calculate the semantic similarity between words or sentences by dictionaries like WordNet. Then, many techniques in the fields of Natural Language Process are tested. Generally, NLP methods performs better than other methods when detecting paraphrased texting. Among NLP methods, Semantic Role Labeling (SRL) is preferred by many researchers. In 2012, the team of Osman et al [30]. is the first to apply Semantic Role Labeling (SRL) in plagiarism detection. They use SRL to analyze the sentences semantically and WordNet thesaurus to extract the concepts or synonymies for each word inside the sentences. Then, Jaccard coefficient is used to calculate a total similarity. The special point of this method is that it considers the

relationships among its terms to capture the underlying semantic meaning. The scheme was tested on PAN-PC-09 data sets and showed a good performance in terms of Recall, Precision and F-measure. In 2015, Merin Paul et al. [31] proposed an improved method combining SRL and sentence ranking. The sentence ranking method can pick suspicious sentence pairs before adopting Semantic Role Labeling. In this way, the time of checking by SRL can be reduced. The experiment proves that SRL with sentence ranking takes less time than SRL-based method. In 2017, Abdi et al. [32] further improved this method by combining the semantic and syntactic information with SRL, this method tries to solve the problem that some sentences share similar bag-of-words (BOW), but have different meanings, which may be easily misjudged when detecting their similarity. The linguistic features like word-order is also used in this method. The plagdet_score reaches 0.737 of PAN-PC-11.

Latent Semantic Analysis (LSA) is also related to semantic analysis. In 2016, AlSallal, Muna, et al. [33] proposed a hybrid detection system with LSA, SVM and the Common Word (MCW). The interesting part is that the item MCW belongs to Stylometric research, which aims to characterize the author's writing style. In this method, first a basic text matrix is built using bag of words mode. Then, LSA is applied in two stages. The first stage gives low weight for MCW using TF-DIF while the second stage gives high weight for MCW using adjusted TF-IDF. The method of Singular Value Decomposition (SVD), the core component of LSA, is then applied for both stages. SVM here is used to build a classifier model.

Machine learning algorithms, especially deep learning methods show great potential in this field. In the field of text classification, machine learning techniques such as Naïve Bayes, Support Vector Machine (SVM) etc. have good performance and can be used. In 2014, Zakiy Firdaus Alfikri [34] used machine learning algorithms Naïve Bayes and SVM for extrinsic plagiarism detection. The features learned by machine learning methods are word similarity, fingerprint similarity, latent semantic analysis similarity and word pairs. This process combines the advantages of the above features. The authors obtained the detection accuracy of 92.86% using SVM, which is better than Naïve Bayes' 54.29%. A conclusion can be made that the machine learning algorithm, especially the SVM model, has a good application prospect in plagiarism detection. In 2016, Yuliang Liu [35] and others proposed a new method based on neural network. This method also uses a string matching but adopts deep learning algorithms of recurrent neural network (RNN) or convolutional neural network (CNN). The method abstracts the text fingerprint extraction into a coding-decoding problem. In the same year, Erfaneh Gharavi et al. [36] used deep representation of words for plagiarism detection task. The author summarized advantages of using deep learning methods, including fewer number of features, no labeled data. The most important point is that, deep learning can improve the efficiency of traditional NLP methods and thus raise the speed of plagiarism detection system.

When talking about clustering. In 2011, Stamatatos [37] gave a new idea by proposing the Stop Words N-gram (SWNG) method, a copy detection method based on structural information rather than content information, which marked a breakthrough in traditional

copy detection technology. To reduce the number of features and improve the detection efficiency, the traditional copy detection technology often removes stop words in the pre-processing stage, while SWNG removes the real words when extracting the document structure information and retains the stop words. The basic idea of this model is that stop words can maintain the stability of sentences. No matter how synonym replacement or sentence reorganization is conducted, stop words usually remain unchanged. Moreover, SWNG method is also capable of extracting plagiarized passage boundaries. Stamatatos used the PAN 2010 Plagiarism Competition corpus for testing, and it performed well in long text tests, reaching a total score of 0.87. In 2016, Gupta D [38] and his partners used Sentence Bounded Stop Word N-gram method (SBSWNG) to further improve the SWNG method mentioned before. NLTK (Natural Language Toolkit) is chosen for part-of-speech tagging when confirming the structure of the sentence. In the experiment, the author compares this method with the state-of-art method based on Word N-grams and the method based on Stop-Words N-gram. This method chooses PAN-13 (Plagiarism, Authorship and Social Software Misuse) "text-aligned corpus" as its test set and prove to achieve better performance.

