Syntactic and Sementic Based Similarity Measurenent for Plagiarism Detection

Sumathi S, Geetha M P, P Ganesh Kumar, K Pushpalatha, A S Shanthi

Abstract: ***In the world of digital era, there is a high availability of huge amount of online documents which leads to plagiarism. Plagiarism is the act of copying other person work. The paper based documents are stored in the digital libraries for future references. In the olden days, people used the Latin word "plagiarius" to indicate the act of stealing someone else work. Plagiarism is the act of using one's ideas, concepts, words or structures without citing their references where original work is expected from the users. In this paper, the main objective is to compare the contents of original document that matches with the contents in other documents. These matches are considered depending on the syntactic matches and also the semantic similarity. This paper employs Sentence Hashing Algorithm for Plagiarism Detection focusing on complete sentence sequences and calculates hash – sum for the sentence sequences. When the user compares the original document to several documents, if the similarity value of the document is 1, then the contents present in the original document is 100% same in the compared documents, i.e., fully plagiarized. If the similarity value varies from 0.1 to 0.9, then it is partially plagiarized. The similarity value is 0%, then the original document is unique.***

Keywords: Plagiarism Detection, Syntactic and Semantic based similarity, Sentence Hashing, Text Mining

I. INTRODUCTION

In this electronic world era, end users select to accumulate their files in the form of digital documents relatively in the form of paper-based documents (i.e. digital libraries). The process to store, examine, and share documents has greatly impacted by the usage of the internet. We are in the digital era where the information is available in abundance; it is really a complex task to explore the original author. It's easy to locate the plagiarized text in the digital society. Internet contains processed available texts, people are taking the advantage of using the copy and paste method, there are many websites which manuscripts are accessible, this site is preferably appropriate for the plagiarists to exchange the data in various formats (text, audio, video, images) using any tools. In the 1st century, people used the Latin word "plagiarius" to indicate the work of stealing somebody's work. According to Bela Gipp academic, plagiarism is referred as the illegal application of thoughts, contents, text or organization of concepts without suitably granting the cause to gain in a situation where innovation is estimated.

Plagiarism is a deed of stealing or bidding to acquire response or value for a systematic study. The action of depriving other innovator's text and the representing the contents as their own content comprise breaching of copyrights and is a serious abuse of the principles of learning. Rewording somebody's texts by exchanging a few words by synonyms or substituting some sentences in own way are also plagiarism. Even replicating in your own words, thinking or investigation made by someone else may constitute plagiarism. The same still relates to if you bring together bits of work by various authors without citing the sources. In a conventional methodology, the only approach to detect plagiarism was to manually inspecting the whole document. Each manuscript must be examined manually by a domain experts to resolve if it is plagiarized or not, this analysis can is quite slow to process so there is a necessity to devise a plagiarism detector. Plagiarism detection techniques are useful by making a difference between natural and programming languages. A similarity score is dogged for each pair of documents which match considerably. In traditional methods, to detect plagiarism will be focused on words matching but it won't detect the plagiarism automatically by syntactic and semantic based measurements.

The main objective of the proposed methodology is to identify plagiarism to safeguard the intellectual properties of each document. The most commonly used plagiarism techniques comprises of rearranging the words in imitative text, by changing and replacing the paragraphs and passages of original document with the words providing similar meaning and phrases.

Nevertheless, all of the specified techniques cannot be detected with its similarity in an instance, but synonym techniques can be easily detected.

A. Syntactic Analysis

Syntactic analysis is the method of examining a group of codes in normal language or in computer languages in compliance with the regulations of a recognized grammar. Syntactic analysis of a sentence is also referring to the process of distinguishing a sentence and conveying a syntactic structure to it. These syntactic structures are assigned by the context free grammar using parsing algorithms like cocke-kasami-younger (CYK), early algorithm, and chart parser. They are represented in

Revised Manuscript Received on December 05, 2019.

***** Correspondence Author

Sumathi S*, Assistant Professor (Senior Grade), Department of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu, India. Email: ssumathibe@gmail.com

Geetha M.P, Assistant Professor, Department of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu, India. Email: geetha.cse@srit.org

Dr P Ganesh Kumar, Assistant Professor, Department of Information Technology, Anna University Regional Center, Coimbatore, Tamilnadu, India Email: ganesh23508@gmail.com

Dr K Pushpalatha, Assistant Professor, Department of Information Technology, Coimbatore Institute of Engineering and Technology, Coimbatore.

