مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (38) العدد (6) 2016
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (38) No. (6) 2016

Survey Of Traditional And Semantic Plagiarism Detection Algorithms

Dr. Khaled Omar [*]
Sobhi Al-Shikha**

(Received 17 / 11 / 2016. Accepted 29 / 11 / 2016)

□ ABSTRACT □

In this paper we review and list, the advantages and limitations of the significant effective techniques employed or developed in text plagiarism detection. It was found that many of the proposed methods for plagiarism detection have a weakness points and do not detect some types of plagiarized operations.

This paper show a survey about plagiarism detection including several important subjects in plagiarism detection, which is plagiarism definition, plagiarism prevention and detection, plagiarism detection systems, plagiarism detection processes and some of the current plagiarism detection techniques.

This paper compares between different plagiarism detection algorithms, and shows the points of weakness, and points of efficiency, and describe the power of semantic plagiarism detection methods, and shows its efficiency in detect plagiarism cases that another plagiarism detection algorithms don't able to detect these cases, that semantic plagiarism detection methods are developed to get rid of traditional weakness points for all plagiarism detection methods have.

Keywords: Semantic Plagiarism Detection algorithms; Detection Process, Detection Techniques

[*] **Faculty of Information Technology Engineering. Damascus University, Syria.**
****Faculty of Electrical and Mechanical Engineering, Damascus University, Syria.**

مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (38) العدد (6) 2016
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (38) No. (6) 2016

استعراض خوارزميات كشف الانتحال التقليدية والدلالية

د.خالد عمر*

صبحي الشيخة**

(تاريخ الإيداع 17 / 11 / 2016. قُبِل للنشر في 29/ 11 / 2016)

□ ملخّص □

في هذه الورقة العلمية نستعرض ونسرد ، المزايا والقيود المفروضة على التقنيات الفعالة المهمة التي تم توظيفها وتطويرها لكشف الانتحال في النصوص . وقد تبين أن العديد من الأساليب المقترحة لكشف الانتحال لديها نقاط ضعف وعدم الكشف عن بعض الأنواع من عمليات الانتحال.

وتقوم هذه الورقة بدراسة مسحية حول كشف الانتحال بما في ذلك العديد من الموضوعات المهمة في كشف الانتحال، وهي تعريف الانتحال، ومنع الانتحال وكشف الانتحال ، وأنظمة كشف الانتحال، وعمليات كشف الانتحال وبعض تقنيات كشف الانتحال الحالية.

تقارن هذه الورقة بين مختلف خوارزميات كشف الانتحال، وتظهر نقاط الضعف، ونقاط القوة، وتوصف قوة خوارزميات كشف الانتحال الدلالية، وتظهر فعالية هذه الخوارزميات في الكشف عن حالات الانتحال لا تستطيع خوارزميات كشف الانتحال الأخرى اكتشافها،حيث أنه تم تطوير خوارزميات كشف الانتحال الدلالية للتخلص من نقاط الضعف التقليدية التي تمتلها جميع خوارزميات كشف الانتحال الأخرى.

الكلمات المفتاحية: خوارزميات كشف الانتحال الدلالية، عملية كشف الانتحال،تقنيات كشف الانتحال.

* كلية الهندسة المعلوماتية، جامعة دمشق، سورية.

**كلية الهندسة الكهربائية والميكانيكية، جامعة دمشق، سورية.

1. INTRODUCTION

Plagiarism is defined as the unauthorized use or close imitation of the language and thought of authors and their representation as one's own original work [1].

There are many types of plagiarism, such as copy and paste, plagiarism of paragraphs, plagiarism of idea, and cross language plagiarism, which is done through translation from one language to another.

These types have made plagiarism very big problem in academic Education. A modern research in the past few years is done to define plagiarism, and to develop methods for detect it, and to make methods to prevent plagiarism from done

It found that 70% of students confess to a few plagiarisms, with about half being guilty of an earnest cheating offence on a written assignment. Additionally, 40% of students confess to using the "cut- paste" method when completing their assignments [2]. Differentiating between the plagiarized documents and non-plagiarized documents in an effective and efficient way is one main issue in plagiarism detection field. According to Carroll [3], at least 10% of student's work is likely to be plagiarized in USA, Australia and UK universities [4].

