Plagiarism and Source Deception Detection Based on Syntax Analysis

Assistant Prof. Dr. Eman Salih Al-Shamery

Department of software, College of Information Technology, Babylon University, Babil, Iraq.
emanalshamery@yahoo.com

Hadeel Qasim Gheni ALkhafaji
Department of software, College of Information Technology, Babylon University, Babil, Iraq.
hadeelqasem84@Gmail.com

Abstract

In this research, the shingle algorithm with Jaccard method are employed as a new approach to detect deception in sources in addition to detect plagiarism . Source deception occurs as a result of taking a particular text from a source and relative it to another source, while plagiarism occurs in the documents as a result of taking part or all of the text belong to another research, this approach is based on Shingle algorithm with Jaccard coefficient , Shingling is an efficient way to compare the set of shingle in the files that contain text which are used as a feature to measure the syntactic similarity of the documents and it will work with Jaccard coefficient that measures similarity between sample sets . In this proposed system, text will be checked whether it contains syntax plagiarism or not and gives a percentage of similarity with other documents , As well as research sources will be checked to detect deception in source , by matching it with available sources from Turnitin report of the same research by using shingle algorithm with Jaccard coefficient. The motivations of this work is to discovery of literary thefts that occur on the researches , especially what students are doing in their researches , also discover the deception that occurs in the sources.

الخلاصة

في هذا البحث، خوارزمية التسقيف وطريقة Jaccard تم استخدامهما كطريقه جديدة للكشف عن الخداع بالمصادر بالإضافة الى الكشف عن السرقة الادبية. الخداع بالمصدر يحدث نتيجة لأخذ نص معين من مصدر ونسبه الى مصدر اخر، بينما السرقة الادبية تحدث في الوثائق نتيجة لأخذ جزء أو كل من النص التابع لبحث آخر، هذه الطريقة تستند على خوارزمية التسقيف مع معامل Jaccard ، التسقيف هي طريقه فعالة لمقارنة مجاميع التسقيف في الملفات التي تحتوي على نص والتي تستخدم كميزة لقياس التشابه النحوي للوثائق والتي ستعمل مع معامل Jaccard الذي سيقيس التشابه بين المجاميع العينة . في هذا النظام المقترح ، سيتم فحص النص فيما اذا كان يحتوي على سرقة ادبية نصية أم لا، وإعطاء نسبة مئوية للتشابه مع غيرها من الوثائق، وكذلك سيتم فحص مصادر البحث لكشف الخداع بالمصدر بواسطة مطابقتها مع المصادر المتوفرة من تقرير Turnitin لنفس البحث باستخدام خوارزميه التسقيف ومعامل Jaccard. دوافع هذا العمل هو اكتشاف السرقات الادبية التي تحدث على البحوث وخصوصا ما يفعله الطلاب في بحوثهم ، وكذلك اكتشاف الخداع الذي يحدث في المصادر .

Keywords: Source deception , Syntax plagiarism , Plagiarism detection , Shingling algorithm, Jaccard coefficient.

الكلمات المفتاحية : خداع المصدر، السرقة الادبية النصية ، كشف السرقة الادبية، خوارزمية التسقيف ، معامل Jaccard

1. Introduction

Recently plagiarism problem increased because of the digital age to the sources available on the internet [Salha Alzahrani , 2011]. Plagiarism can be more vague than a clear , could be more complex than copying and pasting .[Salha Alzahrani1, Naomie Salim , 2010]

Internet allows the students to find numerous examples of programming source code, so it is not hard for these resources to be a resource of plagiarism. Also, because it has become more commonly for students to edit or provide training electronically, It has become easier for the student to copy the work of another student and give it as his / her own work . [Asako Ohno and Hajime Murao , 2010]

So the reality is that the internet is visible treasure of information containing all the threads known to humans . [Amit Prakash, Sujan kumar Saha , 2014]

There is no protection law to protect against scientific fraud , stress factor for researcher that generates as a result of lack of time to do the work required , as well as the researcher culture for copyright and scientific work leads to plagiarism, therefore , Plagiarism is defined as taking or attempting to take part or all of someone else's work without reference to the source and considered him as the author of that work. [Asim M. El Tahir Ali , 2011] [Demetrios Glinos , 2014].

Plagiarism is used in the strict sense to refer to the texts of documents were copied and pasted directly from another source without permission approval, this is unacceptable because even though it was pointed to the source, it will be considered plagiarism , while change or modify some of the words belonging to the original text, taken from researches, magazines, books, newspapers or even ideas considered acceptable, this what is called paraphrasing .