Dr A S Shanthi , Associate Professor, Department of Computer Science and Engineering, Tamilnadu College of Engineering, Coimbatore.

a tree structure. These parse trees serves as an intermediate stage of representation for semantic analysis. In computational linguistics, the term syntactic analysis refers to the correct analysis done by a computer or other sequence of words. Output generated by the syntactic analysis is the syntax tree representing the syntactic correlation among the words.

B. Semantic Analysis

In computational linguistics, relation of the syntactic structures among words, sequence of text, sentences is identified by using the semantic analysis. Semantic analysis is the process of finding the correlation among the words.

II. LITERATURE SURVEY

Alexandr Andoni and Piotr Indyk [1] proposed Near – Optimal Hashing Algorithms where the objects are represented as the points in the d-dimensional space. Andrei Z. Broder et al.,[2] proposed Syntactic clustering of the Web, where the user have established an efficient way to examine the syntactic similarity of files and have smeared it to every document on the World Wide Web. By using this suitable mechanism, we built a clustering of all the syntactically similar documents. Steven Burrows et al.,[3] stated that, well organized identification of breach of copyright for huge code warehouse. In most of the educational organization illegal reuse of code is a common problem. Detection of breach of copyright manually is highly complicated and plagiarism detection methods currently available are not applicable for huge size of code repositories. He also added methods for exploring the similarity in source code using content or word similarity measures and local alignment. Dariusz Ceglarek et al.,[4] proposed that semantic compression is viable concept for English language. More accurate results can be produced by applying WiSENet information gaining method.

III. METHODOLOGY

A. Document Extraction

Extraction of the document is done by the system, only after the user select his/her desired file and its format. It is the primary process used to read or extract content from the selected document that helps the user to identify whether the document is empty or not. The extracted content from the original document and the comparative document is given as the input for the preprocessing step where the removal of the stop words takes place.

The stop words are the unnecessary words (i.e. removal of these words will not change the meaning of the sentence.) in the document that are removed and remaining main words are stored in the database for later and future use.

B. Similarity Words Reduction

Reduction of the similar words is used to reduce the redundant information occurs in the files (original document and comparative documents). It is used to eliminate the same redundant content has been inserted into the data set more than once. It also used to reduce the quantity of words involved in the comparative operation. Dataset is the collection of words which is obtained after the elimination of the stop words and decrease of the similar texts, which are stored in the database for next process.

C. Clustering

Clustering is the method of grouping of similar words into a single group or single package that helps in easy identification of similar words in the selected document (It may be original or comparative document). Clustering is used to increase the speediness of finding the similar words in the document which helps to measure the similarity value of the documents. The words are grouped into different groups based on content in the respective selected documents. The names of the groups are original document and comparative document (names of the documents) which contains respective cluster (group of similar words) based on their content in the original and comparative document.

D. Similarity Calculation

Similarity Calculation is the final step to calculate the similarity value of the content that is present in the original and comparative document. Our goal is to find whether the same content in the original document is also present in the comparative document, if so then percentage of similarity value is calculated.

The most important part is the generation of similarity values which is mandatory to identify the similarity of files. If the similarity value of the document is1, when original document is compared to several document then the content present in the original is 100% same in the comparative document (i.e. the original document is copied from other documents). If the percentage value changes from 0.1 to 0.9 % then based on percentage the content in the original document is copied from other documents. If the similarity value is 0%, then the original document is unique (not copied from any other documents).

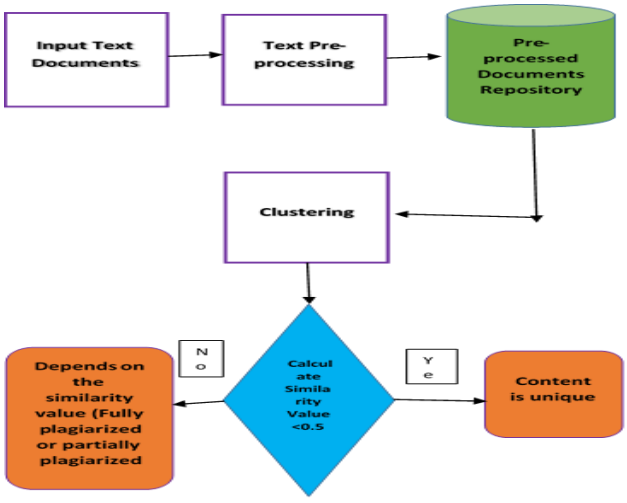

Fig. 1. Proposed System

IV. ISHAPD2 ALGORITHM

improved Sentence Hashing Technique for Plagiarism Detection3 (iSHTPD3) Algorithm emphases on entire sentence sequences and calculates hash-sums for the sentence sequences. Hash – index is arranged in order by applying a new technique and then searching the hash – index is implemented. Correspondence List (CL) is utilized in the method of indexing.