The rest of the paper is organized as follows: Section 2 provides a description of Plagiarism detection Systems. Section 3 discusses how to reduce plagiarism; Section 4 describes the traditional plagiarism detection algorithms. Section 5 describes the semantic plagiarism detection algorithms, Full comparison between plagiarism detection algorithms properties are presented in Section 6 whereas Section 7 concludes the paper.

2. PLAGIARISM DETECTION SYSTEMS :

1) Web-enabled systems: Developing web systems for plagiarism detection overcomes machine capability problems, facilitate the availability of the system to many users and extend the search of plagiarized resources to the World Wide Web easily. Here is discussion of two: First Turnitin [5, 6] is the most well-known commercial plagiarism detection system to which many universities from UK and USA subscribe. It uses an enormous database from the Internet and previous student works to be compared with the query document. Second SafeAssign [7] checks all submitted papers against the following databases: (i) the Internet. (ii) ProQuest database. (iii) Institutional document A Survey on Plagiarism Detection Systems A. S. Bin-Habtoor and M. A. Zaher International Journal of Computer Theory and Engineering Vol. 4, No. 2, April 2012 185 archives containing all documents submitted to SafeAssign. (iv) Global Reference Database containing documents that were volunteered by students to help prevent cross-institutional plagiarism.

2) Stand-alone systems: Stand-alone software is developed to be installed on computers. Two systems will be explored here, EVE [6, 8, 9] and WCopyFind [6, 9, 10]. First EVE (The Essay Verification Engine) is a desktop application but it has the capability to make large number of searches on the Internet to locate matches between sentences in the query document and suspected websites. Thus, in order for EVE to work, the machine should be connected to the Internet. Second WCopyFind developed by University of Virginia, finds plagiarism between two or more assignments. The user can set or change some of the parameters that may influence the detection process such as the number of words used for detecting similarity among statements.

3. REDUCING PLAGIARISM :

there are two ways to reduce plagiarism as it shown in the following figure[11]:

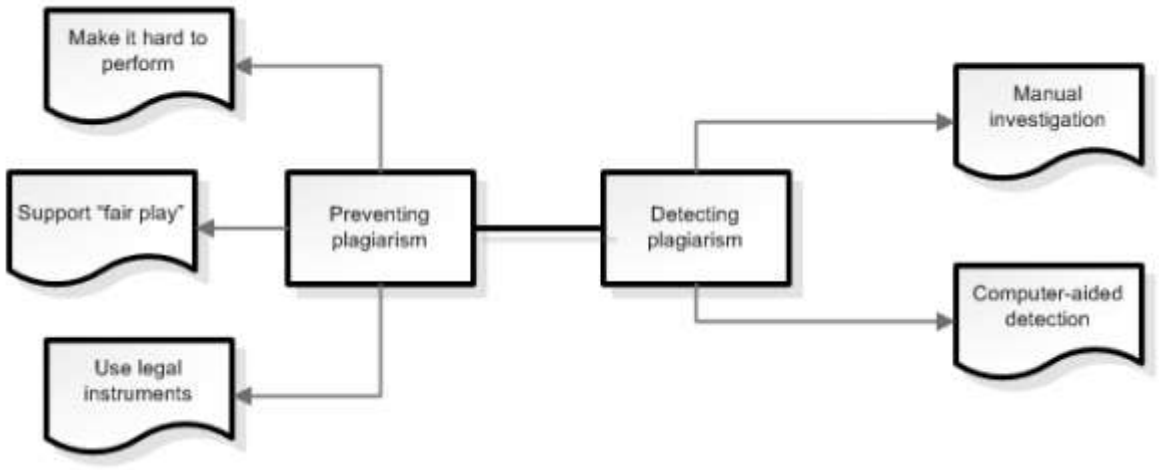


Fig (1): Plagiarism Preventing and detecting [11]

Plagiarism prevention : including make plagiarism hard to perform by preparing individual exercises, concentrating on the classroom work, and by using software tools, to support "fair play" among students by forming their positive attitude to academic honesty principles , and to explicitly formulate "fair" and "unfair" techniques[11].

