Visualization of Plagiarism Detected in Documents

T.Mala, T.V.Geetha
Department of Computer Science and Engineering
Anna University, Chennai
mala@cs.annauniv.edu,rctamil@annauniv.edu

Abstract

Tremendous amount of data is available digitally now a days and therefore need arises for a system which can automatically track any copying of data without permission. Copying of others copyright material is an illegal act. Plagiarism as it is called, is an act of copying others literary works and posing as though it is ones own work. Here in this paper we have developed a system which can automatically detect plagiarism and also uses the data to visualize the plagiarized document. A detector cell is created which will collect all the detected plagiarized information. Detector cell is created based on three criteria namely Copying, Reordering and Replacing. Exact copy of the sentences are detected by checking sentence replication of certain sentences where keywords occur frequently. Reordering is possible by structurally modifying the sentences by changing the order of words. Replacement is done by replacing certain words with their synonyms in the copied sentences. As soon as the plagiarized sentences are detected the detector cell collects the information and updates itself for future references. Later the entire document is modeled as a plain three dimensional terrain surface over which the detector cell information is applied and therefore the plagiarized sentences appear as elevated surfaces and all other sentences appear as plain flat surface. The exact copy of sentences will have more elevation when compared to structurally reordered sentences and replacement sentences which are modeled as having less elevation. This modeling helps users in analyzing the document with respect to the amount of information copied, Spatial location of copied sentences and the behaviour of plagiarism. This also helps in knowing the sentences which are plagiarized the most from multiple sources. The size of the document is also mapped to the size of the terrain surface so that spatial location information is maintained. Gestalt perception laws of Similarity, Proximity and Continuity are also applied over the plagiarized data to obtain a gestalt based visual model.

1. Introduction

Plagiarism is copying others work and claiming it as ones own work. This can be detected using various techniques. Detection of plagiarism in students programs can be done using simple string comparison algorithm [1]. VAST a tool to numerically score the students plagiarized work helps in visualizing the plagiarized information but multiple source plagiarism cannot be visualized using this tool [3]. A metric was formulated based on Kolmogorov complexity to measure the amount of shared information between two computer programs. Software Integrity diagonostic system was developed which uses the metric to detect plagiarism [9]. Many tools like Turnitin and Word check were developed to detect plagiarism and are available as web enabled plagiarism detection tools. These tools perform word checking and phrase matching with a set of already existing papers [2]. Keyword centered core detectors were used to detect plagiarism in natural language text. These detectors were incremental and dynamic in nature [4]. A document sketch is prepared for each test document and a hash

0-7695-3050-8/07 $25.00 © 2007 IEEE
DOI 10.1109/ICCIMA.2007.109

value is generated. The amount of plagiarism depends on the similarity of this hash value with a set of training documents. A system thus developed helps in identifying the similarity candidates in the case of news articles [8]. Plagiarist behavior at various level of document structure was captured . The plagiarist behavior considered are Insertion, Deletion and Change [5]. Levenshtein distance was used to find the similarity between two documents at various levels. Some of the issues in all the above cases were visualizing the plagiarized information, detecting multiple source plagiarism and analysis of plagiarized information. All the above issues were tackled in the new system proposed here. The system takes into account multiple sources, visualizes the plagiarized information and helps users in analyzing the plagiarized document applying Gestalt perception laws. Section 2 of the paper deals with the methodology used in plagiarism detection, Section 3 discusses the implementation details of detector cell creation and Section 4 talks about modeling the plagiarized information, Section 5 applies Gestalt perception laws over the model generated for document analysis and Section 6 gives conclusion and future extensions of the system.

2. Detector Cell Creation

Detector cell creation was a dynamic incremental process. The set of training documents are pre processed first. Pre processing involves removal of stop words, other irrelevant words like web site address, E_mail id and irrelevant symbol. Later the preprocessed training documents are given to sentence splitter for splitting the sentences. A detector cell is created for each training document. This cell consists of statistical information of all words in the pre processed document like frequency of occurrence and its location information. Later Brill tagger is used for parts of speech tagging of the training document set. Then the words from the training set are taken and synonym information of each word is extracted from wordnet dictionaray and was stored. All the above processes are done automatically as soon as the training set of documents were given into the system. The detector cell also collects the synonym information and tagged sentence information. Then the test document is given. The test document is also pre processed like training documents and later checked for the following conditions.

2.1 Copying

The words from the pre processed test document are collected and its frequency noted. This is checked with the detector cell of each training document. If the word matches and match is above a threshold level the sentences where the word occurs in both training and test document are extracted. These sentences are mutually compared for exact copy of sentences. If an exact copy occurs it is marked. Then, the document was checked for structural reordering.