Compared with other methods focusing on the preprocessing process or comparison process, Cluster-based method pay more attention to the candidate retrieval stage. This method was first proposed in 2014 when Vani [39] et al. chose K-means algorithm. It is observed in the test of PAN-2013 that K-means method gives promising results when dealing with highly obfuscated data. However, K-means clustering creates K non-overlapping clusters of the document features and thus cannot easily recognize sentence boundaries. To solve this problem, in 2016, Ravi et al. [40] proposed Fuzzy C Means Clustering Algorithm and showed a better performance. In the same year, Alzahrani et al. [41] made some improvements by using logical tree-structured features and multi-layer clustering. The top layer features are used to find similar clusters and perform candidate retrieval while the bottom layer features are used to cluster structural components and to detect plagiarism. Finally, detailed analysis and similarity calculation are performed to find the structural components that are highly similar. This approach proves to be better than K-means clustering algorithms.

Word Embedding is welcomed in recent two years. In 2017, Kensuke Babaa et al. [42] used a distributed representation obtained from word2vec for the word similarity and combined it with the Longest Common Subsequence (LCS). The experiment shows that the length of a "weighted" local LCS is the best. In 2018, two new methods proposed use word embedding. Erfaneh Gharavi et al. [43] combine word embedding with Jaccard. Sentences are represented by Composition Function, including Paragraph Vector, RAE (Recursive Auto-encoder) and Matrix Vector Recurrent Neural Networks while Jaccard coefficient is used to calculate the lexical similarity. The system runs fast because it does not need lexicon and preprocessing stage, including POS-tagging, stemming, etc. In the same year, Khorsi et al. [44] combined fingerprinting with word embedding to construct a two-layer plagiarism detection system. Fingerprinting is used to detect verbatim plagiarism in lexical level, while word embedding is applied later to recognize intelligent plagiarism. The

authors put forward a new selection strategy, which only stores the less-frequent n-grams to reduce the n-gram inverted size. Also, CBOW (Continuous Bag of Words Model) is used in word representation and each word is represented by a vector of 300-dimension. At the second level, word-alignment based on semantic similarity is used to improve similarity results. Two weighting functions (IDF/POS) are used to weight the aligned words.

Latent Dirichlet Allocation (LDA) can also be found in designing copy detection system. In 2018, Naif Radi Aljohani [45] and others used natural language processing techniques to perform external plagiarism detection based on semantic-syntactic methods, the combination of Latent Dirichlet Allocation (LDA) and Part of Speech Tags (POS). Semantic information is added even if the part-of-speech features alone can be used satisfactorily. LDA is used to capture semantic similarity while POS is used to compare syntactic similarity. In the experiment, the author compares this method with the state-of-art method based on Word N-grams and the method based on Stop-Words N-gram. The corpus used is PAN-13 (Plagiarism, Authorship and Social Software Misuse) Championship "text-aligned corpus". Experiments have shown that this method has achieved better results.

When extracting syntax-semantic concept, genetic algorithm (GA) can also be used. In 2018, Vani et al. [46] proposed an idea using this method. Genetic Algorithm is applied to find out the interrelated cohesive sentences that can convey the concept or idea of the source document. Both passage level and document level plagiarism detection are tested in PAN13-14 Corpus, and combined plagdet score is 0.7663. In this method sentences are represented in vector space model (VSM) with term frequency-inverse sentence frequency (TF-ISF) weighting instead of TF-IDF weighting.