To publish legal documents such as honor codes and university regulations, stating strict punishments for plagiarism[11].

Plagiarism detection: it includes manual plagiarism detection, and computer – aided detection by developing a system for automatic plagiarism detection[11].

4. TRADITIONAL PLAGIARISM DETECTION ALGORITHMS: This section mainly discusses all category of plagiarism detection algorithms ,and describes the weakness points of every type of these algorithms.

4.1. fingerprinting algorithms: these algorithms generate unique code for every file and this code is called as file fingerprint , and then these algorithms use fingerprints codes to compares between files (origin and suspected files) ,there are three types of fingerprint algorithms as the following[12]:

- Char based fingerprint
- Word based fingerprint
- Sentence based fingerprint

The most famous fingerprint plagiarism detection algorithm is "Winnowing" algorithm, which runs as explained in the following example:

Example[13]: let the following is a sentence of origin file:"A do run run run, a do run run"

The algorithm removes the spaces between words: "Adorunrunrunadorunrun"

Then it generates windows (its length is k =5) which called 5-grams windows, the windows are generated by one alignment in each iteration, after generating these windows the text will be:

"adoru dorun orunr runru unrun nrunr runru unrun nruna runad unado nador adoru dorun orunr runru unrun"

Then the algorithm generates a hash code for every generated window then it formulates the file fingerprint code as the following:

{ 77 72 42 17 98 50 17 98 8 88 67 39 77 72 42 17 98}

Then the algorithm apply a heuristics for choosing set of codes to be the representing for the file, winning apply the following function : (0 mod 4)

After applying this function, the final fingerprint code is: {72 8 88 72}

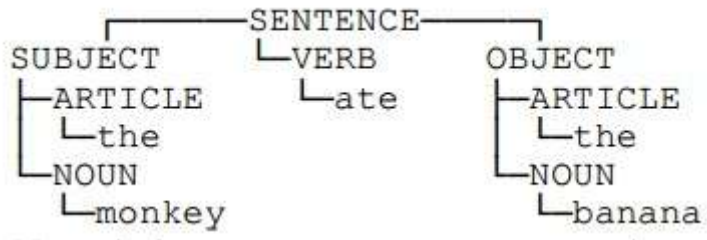
Then the algorithm generates the fingerprint for the whole origin and suspected file and then it compares between these fingerprints (LCS algorithm is most used algorithm for fingerprint matching[14]).

The most weakness point in fingerprint plagiarism detection methods in all versions that it strongly affected by words rearrangements and word Synonyms replacements.

4.2. String matching algorithms: which are based on matching between origin and suspected files strings, the most famous one is "Karp-Rabin Matching" algorithm; these types of algorithms have serious weakness point, which called "Split Match" Problem, which is setting the optimal minimal length of substrings to be matched between files[15].

The use of smaller "shortest string length to match" constant can effectively fight against swaps as well, but it also significantly increases the possibility of false matches. Furthermore, larger values of this constant make detection algorithm work faster [15].

4.3. Tree matching algorithms: these types of algorithms make new representation of the text using trees, like computer programs, natural language sentences have syntactic and semantic structure. There are software tools available that can be used to build parse trees for individual sentences. Most automatic English parsers use Chomsky-styled Penn Treebank grammars [15], based on the traditional linguistic approach to the syntax analysis, producing phrase structure-styled analyses. For example, the phrase "the monkey ate the banana" will be represented as[15]:



Natural language parsers can recognize noun phrases, homogeneous parts of the sentence, etc. It is clear that word swapping can occur, in particular, in sentences with conjunctions, such as "and", "or", "but", etc. For example, the phrase "I ate the pizza, the pasta and the donuts" can be reworded as "I ate the pasta, the donuts and the pizza". Instead of comparing sentences as word strings, we can first analyze them by a parser that recognizes the syntactic structure. These syntactically tagged structures normalize differences between sentences with the same proposition expressed with different word order, thus revealing potential plagiarism[15]. The most weakness point of these algorithms that generated from the ambiguity which exist in the natural language that this ambiguity leads that we have more than one tree representing the same sentence, and these types of algorithms are so weak in languages which have Diacritics to words(such as Arabic Language), that the absence of these Diacritics leads to fail these algorithms to generate the representing trees.