The (iSHTPD3) technique comprises of two sets of input documents to work with: i) a set of input original source files F= $\{f_1, f_2,\ldots..f_n\}$ ii) a set of doubtful files which has to be compared with the original files for similarity, S = $\{s_1,s_2\ldots s_m\}$

Primary step of this algorithm is to do pre processing of each file in which text enhancement process is implemented. Preprocessing of files starts with identification of lexical units called tokenization. Next step in preprocessing comprises of removal of stop word which is also called as delimiter, detection of multi word concepts and fetching back to the root word called stemming or lemmatization

For certain languages such as Polish, French preprocessing of text document is comparatively a huge process. Final step in this procedure is the concept of removing ambiguity. The system finally generates an output vector representing arranged list of concept descriptors deduced from the files. Here, for all the set of files F, a hash table H is generated, the following function (3.1) is used to store the values in the index

$$H\,[k_{i,\,i}] = < i,\, j > \qquad (3.1)$$

Similarly for all files in the suspected document database S, the hash values are generated. Finally all the files in the suspected documents S and the original source files are denoted as the hash values altogether.

It is noteworthy, that (iSHTPD3) Algorithm is capable to implement and generates the more similar results as the output in a short time.

V. RESULTS AND DISCUSSIONS

The research work has been implemented with the aim of producing similarity value between the original and comparative documents using iSHTPD3 algorithm.

Extraction of the document is done by the system, only after the user select his/her desired file and its format. It is the primary process used to read or extract content from the selected document that helps the user to identify whether the document is empty or not and it is the input for the preprocessing step stop words removal. Reduction of the similar words is used to reduce the redundant information occurs in the files. It also used to reduce the amount of text involved in the comparative operation. Clustering is the method of grouping of similar text into a single group or single package that helps in easy identification of similar words in the selected document (It may be original or comparative document).

Clustering is used to quicken the process of finding the similar words in the document which helps to measure the similarity value of the documents. Similarity Calculation is the final step to calculate the similarity value of the content that is present in the original and comparative document. Our goal is finding whether the content in the original document is same or not. The most important part is the generation of similarity values which is mandatory to identify the similarity of files. If the similarity value of the document is1, when original document is compared to several document then the content present in the original is 100% same in the comparative document (i.e. the original document is copied from other documents). If the percentage value changes from 0.1 to 0.9 % then based on percentage the content in the original document is copied from other documents. If the similarity value is 0%, then the original document is unique (not copied from any other documents).

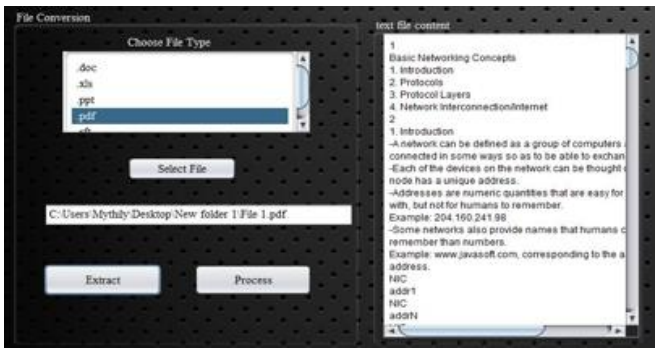

Fig.1 File Transformation

The above Fig 1 describes the next page of the plagiarism detection tool where the user can choose their desired format (.ppt, .pdf, .doc, .xls) of the input file (the original i.e. file which is need to be uploaded and check for plagiarism) under choose file type drop-down box. The select file option in the middle screen helps to choose the file from the local machine (user machine). The text box below the select file option will display the path of the selected file in the screen for user verification and confirmation. In case if the user chooses the wrong file then it can be rectified by choosing the correct file by using select file option again. The Extract button in the bottom left corner of the screen is used to extract files (i.e. read the content of the file in the desired format) and display the content of the file in the left side text box (text file content) in the screen where user can cross verify the content of the file and can change the file using select file option again, the process button in the right bottom corner of the screen will process the file to the next stage and redirect page to next page.

Fig.2 Stopword Removal from input documents

The above Fig 2 describes the stop word removal process. Stop word removal process is used to remove unwanted things like articles, preposition from the input (original) document. Input file textbox shows the selected file name and its path. Remove stop words button in the left side of the screen helps to remove the stop words from the file and those words are displayed in the small screen of the page. Store stop words button is used to store words, which is listed in the screen into the database. If insertion of words in the database is successful, the dialog box with the message "Data is successfully inserted into the database" is popped in the screen by clicking the ok button the dialog box disappears from the screen.

If data is not inserted successfully in the database, the dialog box with the message "Data is not successfully inserted into the database" is popped in the screen. Data are stored in the database with the help of JDBC connection between the front end (UI-User Interface) and backend (database) using MYSQL and Wamp. Proceed button in the right side of the screen help user to navigate to next page.

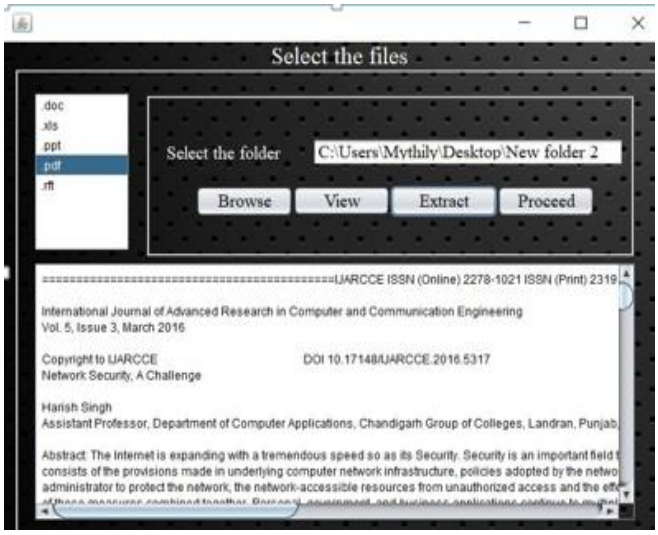

Fig.3 Comparative Document Selection

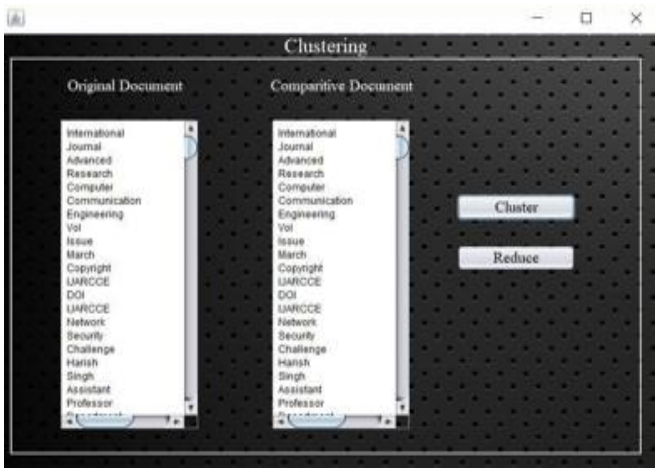

Fig 4 Clustering

The above Fig 4 describes the clustering process in the plagiarism detection tool which using iSHTPD3 algorithm for similarity detection among documents. Clustering is the method of grouping of similar words into a single group or single package that helps in easy identification of similar words in the selected document (It may be original or comparative document). Clustering is used to increase the process speed in finding the similar words in the document which helps to measure the similarity value of the documents. The words are grouped into different groups based on content in the respective selected documents.

The names of the groups are original document and comparative document (names of the documents) which contains respective cluster (group of similar words) based on their content in the original and comparative document. By clicking the cluster button, the clustering process is done and the results are displayed in the screen. Original documents contain clustered word based on words in the original document and comparative documents contain clustered words based on words in the comparative documents.

Fig 5 Similarity Results

The above Fig 5 describes Similarity Calculation which is the final step to calculate the similarity value of the content that is present in the original and the comparative document. If the similar content is found, based on it the percentage of similarity value is calculated. Similarity values are very effective to disclose whether the information in the document is identical or not. If the similarity value of the document is1, when original document is compared to several document then the content present in the original is 100% same in the comparative document (i.e the original document is copied from other documents). If the percentage value changes from 0.1 to 0.9 % then based on percentage the content in the original document is copied from other documents. If the similarity value is 0%, then the original document is unique (not copied from any other documents).After clicking the result button in the screen, the page is displayed where the file name text box contains all files name which the input (original) document is compared.

Similarity textbox contains the similarity value of the input file which matches with the files in the dataset (comparative documents). Similarity displays the similarity measurement in the range 0 to 1. Result textbox contains the similarity value in percentage.

VI. CONCLUSION

The proposed work helps everyone to check whether their document is plagiarized or not using improved Sentence Hashing Algorithm for Plagiarism Detection2 (iSHTPD3) algorithm that gives more accurate result than w-shingling algorithm. iSHTPD3 produces similarity value in terms of percentage which is easy for user understanding.

With the help of syntactic and semantic based similarity measurement user can easily understand how much percentage the content is similar when compare to contents in the other documents. It also helps students to escape from most extreme penalties like academic suspension; violation of rules that may affect degree completion if plagiarism is detected in their documents is high.

In this proposed work, user has to choose the selective domain and selective documents to which the original document is checked for similarity content.

Table- I: Similarity Results

Original Document	Comparative Document	Similarity	Results
Source Document	Comparative Document 1	1	100% Same
	Comparative Document 2	0.12614138	10% Same
	Comparative Document 3	0.12741111	10% Same

In the future, user can check his/her documents for similarity in the entire domain not only in a selected or particular domain. And, user can extend the algorithm in an effective way so that it can compare n number of documents and produce more accurate results**.**

REFERENCES

1. A. Andoni and P. Indyk, "Near-optimal hashing algorithms for a approximate nearest neighbor in high dimensions", Communication of the ACM, 51(1), pp. 117-122, 2008.
2. A. Z. Broder, "Syntactic clustering of the web", Computer Networking ISDN Systems, vol. 29 (8-13), pp. 1157-1166, 1997.
3. S. Burrows, S.M. Tahaghoghi and J. Zobel, "Efficient plagiarism detection for large code repositories", Software: Practice and Experience, vol. 37 (2), pp. 151-175, 2007.
4. D. Ceglarek, K. Haniewicz and W. Rutkowski, "Towards knowledge acquisition with WiseNet", vol. 351 of Studies in Computational Intelligence, pp. 7584, Springer Verlag, Berlin, 2011.
5. N. T. Nguyen, R Katarzyniak and S.M. Chen (eds.), "Semantic compression for specialised information retrieval systems", Advances in Intelligent Information and Database Systems, vol. 283 of Studies in Computational Intelligence, pp. 7584, Springer Verlag, Berlin, 2011.
6. K. Erk, and S. Pado, "A Structured Vector Space Model for Word Meaning in Context", pp. 897-906, ACL, 2008.
7. C. Grozea, Ch. Gehl and M. Popescu, "Encoplot: Pairwise sequence matching in linear time applied to plagiarism detection", Time, pp. 10-18, 2009.
8. R. Lukashenko, V. Graudina and T. Grundspenkis, "Computer-based plagiarism detection methods and tools: an overview", In: Proceedings of the 2007 international conference on Computer Systems and Technologies, CompSysTech '07, pp. 40:1-40:6, ACM, New York, NY, USA, 2007.

AUTHORS PROFILE

Sumathi S received the B.E., degree in the year 2004 and M.E., degree in the year 2011 from Periyar University, Salem and Government College of Technology Coimbatore respectively. Currently, she is working as an Assistant Professor in the Department of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore. Her research interest includes Machine Learning Techniques and Text Mining.

M.P.Geetha received the B.E., degree in the year 2004 and M.E., degree in the year 2009 from Bharathiyar University, Coimbatore and Kumaraguru College of Technology Coimbatore respectively. Currently, she is working as an Assistant Professor in the Department of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore. Her research interest includes Big Data Analytics and Data Mining.

Dr. Pugalendhi GaneshKumar received the BTech, MS (by research) and PhD degrees in Information Technology in the year 2003,2008 and 2012 respectively from University of Madras, Anna University , Chennai and Anna University Coimbatore. Currently, he is an Assistant Professor in the Department of Information Technology, Anna University, Regional Center, Coimbatore. His research interest includes application of soft computing techniques in data mining and bio informatics.

Dr.K.Pushpalatha has received her B.E. degree in Information Technology from Bharathiar University, Coimbatore, India and her M.E degree in Computer Science and Engineering from Anna University. She has also received Ph.D Degree in Faculty of Information and Communication Engineering from Anna University, Chennai, India. Her research interests include Wireless communication, Cloud computing and Data mining. She is a life member of ISTE. She is currently working as an Assistant Professor of Information Technology department at Coimbatore Institute of Engineering and Technology , Coimbatore India.

A.S.Shanthi has received her B.E.,Degree in Computer Science and Engineering form Bharathiar University,Coimbatore,India,during 2000.She has also received M.E.,Degree in Software Engineering and Ph.D Degree in Faculty of Information and Communication Engineering from Anna Uniersity,Chennai,India.Her research interests include Data mining,Image processing,Networks,Wireless Communications.She is currently working as an Associate Professor of Computer Science and Engineering Department at Tamilnadu College of Engineering,Coimbatore, India.