Plagiarism has several types , including Direct Plagiarism which means word-for-word transcription of a part of someone else's work without attribution and without the quotes . Self-Plagiarism which means re-use of one's own prior research or use parts of it without recognizing that the piece was citing from original research , concealing sources which means taking ideas from a person and put them in your words without reference to the source , or using the same source several times but referred to it only once .

In this research, shingling method is used , that has major impact on the conduct of detective work because it is very efficient in reducing the amount of non-related files, and effective to find documents that are similar to the files, which have already been identified as related. shingling (continuous sequence of symbols in the document) has a significant improvement in the efficiency of a decision on the similarity level of two documents in a number of joint shinglings. [Dariusz Ceglarek , 2012]

Shingling method has been used in several ways in different researches, but employed here with Turnitin report . Size of shingle set must be more than one word, take into consideration the length of the document and the ideal amount of the word is commonly used in the document. They should choose the size to be large enough to ensure a low probability accidental similarity in the specified document(s) [Andre Ross , 2014].

The similarity of two shingle sets can then be determined by Jaccard coefficient , Jaccard coefficient is one of the common metrics that used to measure the overlap of two groups [Paul Ginsparg , 2011] .

2. Proposed method

The proposed system has been developed in three main phases. The first is the pre-processing phase which includes tokenization, stop words removing. The second phase is a list of candidate documents retrieved for the input document using shingle with Jaccard coefficient. The third phase is the matching between sources of research and Turnitin report for the same research using unification and sifting operations. The proposed system has been shown in the figure below :

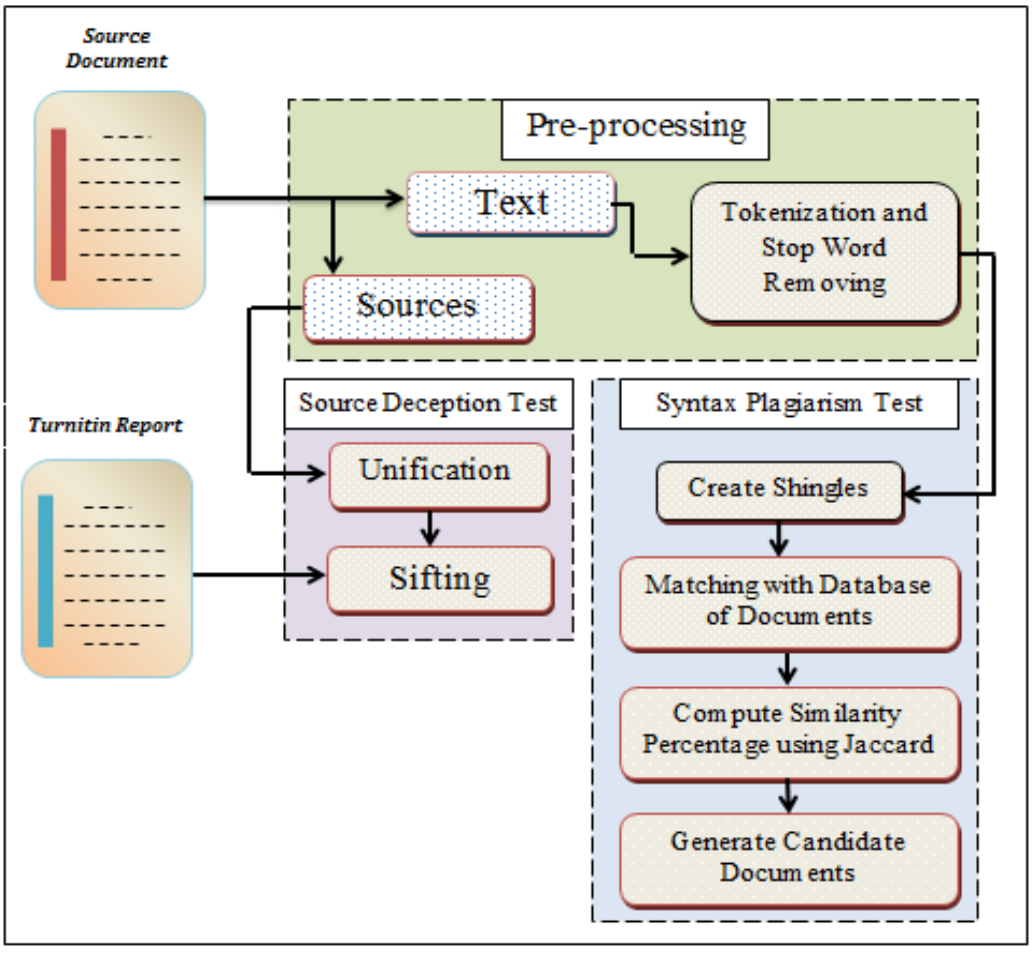


Figure 1. Architectural proposed system to detect source deception and syntax plagiarism .