5. SEMANTIC PLAGIARISM DETECTION ALGORITHMS : the weakness points in traditional plagiarism detection algorithms make the research towards developing a new plagiraims detection algorithms that that try to overcome on these weakness points .

5.1.semantic plagiarism detection algorithms which based on semantic dictionaries :these types of algorithms use semantic dictionary as a resource to find the

relation between vocabulary of language , there is a famous semantic dictionary for English language called as "WordNet" , in this dictionary every word has its related words into set which called as synset of word, the main usage of these dictionaries is to find the words Synonyms **,**that when a person do the plagiarism operation he usually use word replacements with its Synonyms.

these algorithms usually use the following function to calculate the similarity between the origin text and the plagiarized text [16]

$$F_{q,k} = \begin{cases} 1 & \text{if } w_k \text{ and } w_q \text{ are identical} \\ 0.5 & \text{if } w_k \text{ is in the synset of } w_q \\ 0 & \text{otherwise} \end{cases}$$

That the function return the value /1/ if the two words are identical.

In addition, the function returns the value /0.5/ if the two words are in the same synset, this means that one word is Synonym to the other word.

In addition, the function return the value /0/ if the tow words are different.

Example: calculate the similarity between "the car consume oil" and "car consume petrol"[16].

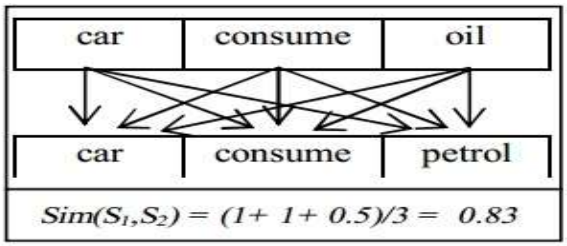


Fig (2): similarity calculate using synonyms [16]

In addition, there is another example in similarity calculation between the following two sentences

"The teacher gives each student a text that he authored"

"A textbook authored by the instructor is given to his pupils"

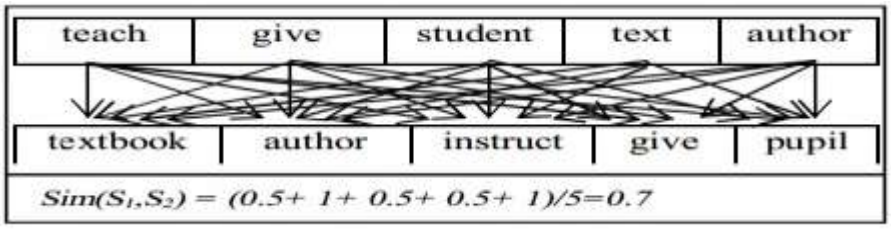


Fig (3): similarity calculate using the synonyms and different structure [16]

The most weakness point of these algorithms is the limitation of the size of the vocabulary of the semantic dictionaries.

5.2. Semantic plagiarism detection algorithms which based on semantic web languages:

these types of algorithms are based on the languages which describe the semantic web such as OWL,RDF, and these algorithms are based on the mapping between the ontologies of the origin files and the ontologies of the suspected ones ,as it shown In the following figure[17] :