In 2015, Suhong Wang [47] and others proposed a plagiarism detection method based on information retrieval and VSM. The performance was tested with PAN-2010 corpus. The core content of this approach is to use the information retrieval system to retrieve the source documents corresponding to the licensable documents from the reference document set, and to form the <suspicious document, candidate documents> for feature extraction, and further obtain feature values expressed by vector. These features are further used to train VSM classifier. The author compared the method with the top three methods of PAN@ CLEF 2012 and obtained a total score of 0.708, which further verified that the VSM method can be effectively utilized in copy detection.

According to the comparison, we can see that in recently years, more and more NLP techniques are used in plagiarism detection system, including POS, SRL, etc. The main goal is to comprehensively compare similarity especially in terms of syntax and semantics. Moreover, hybrid methods are preferred and multi-layer systems are preferred to increase the accuracy. Apart from traditional methods, machine learning methods and deep learning methods can be used in some phases of the system to speed up the whole process by reducing some unnecessary stages.

4. **Comparisons of Common Plagiarism Detection Systems at Home and Abroad.** The commonly used check-up systems abroad include Turnitin, CrossCheck, Dupli Checker,

etc., Considering three most commonly used commercial paper citation detection systems in China, Vip-Tongda Paper Citation Detection System (VTTMS), the CNKI Dissertation Academic Misconduct Detection System (Referred to as AMLC), Wanfang are the most welcomed. The remaining paper detection systems such as PaperTime, PTCkeck, etc. are also based on the web platform, and provide various functions such as online modification, robot weight reduction and so on. Due to the difference between Chinese characters and English characters, plagiarism detection systems at home and abroad are difficult to be universal.

Table 5 compares the replication detection techniques mainly used in plagiarism detection systems commonly used in domestic and international papers.

TABLE 5. PLAGIARISM DETECTION AT HOME AND ABROAD

Name of Software	Techniques or Methodologies
Home(China)	
VTTMS	"F&V" algorithm: VSM+、Semantic Fingerprint; Automatic Classification
AMLC	Multi-level Adaptive Fingerprint Analysis, Semantic Comprehension Technology
INFOSOFT	Low-frequency feature partial matching algorithm based on sliding window, batch detection simplification technology
PaperTime	Multi-level fingerprint contrast technology; deep semantic exploration and recognition technology; fingerprint encryption
GeZiDa	Fingerprint comparison; VSM+; the semantic matching algorithm
PTCheck	Dynamic fingerprint over-the-level scanning technology, semantic matching database
PaperRight	Dynamic semantic cross-domain recognition technology, RSA encryption technique
Abroad	
Turnitin	Fingerprint based
Dupli Checker	String matching
Copyleaks	String matching, word frequency (VSM)
PaperRater	String matching
Plagiarisma	String matching
PlagTracker	String matching
Quetext	String matching; Re-score tokens based on context

After comparison, it can be found that most of the paper plagiarism detection software at home and abroad still use the method of string matching. The commonly used check software in China tends to combine digital fingerprints and semantic understanding technology. The method of Wanfang data base is quite special. It self-proclaims to use self-developed low-frequency feature partial matching algorithm based on sliding window. Most of the foreign paper plagiarism detection systems choose the method of string matching, but there is no specific explanation in the official website. The use of vector space model for word frequency statistics and calculation of similarity has also been

applied.

VTTMS [48] adopts the so-called self-developed "F&V" algorithm - a collection model of VSM+, semantic fingerprinting and automatic classification. The semantic fingerprint is used to detect the entire text; VSM is used to automatically analyze the semantic segment; the automatic classification is used to automatically detect the detected document to a professional comparison source for detection. VIP's detection granularity supports a minimum of phrase level, and the detection granularity is smaller than other similar products.

AMLC [49] adopts multi-level adaptive fingerprint analysis and semantic understanding technology. For any document that needs to be detected, the system firstly processes it hierarchically, and creates fingerprints according to chapters, paragraphs, sentences, etc., and compares them. The comparison literature in the resource library also uses the same technology to create a fingerprint index and builds a powerful semantic analysis framework to achieve semantic analysis of words, sentences, sentence groups, and chapters.

PaperTime [50] adopts a fingerprint matching algorithm based on big data. Under the pre-processing of all fingerprints of the paper, multi-level fingerprint comparison technology combined with deep semantic exploration and recognition technology is used to quickly and accurately find all similar fragments by fingerprint index. It is said to increase the speed by 10 times compared with the conventional speed. In the case of ensuring the quality of the check, the result can be checked in a few seconds. In terms of security, fingerprint comparison is used, and the original text is converted into an encrypted fingerprint after uploading. There is no problem of original text leakage.

GeZiDa [51] plagiarism detection software uses semantic fingerprint multiple recognition technology, minimal support for phrase-level particles, and a unique dynamic recursive semantic comparison algorithm. The minwise hash algorithm is used in fine-grained text extraction. The system selects phrase-level particles in the multiple fingerprint recognition of semantic fingerprints, which is also more breakthrough. In the process of extracting the feature set of the document, the word segmentation, the stop word and the extracting the shine feature are mainly included.

Turnitin [52] has its unique techniques and iThenticates database. Generally, it adopts the methods of string matching, specifically digital fingerprinting. To cover more database and increase its detection accuracy, Turnitin also uses the method applies in the field of information retrieval. Now, Turnitin is widely used around the world by providing copy detection service for various languages.

According to the analysis, some domestic check-in systems add semantic understanding technology at the basis of fingerprint analysis, combining the structural information of the paper with the semantic information. They also tend to introduce context information. At the same time, it is devoted to researching the catalogue and originality declaration. Automatic identification by reference documents, etc., excludes irrelevant information from the detection range, and further improves the accuracy of the check. The method of machine learning has not been clearly reflected in the application of the check-up software on the market. In terms of improving the speed of checking, the existing paper plagiarism

detection system tends to optimize the calculation method, move the calculation to the cloud, use distributed computing to improve the detection efficiency, and shorten the examination time of the paper.

5. **Development Trend and Future of Plagiarism Detection System.** The copy detection techniques have been developed since the SIF prototype system in 2003, and has been continuously improved in all directions, and gradually achieved good results in the practice of plagiarism detection. With the analysis of comprehensive texts, the copy detection techniques will present a diversified development trend in the future, including various detection technologies and improved detection speed. At the same time, the plagiarism detection system of the paper is gradually becoming more personalized, such as providing a dual detection mode of "resource library" + "self-built library", providing a range of comparison resource pools according to different needs of users; Also, many websites use artificial intelligence method to help users reduce repetitive rates, which has caused some changes in the purpose of plagiarism detection.

5.1. **Diversification of Plagiarism Techniques.** In terms of detection technology, copy detection systems tend to apply the diversity of methods, which means it is not limited to the traditional three basic methods, namely semantic based, grammar based and semantic and grammar-based hybrid detection methods. But the traditional methods have not been abandoned, such as fingerprinting, until 2018, still researchers try to improve and use this kind of method.

New detection methods are emerging, such as fuzzy-based methods, cluster-based methods, character n-gram based methods, structural based methods, cross-lingual methods, from the perspective of the semantics, grammar, text structure, context, etc. of the article, even cross-language plagiarism detection. According to the analysis of the text, most plagiarism detection systems tend to combine two or three detection methods to improve the accuracy of the system. Some systems have adopted different ideas. According to different plagiarism methods, the system uses more than 1 layer and the selection algorithm is selected to improve the detection accuracy [53].

With the development of natural language processing, more natural language processing techniques will continue to be tried to design copy detection systems, such as POS, SRL, LSA and LDA. Advanced techniques in the field of natural language processing can also continuously improve the performance of plagiarism detection systems. More and more machine learning, especially deep learning algorithms, will also emerge in the development of plagiarism detection system, further promoting the development of cross-language detection.