2.1 Pre-processing Step : The input document will split into text and sources.

- ✓ Tokenization and stop word removing : here , the text of the document that consists of paragraphs will be divided into set of tokens in a process called tokenization , remove blanks , punctuation and stop word, puts everything in lower case .

2.2 Syntax Plagiarism Test :

A. Create shingles : In this step, each paragraph will be divided into small groups of words based on a key specified within the program , this key determines the number of words in each group, which is called shingle sets . The key must be appropriate to the length of the document, if the document is short then the key length could be 3 or 4 but if the length of the document is too long, the key length could reach to 10.

B. Matching with Database of documents : now the input document will matching with documents pre-stored in database table, each document in the database will pulling and segmenting into shingling sets with the same key that used to grouped the input document , then matching all the sets in both documents , this process will repeat for all documents in the database.
Matching process occurs on every word exist in the shingle sets, this meaning, it compares the words of the first shingle set with the rest of sets and so on, before entering this stage, all the characters have been converted to lowercase to prevent the case of similarity of two words, one with lowercase letter , while the other with uppercase letter.

C. Compute similarity percentage using Jaccard : After matching process , the percentage of similarity between the shingle sets will calculate according to Jaccard coefficient .
The Jaccard similarity is defined as : [Jeff M. Phillips, 2013]
$$Jaccard(A,B) = |A \cap B| / |A \cup B| \quad \text{..... equation(1)}$$
This means, the number of similarities to the total number in the document . The result ranges from 0 (with no elements in common) to 1 (identical elements).

D. Generate candidate documents : The last step in this stage is to find candidate documents that be related to the input document, this means , find the documents that have texts found in the input document , thus , a syntax plagiarism has been found.

2.3 Source Deception Test :

A. Unification step : In this step , research sources written in any format will unite and stored in a database table .

B. Turnitin report : also , Turnitin report of the sources to the same research will saved to PDF , convert it to text then stored in a new table in the database .

C. Sifting : after that, a matching process will be done between the sources of the research and Turnitin report, then find the similarity percentage of the matching and are they identical or not ?

3. Algorithms

A. Algorithm1 : Shingle algorithm with Jaccard coefficient for detecting syntax plagiarism is shown below:

***Name** : Shingle Algorithm with Jaccard Coefficient*
***Input** : Document file .*
***Output** : Candidate similar documents .*
Begin
***Step1 :** Extract the plain text from the input document*
***Step2 :** Removes all characters except letters and digits and puts everything in lower case*
***Step3 :** Splits the text into tokens(overlapping groups of words)*
***Step4 :** Examine Database Documents*
1) *For I= first document to the last document in database table Splits the text into tokens*
2) *Compares the sets of shingles generated from the input document and database documents using Jaccard coefficient to check how similar two document are .*
3) *If the result = 0 then*
No elements in common
else
There identical elements
End if
***Step5 :** Extract the candidate document that related to the input document*
End

B. Algorithm2 : Source Deception algorithm is shown below :

***Name** : Sources Deception*
***Input** : Document file .*
***Output** : Detecting Source Deception*
Begin
***Step1 :** Extract the sources from the input document*
***Step2 :** Unite the sources and store it in a database table*
***Step3 :** Store Turnitin report in another database table*
***Step4 :** Compare both sources by Sifting process and compute the similarity*
If the result = 0 then
Both Sources are Identical
else
There is Source Deception
End if
End

4. Experimental Result

The program has been built using Java NetBeans IDE 8.0.2 , The results of proposed system in this research have been applied to many of files , the size of each file can be between 1 – 50 page , whatever the type of file , whether Word or PDF, it will be converted to text then it will be treated according to the proposed system . The database was built in MySQL workbench program and has a storage capacity of nearly 500,000 files, which is stored currently nearly 100 file, This database has two tables , one for storing the researches while the second is for storing the research sources .

When building the database several limitation was encountered including :

First : the files stored in the database, a lot of research published in international journals cannot be obtained free of charge, and so we cannot take advantage of that research to detect plagiarism, the more the large size of the database that contains a lot of research will be more accurately for detect plagiarism .

Second : converting research formula from PDF or Word format to text format , in this research, research needs to change from formula to another but always we find there are problems in the changes process , for example, the end of each line in the original search will be considered as paragraph after change formula process and these are a problems at work and must work order of each search after changes.

Researches usually contain images may be for algorithms where the researcher put a complete algorithm as image, in the plagiarism detection process, these images are not comparable because it lies under the topic of image processing, in this case only the image address will match , while tables will match naturally .