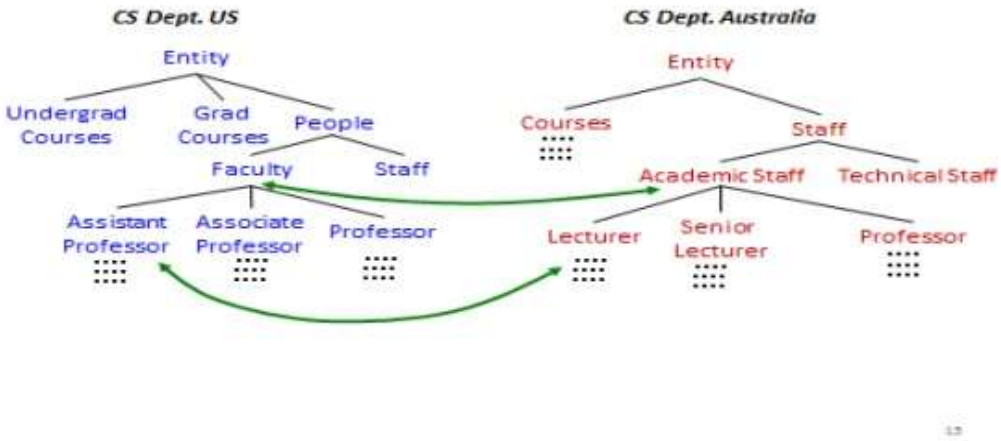


Fig (4): Ontology mapping Example [17]

and the following figure shows the general structure of these types of algorithms :

Fig (5): Architecture of detection system [17]

these algorithms as described in the figure above depends on learning the ontology for the origin documents, and from the suspected documents, and then it make the mapping stage between the resulting ontologies from the learning stage, The mapping between ontologies is done as described into the following figure:

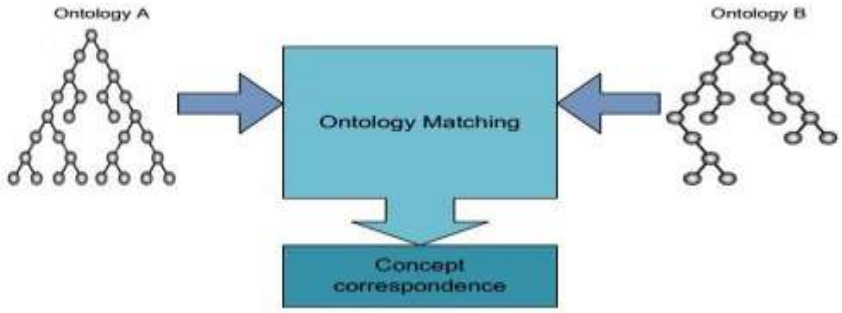


Fig (6): Ontology Matching [18]

The main weakness point of these algorithms is in the learning of ontologies from the web pages, that the automatic ontology learning process is not valid yet, and it requires a lot of manual validation to be valid 100%.

5.3. Semantic citation based plagiarism detection algorithms: these type of plagiarism detection methods are based on the analysis of the citations in the origin and suspected files ,as shown in the following figure[19] :