2.2 Reordering

The pre processed test document is also tagged as training document set. In the reordering check of a test document, sentences whose keywords match with the training detector cell more than a threshold level are extracted. If they fail for exact match they are then checked structurally. The structure of these sentences are identified from the tagged file and checked for consistency. Same sentences may be repeated with a change in word order. This structural reordering can be identified and is captured.

2.3 Replacing

Finally the test document is checked for synonym replacement. The words in the test documents are checked not only with the words in the training set but also with the word synonyms present in detector cell. If any number of matches goes beyond the threshold level the corresponding sentences are checked. If both matches with only change in synonym then the sentence is marked. In all the above cases the detector cell is incremented with the plagiarized sentence as and when a match occurs. This incremental detector cell helps in knowing the sentences which are plagiarized most often. This helps in quick detection where these sentences can be compared first as soon as a new test document is given into the system.

3. Implementation Details of Detector Cell

The user interface after authentication gives option for inserting authenticated documents to database so that it can be trained with other suspected documents. Simultaneously, a detector cell table is created with the table design with a unique name, created from unique id assigned for each entry of the table as shown in figure 1. Later the following actions are performed over the test document – Gate initialization, stop word removal, sentence splitting, POS tagging, temporary word_table creation, communication with wordnet, synonym retrieval, exact line matching, comparison of simply modified sentence in order and in words by changing with synonyms and generation of visualized form of result of text intrusion detection. The detection module creates a temporary table for storing all relevant words of input file, with its position indicated by the sentence number. The words are sorted and not allowed to repeat in table. Multiple occurrence of a word entered to table by appending the new sentence number. The table entry is written into a text file. This text file with plagiarized sentence information is used for terrain modeling.

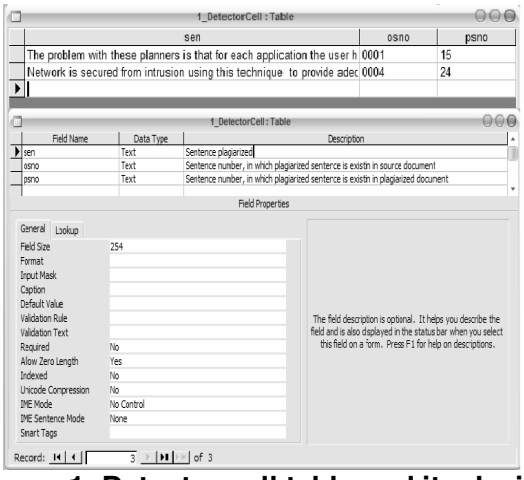

Figure 1. Detector cell table and its design

4. Three dimensional terrain modeling

The plagiarized information from the detector cells are collected and are mapped onto three dimensional terrain surface model . The three dimensional terrain surfaces are modeled using polygonal meshes. Here triangular polyhedra are modeled and are

then patched to form a three dimensional terrain surface. The surface area of the terrain is modeled based on the size of the test document. The triangular strips are generated over the entire area and are then patched together. These strips will have same height information. The location where sentences are copied exactly will be mapped onto triangular strip with greater height and the locations where there is reordering of sentences or replacing will have height lesser than exactly copied sentences but more than normal sentences. The terrain so created will have elevated surfaces wherever there is plagiarized information. The terrain surfaces so modeled for two journal articles are shown in figure 2 and in figure 3. Figure 2 represents excessive plagiarized document with mostly exact copy but less of reordering or replacement. Moreover the plagiarized information lies evenly throughout the journal article and the reordered and replaced sentences occur in the starting position of the document. On the other hand figure 3 shows a document with less plagiarized information and the behavior of plagiarism tend to be more of replacement and reordering rather than exact copy.

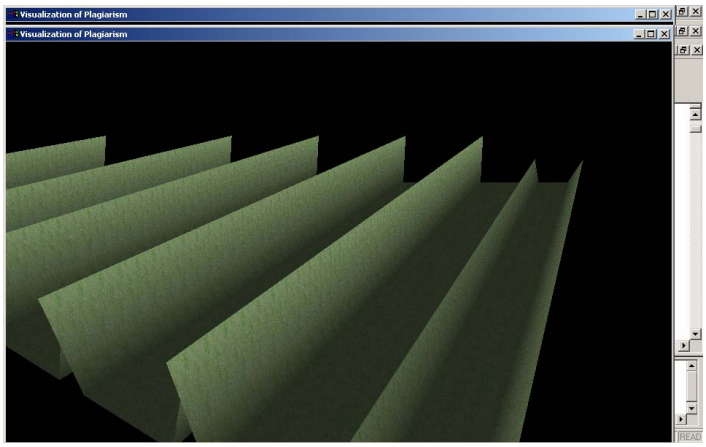

Figure 2. Document with excessive plagiarism

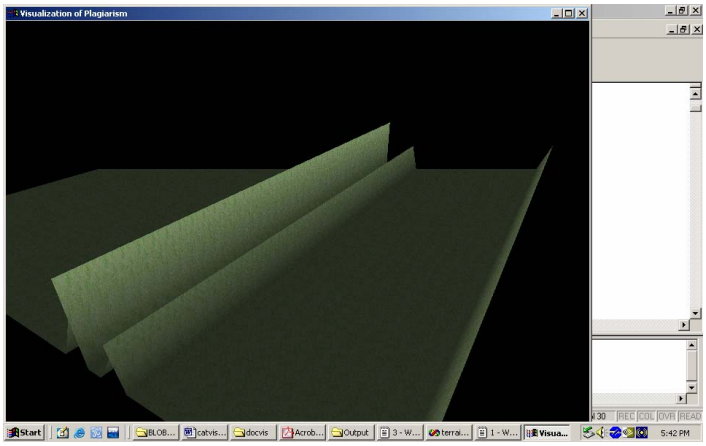

Figure 3. Document with less plagiarism

There is only a single case of exact copy in the article and it occurs in the middle of the article. The reordered or replaced sentences occur in the starting position of the article and in the middle too. Later Gestalt perception model is applied over the model for better perception and understanding of the generated visual.

5. Gestalt based visual perception Model

Gestalt law states that whole is greater than sum of its parts [6]. Gestalt perception laws like Similarity, Proximity, Continuity, Closure and Symmetry were applied over the category information and the visuals obtained were analyzed for better understanding of the documents [7]. Here the whole considered is the entire document and the perception laws are applied over this visual. Similarity law states that similar looking objects form a group and applying similarity here similar height terrain locations form group either plagiarized or normal group. By proximity nearer objects form one single group and here nearer elevated locations visualizes nearer sentences to be plagiarized preserving the spatial location information. Continuity law states that we perceive continuous object even though discontinuity exist and here the elevated discontinuities in the terrain surface makes the whole surface to be perceived as a one single document. This gestalt based perception model is useful for analyzing the plagiarized document and understanding the plagiarized information better.

6. Conclusion & Future work

The system developed helps automatically find out plagiarism present in documents. The method used considers exact copy, reorder and replacement strategies into account from multiple sources. The gestalt based visual terrain model developed helps to know the kind of plagiarism, the amount of plagiarized information contained in the document and the location where exactly the plagiarism has occurred. Further the system can be improved by applying other plagiarist behavior like modification of sentences and insertion of new words. The visual can be made aesthetically appealing by applying different texture over the figure for differentiating the locations.

7. References

[1] Allan Parker, J. 0. Hamblen, Computer algorithms for Plagiarism detection, I*EEE Transactions on Education,* Vol 32, No 2, May 1989.

[2] Colin J. Neill and Ganesh Shanmuganthan, A Web-Enabled Plagiarism Detection Tool, IT Pro, September | October 2004

[3] Culwin F, Lancaster T, Visualising Intra-Corpal Plagiarism, Fith International conference on Information visulization, 2001.

[4] Heon Kim, Yang-koo Kang, Pyung-Jin Kwon, Moon-Hyun Kim, An Application of DICOM Architecture for Detecting Plagiarism in Natural Language, The 9th Intemational Conference on Computer Supported Cooperative Work in Design Proceedings.

[5] Manuel Zini, Marco Fabbri, Massimo Moneglia, Alessandro Panunzi, Plagiarism Detection Through Multilevel Text Comparison, Proceedings of the Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS'06)

[6] Max Wertheimer, Gestalt theory, Gestalt journal press,1924.

[7]T.Mala, T.V.Geetha, Multilevel categorization Visualization based on Gestalt Perception model, 15 th Scientific Convention on Relations and Structures Developments of Gestalt theory in psychology and adjacent fields, marcerata, Italy May 25-27, 2007.

[8] Wolfgang Kienreich, Michael Granitzer, Vedran Sabol, Werner Klieber, Plagiarism Detection in Large Sets of Press Agency News Articles, Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06)

[9] Xin Chen, Brent Francia, Ming Li, Brian McKinnon, Amit Seker, Shared Information and Program Plagiarism Detection, IEEE transactions on Information theory, vol. 50, No. 7, July 2004