In this research , work was based on shingle algorithm and Jaccard coefficient , in source deception detection process, shingle algorithm will partitioning the sources of research and Turnitin report to sets based on key , then this sets will be subjected to matching test between them and finally compute the percentage of matching using Jaccard coefficient according to equation (1) . Figure (2) shows the percentage of matching 50 document sources and the value chosen for the key is 3 .

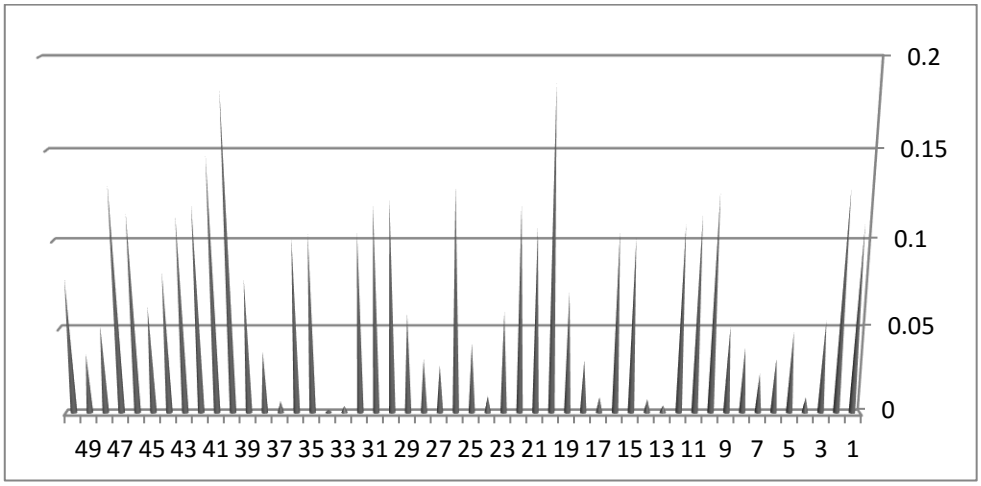


Figure (2) : Source deception detection process for 50 documents .

In figure (2), high percentages represent many numbers of matches between research sources and Turnitin sources, and vice versa, where the few percentages represent few numbers of matches . The time it takes to complete the matching process for one research sources when key = 3 is one second .

the same process of shingle and Jaccard will repeated in syntax plagiarism detection process but in the end, the result of Jaccard equation will be subject to testing a certain threshold , a threshold value of Jaccard= 0.05 was used to filter out non – candidate documents , If the percentage of similarity exceeded this threshold then there will be plagiarism . Figure (3) shows the percentage of matching one document with a sample containing 100 documents stored in the database and the value chosen for the key is 3 .

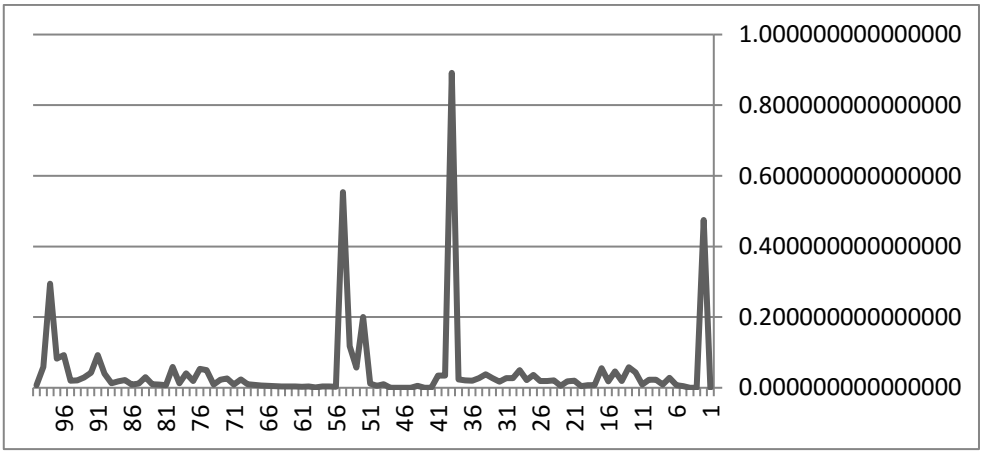
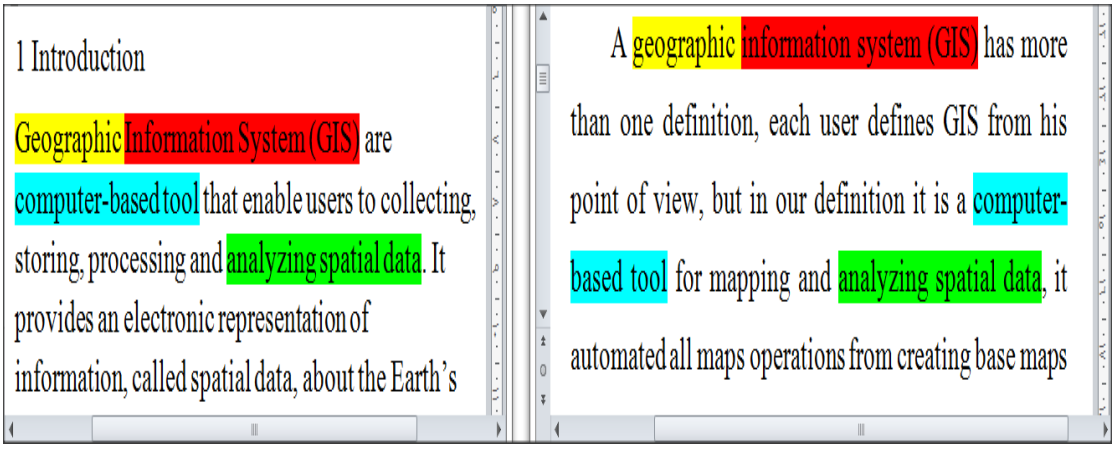