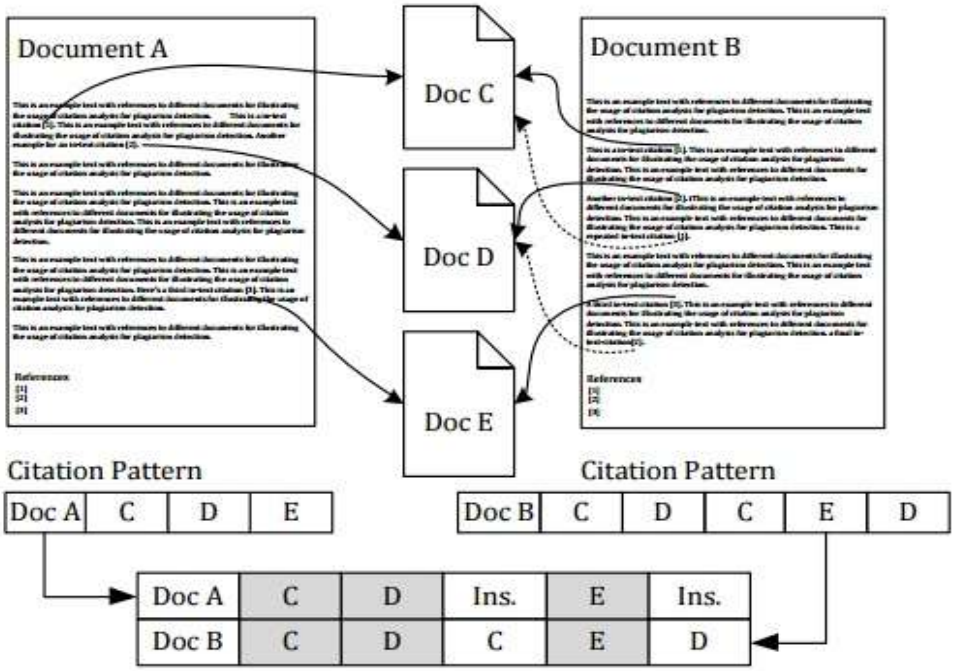


Fig (7): citation based plagiarism detection algorithms: [19]

These types of algorithms are language independent [20] that means the origin and suspected files maybe in tow different languages, the main weakness point in these algorithms that it cannot discover or detect plagiarism in the absence of citation marks in the origin and suspected texts parts.

5.4. Semantic plagiarism detection algorithms basing on concepts similarity [21]:

These algorithms for plagiarism detection composes of two phases: semantic analyzing and semantic comparison (each of them contains multi steps to finish). In this section, we will give a brief description of the algorithm.

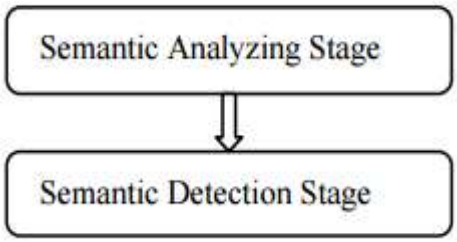


Fig (8): Algorithm main stages [21]

- **Semantic Analyzing:** the semantic stage is described in the following figure :

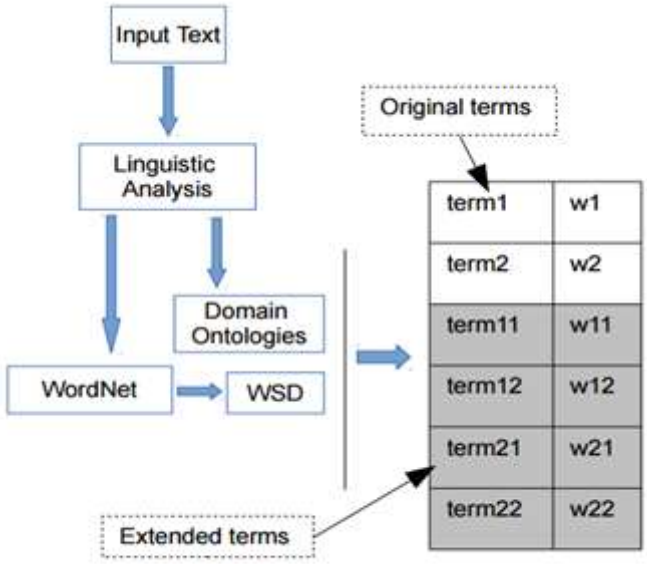


Fig (9): semantic analyses [21]

In this phase, the aim is to get semantic representation of the text using global semantic resources like WordNet ,and domain Ontologies. As domain ontology EMBLEBI an ontology lookup service was used, which provides a unified service for about 93 medical Ontologies [22], the biggest Ontologies in this set are the following ones:

- Gene Ontology (GO): GO ontology has developed three structured Ontologies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner [23].
- Infectious Disease Ontology (IDO): The IDO Ontologies are designed as a set of interoperable Ontologies that will together provide coverage of the infectious disease domain [24].
- Foundational Model of Anatomy (FMA): The FMA is reference ontology for the domain of anatomy. It represents the canonical, phenotypicstructure of the human body, spatial-structure and relations that characterize the Physical organization of the body at all salient levels of granularity [25]:
- Human Disease (DOID): The Disease Ontology has been developed as a standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms [26].
- Cell Type (CL): The Cell Ontology is designed as a structured controlled vocabulary for cell types. This ontology was constructed for use by the model organism and other bioinformatics databases [27]. A free Ontologies OWL files could be downloaded from the OBO Foundry [28].
- **Semantic comparison:** After extracting semantic representations of the origin text and the suspected text, semantic comparison goes through two stages: Domain comparison and similarity measure. Domain comparison ,It is very clear that if the two texts are from different domains, then any further investigations will be meaningless. In response, the following function was developed and used to compute domain closeness of two texts depending on their semantic representation, the function definition is described as :

$$F\left(SR_{t_1}, SR_{t_2}\right) = \max_{O \in domainontologies}^{M} \left\{ \sum_{c \in \left(O \cap SR_{t_1} \cap SR_{t_2}\right)} W_{SR_{t_1}}(c) * W_{SR_{t_2}}(c) \right\} \text{T}$$

hat: SR_{t_i} is a semantic representation of text I.

$W_{SR_{t_i}}(c)$ is a function returning the weight of concept c in SR_{t_i} .

$c \in \left(O \cap SR_{t_1} \cap SR_{t_2}\right)$ Are all shared concepts/terms between the two texts and domain ontology O.

Then if the tow files are in the same domain a Similarity Measure function is used to compute the semantic similarity between origin text and suspected text sentences ,this function is described as the following [29] :

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where: A: is the first terms vector.
B: is the second terms vector.
N: is the count of the shared terms

These types of plagiarism detection methods overcome to most of weakness point of traditional plagiarism detection methods, but these types have high complexity more than traditional plagiarism detection methods.

6. COMPARISON BETWEEN PLAGIARISM DETECTION ALGORITHMS

This table compares between weakness points, and features of plagiarism detection algorithms, the comparison includes the capacity of algorithms to detect plagiarism in cases of Words Synonyms replacements, Words order changements,and it includes The most important weakness point.

	Words Synonyms replacements	Words order changements	The most important weakness point
Fingerprint algorithms	**x**	**x**	Fail to detect plagiarism in words Synonyms replacements
String matching algorithms	**x**	**x**	Setting the optimal strings length when comparing
Tree matching algorithms	✔	✔	Generating more than one representing tree to the same text
	Words Synonyms replacements	Words order changements	The most important weakness point

Dictionary based semantic detection algorithms	✔	✔	The small size of the semantic dictionaries vocabularies
semantic detection algorithms based on semantic web languages	✔	✔	The automatic ontology learning which produce not valid ontologies
Citation based algorithms	**x**	**X**	Does not work in the absence of citation marks
semantic plagiarism detection algorithms basing on concepts expansion	✔	✔	The complexity of these algorithms

7. CONCLUSION : in this paper we present plagiarism definition , plagiarism detection systems, plagiarism reducing ,then we described the traditional plagiarism detection algorithms and the semantic plagiarism detection algorithms, and we have discussed the weakness points in every type of plagiarism detection algorithms, and finally we have compared between all plagiarism detection algorithms to know capacity of it to detect plagiarism in cases of Words Synonyms replacements and Words order changements, and this paper has shown that the semantic plagiarism detection algorithms are the most efficient detection algorithms, but it has complexity more than other plagiarism detection algorithms due the usage of semantic web resources in plagiarism detection.

REFERENCES

[1]. Plagiarism .[online] ,http://www.plagiarism.com, [Accessed 1-August 2015].

[2]. D. McCabe, "Research Report of the Center for Academic Integrity," 2005

[3]. J. J. G. Adeva, et al., "Applying plagiarism detection to engineering education," 2006, pp. 722-731.

[4]. C. Lyon, et al., "Plagiarism is easy, but also easy to detect," Plagiary: CrossDisciplinary Studies in Plagiarism, Fabrication, and Falsification, vol. 1, 2006.

[5]. L. Chao, L., et al., "GPLAG: detection of software plagiarism by program dependence graph analysis," the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006, ACM: Philadelphia, PA, USA.

[6]. C. J. Neill and G. Shanmuganthan, "A Web-enabled plagiarism detection tool." IT Professional, 2004.

[7]. M. Ginger and C. Christian, "K-gram based software birthmarks," the 2005 ACM symposium on applied computing. 2005, ACM: Santa Fe, New Mexico.

[8]. C. Hung-Chi, W. Jenq-Haur, and C. Chih-Yi, "Finding Event-Relevant Content from the Web Using a Near- Duplicate Detection Approach," the IEEE/ACM International Conference on Web Intelligence. 2007, IEEE Computer Society.

[9]. H. Dreher, "Automatic Conceptual Analysis for Plagiarism Detection." Issues in Informing Science and Information Technology, 2007.

[10].L. J. Edward, "Metrics based plagarism monitoring." Consortium for Computing Sciences in Colleges, 2001.

[11].Maxim Mozgovoy," Enhancing Computer-Aided Plagiarism Detection' , Department of Computer Science and Statistics University of Joensuu, University of Joensuu, Computer Science and Statistics, Dissertations 18 Joensuu, 2007, 131 pages

[12]. Y. HaCohen-Kerner, A. Tayeb, and N. Ben-Dror, "Detection of Simple Plagiarism in Computer Science Papers," *Proceedings of the 23rd International Conference on Computational Linguistics (Colinh 2010)*, no. August, pp. 421–429, 2010.

[13]. S. Schleimer, D. S. Wilkerson, and A. Aiken, "Winnowing: Local Algorithms for Document Fingerprinting," *Proceedings of the 2003 ACM SIGMOD international conference on on Management of data - SIGMOD '03*, pp. 76–85, 2003.

[14]. E. Perleka, "Erisa Perleka," Plagiarism Detection An Overview of Text Alignment Techniques. Norwegian University of Science and Technology July, 2013.

[15]. M. Mozgovoy, T. Kakkonen, and E. Sutinen, "Using natural language parsers in plagiarism detection," *Proceedings of SLaTE' 07 Workshop. Farmington. Pennsylvania*, pp. 7–9, 2007.

[16]. Salha Alzahrani and Naomie Salim 2010, " Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection - Lab Report for PAN at CLEF 2010 ", Available: http://clef2010.org/resources/proceedings/clef2010labs_submission_21.pdf ,2010.

[17]. M. Shenoy, K.C.Shet, and U. D. Acharya, "Semantic Plagiarism Detection System Using Ontology Mapping," *Advanced Computing*, vol. 3, no. 3, pp. 59–62, 2012.

[18]. M. M. Taye, "Ontology alignment mechanisms for improving web-based searching," De Montfort University United Kingdom, England 2009.

[19]. B. Gipp, N. Meuschke, C. Breitinger, M. Lipinski, and A. Nuernberger, "Demonstration of Citation Pattern Analysis for Plagiarism Detection," in Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, UK, 2013.

[20]. plagiarismtoday. [online] Available at: http://www.sciplore.org/projects/citation-based-plagiarism-detection/ [Accessed 25-may 2015].

[21].Omar, K., Alkhatib, B., Dashash, M.,and Alhassan, F. (2016), - "TheImplementation of using Medical Ontologies in Plagiarism Detection",International Review on Computers & Software, Vol. 11 No. 3.

[22].EMBL-EBI, Ontology Lookup Service.[online] http://www.ebi. ac.uk/ontology-lookup/init.do#soft [Accessed 15-August 2015].

[23].Gene Ontology Consortium, [online] Available at: ,http://geneontology.org/ [Accessed 1-March 2016].

[24].The Infectious Disease Ontology, [online] Available at: http://infectiousdiseaseontology.org/page/Main_Page[Accessed 1- March 2016]

[25].Ontology Design Patterns .org(ODP), [online] Available at: http://ontologydesignpatterns.org/wiki/Ontology:Foundational_M odel_of_Anatomy_(FMA) [Accessed 1-March 2016].

[26].Disease Ontology, [online] Available at: http://diseaseontology.org/[Accessed 1-March 2016].

[27].Cell Ontology, [online] Available at: http://obofoundry. org/ontology/cl.html [Accessed 1-March 2016].

[28].The OBO Foundry. [online] Available at: http://www.obofoundry.org/ [Accessed 15-March 2015].

[29]. Wikipedia, Cosine similarity.[online] https://en.wikipedia.org/wiki/Cosine_similarity / [Accessed 1-August 2015].