5.2. **Constantly Increasing Detection Speed.** Plagiarism detection takes place in a large collection of documents, which requires a lot of time and resources. In terms of improving the detection speed, in addition to the improvement of the replication detection method itself, the commercial paper plagiarism detection system can also utilize distributed computing frameworks such as Hadoop and Spark in computing. The use of distributed computing can effectively improve the detection efficiency and shorten the examination

time of the paper. In the existing software, VTTMS, PaperRight, etc. adopted a distributed computing method, which greatly shortened the paper detection time. With the development of cloud computing technology, it has become an inevitable trend to transfer the data processing part of plagiarism detection to cloud computing.

5.3. **Application of Web Resource.** In the process of selecting comparison resources, some check software establishes a multi-dimensional comparison resource system, which combines professional database, network database, shared database and user-built database. For example, VTTMS adopts four-dimensional comparison source, which is VTTMS professional Database - the largest and most complete Chinese scientific and technical journal full-text database, currently has more than 26.7 million full-text articles; Web resources - monitoring billions of pages included in Google, updated weekly; Tonda shared database - including more than 2 million papers which are updated weekly; users build their own libraries to meet user-specific matching needs. Multi-dimensional resources make the resource library larger, and the accuracy of checking is improved. At the same time, the Web-based resources have the characteristics of real-time updating, which improves the accuracy of checking. In an environment where network resources are rapidly updated, future plagiarism detection will continue to be integrated into network resources, and the use of network search engines will expand the scope of comparison.

5.4. **Personalized Service**

5.4.1. **"Resource Library plus Self-built Library"- Dual Detection Mode.** "Self-built library" refers to the user's own upload of document resources to establish a comparison library. Users can upload all the documents (document format supporting doc/docx/txt, etc.) referenced in the writing process to the self-built library, and effectively compare them by checking the self-built library on the comparison source selection page. The “resource library + self-built library” method can help users customize the scope of literature search and make the check more personalized. In the existing software, VTTMS, PTCheck, etc. all support the user-built library function. “Resource Library + Self-Building Library” gives different audiences the opportunity to self-select, making plagiarism detection more practical and a wider range of applications.

5.4.2. **Artificial Intelligence-based Services for Reducing Replication Repetition Rate.** The purpose of plagiarism testing is to resist academic misconduct and protect intellectual property rights. According to the research in this paper, more and more domestic websites are beginning to provide manual or artificial intelligence-based weight loss services. Does this deviate from the original intention of plagiarism testing, and it is worthy of further discussion on whether the academic talents in colleges and universities have adverse effects.

REFERENCES

[1] CNKI Scholar: http://elib.cnki.net/grid2008/Help/AssistDocument/036/html/main.htm

[2] https://www.plagiarism.org/article/plagiarism-facts-and-stats

[3] Yanjun Shi, Hongfei Teng, Bo Jin et.al. Research and Development of Plagiarism Paper Recognition. Journal of Dalian University of Technology. 2005, 45(1):50-57.

[4] Jianhua Su. A Review of Research on Plagiarism Recognition Technology [J]. Digital Library Forum, 2007(11): pp. 61-64.

[5] D Namdev, J Surana et al. A Survey Paper on Plagiarism Detection Techniques [J]. International Journal of Computer Applications (0975 – 8887), 2015.

[6] Jianhua Su. A Review of Research on Plagiarism Recognition Technology [J]. Digital Library Forum, 2007(11): pp. 61-64.

[7] Ottenstein K J. An Algorithmic Approach to The Detection and Prevention of Plagiarism [M]. ACM, 1976.

[8] Karp R M, Rabin M O. Efficient Randomized Pattern-Matching Algorithms [J]. IBM Journal Of Research & Development, 1987, 31(2):249-260.

[9] Schleimer S, Wilkerson D S, Aiken A. Winnowing: Local Algorithms for Document Fingerprinting[C]// Proc. ACM Sigmod Conference, June. 2003:76-85.

[10] Duan X, Wang M, Mu J. A Plagiarism Detection Algorithm Based on Extended Winnowing[C]// 2017:02019.

[11] Junpeng Bao, Junyi Shen, Xiaodong Liu Et Al. A Survey of Research on Natural Language Document Replication Detection [J]. Journal of Software, 2003, 14(10):1753-1760.

[12] El Moatez Billah Nagoudi,Ahmed Khorsi,Hadda Cherroun,Didier Schwab. 2l-Apd: A Two-Level Plagiarism Detection System for Arabic Documents[J]. Cybernetics and Information Technologies,2018,18(1).

[13] Manber U. Finding Similar Files in A Large File System[C]// Usenix Winter Technical Conference. 1994:1--10.

[14] Brin S, Davis J, Garcíamolina H. Copy Detection Mechanisms for Digital Documents[J]. ACM Sigmod Record, 1995, 24(2):398-409.

[15] Monostori K, Zaslavsky A, Schmidt H. Document Overlap Detection System for Distributed Digital Libraries[J]. 2000, 22(3):226-227.

[16] Shivakumar N. Scam: A Copy Detection Mechanism for Digital Documents[J]. Proc Dl, 1995.

[17] Garciamolina H, Gravano L, Shivakumar N. Dscam: Finding Document Copies Across Multiple Databases[C]// International Conference on Parallel and Distributed Information Systems. Ieee, 1996:68-79.

[18] An Tonio S, Leong H V, Rynson W H . Check: A Document Plagiarism Detection System [C]// Proceedings of Acm Symposium For Applied Computing. San Jose.

[19] Hoad T C, Zobel J. Fast Video Matching with Signature Alignment[C]// ACM Sigmm International Workshop on Multimedia Information Retrieval. ACM, 2003:262-269.

[20] Bo Jin, Yanjun Shi, Hongfei Teng. Replication Detection Algorithm Based on Chapter Structure Similarity[J] Journal of Dalian University of Technology, 2007, 47(1):125-130.

[21] PAN: http://pan.webis.de

[22] Potthast, Martin, Et Al. "An Evaluation Framework for Plagiarism Detection." Proceedings of the 23rd International Conference On Computational Linguistics: Posters. Association for Computational Linguistics, 2010.

[23] Sanchez-Perez, M., Sidorov, G., & Gelbukh, A. (2014). A Winning Approach to Text Alignment for Text Reuse Detection – Lab Report for Pan at Clef 2014. Proceedings of the 6th International Workshop Pan-14. Sheffield, Uk.

[24] Grozea, C., Gehl, C., & Popescu, M. (2009). Encoplot: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. Proceedings of the 1st International Workshop Pan-09. Universidad Polit¨¦Cnica De Valencia And Ceur-Ws.Org.

[25] Muhr M, Zechner M, Kern R. External and Intrinsic Plagiarism Detection Using Vector Space Models [J]. In: Proceedings of the Sepln 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (Pan 09, 2009.

[26] Kong, L., Qi, H., Shuai, Wang, Du, C., Suhong, Wang, & Han, Y. (2012). Approaches for Candidate Document Retrieval and Detailed Comparison Of Plagiarism Detection—Notebook For Pan At Clef 2012. Proceedings of The 4th International Workshop Pan-12, Rome, Italy.

[27] Oberreuter, G., Huillier, G. L., Ríos, S. A., & Velásquez, J. D. (2010). Fastdocode: Finding Approximated Segments Of N-Grams for Document Copy Detection: Lab Report for Pan At Clef 2010. Proceedings of the 2nd International Workshop Pan-12, Padua, Italy.

[28] Torrejón, D. A. R., & Ramos, J. M. M. (2010). Coremo System (Contextual Reference Monotony) A Fast, Low Cost and High Performance Plagiarism Analyzer System: Lab Report For Pan At Clef 2010. Proceedings of the 2nd International Workshop Pan-10, Padua, Italy.

[29] Osman, Ahmed Hamza, et al. "Plagiarism detection scheme based on Semantic Role Labeling." Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on. IEEE, 2012.

[30] Paul, Merin, and Sangeetha Jamal. "An improved SRL based plagiarism detection technique using sentence ranking." Procedia Computer Science 46 (2015): 223-230.

[31] Abdi, Asad, et al. "A linguistic treatment for automatic external plagiarism detection." Knowledge-Based Systems 135 (2017): 135-146.

[32] AlSallal, Muna, et al. "An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection." Developments in eSystems Engineering (DeSE), 2016 9th International Conference on. IEEE, 2016.

[33] Alfikri Z F, Purwarianti A. Detailed Analysis of Extrinsic Plagiarism Detection System Using Machine Learning Approach (Naive Bayes And Svm) [J]. Telkomnika Indonesian Journal of Electrical Engineering, 2014, 12(11).

[34] Alfikri, Zakiy Firdaus, and Ayu Purwarianti. "Detailed analysis of extrinsic plagiarism detection system using machine learning approach (naive bayes and svm)." Indonesian Journal of Electrical Engineering and Computer Science 12.11 (2014): 7884-7894.

[35] Yuliang Liu, Xiaohua Liu, Lianshuan Li, Et Al. A Method for Detecting Plagiarism of Academic Literature Based on Deep Neural Network: Cn 106095735 A[P]. 2016.

[36] Gharavi, Erfaneh, et al. "A Deep Learning Approach to Persian Plagiarism Detection." FIRE (Working Notes). 2016.

[37] Stamatatos E. Plagiarism Detection Based on Structural Information[C]// ACM International Conference on Information and Knowledge Management. ACM, 2011:1221-1230.

[38] Gupta, D., Vani, K., & Leema, L. M. (2016). Plagiarism Detection in Text Documents Using Sentence Bounded Stop Word N-Grams. Journal of Engineering Science and Technology, 11(10), 1403–1420.
[39] Vani, K., and Deepa Gupta. "Using K-means Cluster based techniques in external plagiarism detection." Contemporary computing and informatics (IC3I), 2014 international conference on. IEEE, 2014.
[40] Ravi, N. Riya, K. Vani, and Deepa Gupta. "Exploration of fuzzy C means clustering algorithm in external plagiarism detection system." Intelligent Systems Technologies and Applications. Springer, Cham, 2016. 127-138.
[41] Alzahrani, Salha, Naomie Salim, and Vasile Palade. "Framework for Plagiarism Detection Using Logical Tree-Structured Features and Multi-Layer Clustering."
[42] Baba, Kensuke, Tetsuya Nakatoh, and Toshiro Minami. "Plagiarism detection using document similarity based on distributed representation." Procedia computer science 111 (2017): 382-387.
[43] Erfaneh Gharavi et al. Plagiarism Detection Using Word Embedding, University of Tehran
[44] Khorsi, Ahmed, Hadda Cherroun, and Didier Schwab. "A Two-Level Plagiarism Detection System for Arabic Documents." Cybernetics and Information Technologies 20 (2018).
[45] Naif Radi Aljohani,Jalal S. Alowibdi,Ali Daud,Jamal Ahmad Khan,Jamal Abdul Nasir,Rabeeh Ayaz Abbasi. Latent Dirichlet Allocation and Pos Tags Based Method for External Plagiarism Detection: Lda And Pos Tags Based Plagiarism Detection [J]. International Journal on Semantic Web and Information Systems (Ijswis), 2018, 14(3).
[46] Vani, K., & Gupta, D. (2017). Identifying Document-Level Text Plagiarism: A Two-Phase Approach. Journal of Engineering Science & Technology (Jestec), 12(12), 3226¨C3250.
[47] Suhong Wang, Huining, Song Yang Et Al. Research on Plagiarism Detection Method Based on Svm [J]. Journal of Applied Science, 2015, 42(5): 51-54.
[48] VTTMS: http://vpcs.cqvip.com/login.aspx?r=%2f.default.aspx
[49] AMLC: http://check.cnki.net
[50] Papertime: http://www.papertime.cn/
[51] GeZiDa: http://www.gezida.com/
[52] Turnitin: http://Turnitin.Com/En_Us/
[53] A Khorsi, H Cherroun, D Schwa.2l-Apd: A Two-Level Plagiarism Detection System for Arabic Documents- Cybernetics and Information. 2018.