Figure (3) : Syntax plagiarism detection process for one document .

In the figure above, high percentages represent many numbers of matches between the input document with database documents, and vice versa, where the few percentages represent few numbers of matches . the time it takes to complete the matching process shown above when key = 3 is 25 seconds .

The following figure shows a text of a research has been compared with research stored in the database, similar text has been misled only to illustrate them, knowing that the key = 3.



(A) (B)

Figure (4) : Comparing two researches .

In figure (4) , (A) represent a text of one of the researches stored in the database, while (B) represent the entrance research text. accordingly to figure (4) and base on the key, the similar texts are illustrated in the following table :

Research (A)	Research (B)
geographic information system	geographic information system
information system (GIS)	information system (GIS)
computer-based tool	computer-based tool
analyzing spatial data	analyzing spatial data

Figure (5) : Similar texts

Of course, the result gained when key = 3 is differ whether = 4 or 5 and so on.

5. Conclusion

This paper, describes a new approach to detect source deception and syntax plagiarism which occurs in researches by using shingle algorithm and Jaccard coefficient, In this approach, the conclusion was that, the shingling algorithm has proven as an effective way to identify the deception in sources and also an efficient way in matching texts to identify plagiarism, size of shingle set effect on the time it takes in the decision-making process, and this approach is different from Turnitin only in the amount of data that can be accessed by Turnitin .

In the future , hoping to enhance the method via using another approach to detect the Semantic Plagiarism which is called paraphrasing, and that does not mean taking the text directly, but rather to change part of it which is acceptable to some extent.

6. References

Amit Prakash & Sujan kumar Saha . (2014) . " Experiments on Document Chunking and Query Formation for Plagiarism Source Retrieval" . Department of Computer Science and Engineering, Birla Institute of Technology, India .

Andre Ross . March 26 2013 . Identifying similar documents . http://digfor.blogspot.com/2013/03/fruity-shingles.html

Asako Ohno & Hajime Murao . (2010) . "A Two-Step in-Class Source Code Plagiarism Detection Method Utilizing Improved CM Algorithm and SIM " . Department of Life Design Shijonawate Gakuen Junior College , Graduate School of Intercultural Studies Kobe University , Japan

Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla & Vaclav Snasel . (2011) . "Overview and Comparison of Plagiarism Detection Tools ". Department of Computer Science, V_SB-Technical University of Ostrava .

Dariusz Ceglarek . (2012) . " Evaluation of the Shapd2 Algorithm Efficiency in Plagiarism Detection Task using PAN Plagiarism Corpus" . Department of Applied Informatics, Poznan School of Banking, Poznan, Poland

Demetrios Glinos . (2014) . " A Hybrid Architecture for Plagiarism Detection". Computer Science, University of Central Florida, Orlando, Florida, United States .

Jeff M. Phillips . (2013) . " Jaccard Similarity and Shingling" . University of Utah .

Paul Ginsparg . (3 Nov 2011) . Slides adapted from Hinrich Schütze's , Linked from http:// informationretrieval.org/ . Cornell University, Ithaca, NY.

Salha Alzahrani, Naomie Salim & Ajith Abraham . (preprint 2011). Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods.

Salha Alzahrani & Naomie Salim . (2010) . "Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection"