International Journal of Engineering & Technology, 7 (2.32) (2018) 14-17

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Context Similarity Strategy for Text Data Plagiarism Detection

[1]Durga Bhavani Dasari, [2]Dr. Venu Gopala Rao. K

[1]Assistant Professor, Dept of CSE, Koneru lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India.
[2]Professor, Dept of CSE, G. Narayanamma Institute of Technology and Science, Hyderabad, India.
**Corresponding author E-mail: bhavani.dd@kluniversity.in.*

Abstract

Advent development of anti-plagiarism solutions has supported varied range of elementary forms of textual recycling, however, considering the magnum of content that is being generated, a tool alone might be ineffective in preventing complex forms of plagiarism. Some of the issues that are envisaged with the plagiarized articles in many of the open-access journals emphasize the point that critical deficiencies of varied kind of solutions that are existing aren't being resourceful in identifying the manipulation that is taking place in the form of paraphrasing and editing. Manipulative editing has become a major menace even in the case of predatory journals and is leading to issues of publication ethics. Certain preventive strategies that have evolved in the recent past are relying on semantic solutions, comprehensive texts evaluation, graphics, reference lists, key words, digital technologies. It is right time for enforcing adherence to global editorial guidance and towards implementing a comprehensive set of strategies to address the issue of plagiarism.

Keywords: *Plagiarism detection, Citation based plagiarism detection, context relevance, Semantic similarity.*

1. Introduction

Plagiarism refers to the copying of someone else's concepts, proprietary efforts and pushing them forward as their self generated notions [1]. This issue is often observed in graduates and undergraduates in addition to research scholars [2].

In the recent past, one copy related issue targeting Germany federal defense ministry circulated in the newspapers. Unintentionally, a law instructor discovered copied paragraphs in the minister's phd papers [3]. However, the minister rejected the allegations as obscure. Few enthusiasts developed the GuttenPlag study [4] to investigate the actual volume of copied work published in his research.

Further, by April 2011, coordinated attempts of these enthusiast researchers successfully depicted that his work consisted copied contents to the extent of 371 in 393 key pages. Overall, around 64% of the sentences have been discovered as copied. The below pictograph depicts the outcome of their investigation.

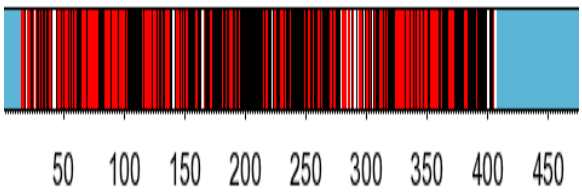

Figure 1: Copied text papers in Germany defense minister's work [4]

In the pictograph, the dark shaded bars imply papers affected by plagiarism out of a single work while the red bars denote the copied text from various research works. The white bars depict genuine work, not copied from any author. Rest of the pictograph (the blue bars) depict the cover page, contents in the work and references.

After presenting this study, the minister failed to convince that his work was error free. Accordingly, he returned his research degree and had to move away from his finance ministry.

This research work implemented the reference driven discovery model to the minister's work so as to depict its capability of discovering intelligently hidden copy work, in specific, in scenarios of otherwise tough to identify language translated copied work. Having presented the brief of diverse types of plagiarisms and existing discovery models, this research emphasizes on Cb PD model. Later on, the research methodology for the assessment is produced. In the last chapter, pros and cons of integrating both textual and reference models are described.

2. Related Research

Different Types of Plagiarism

Analysis of different plagiarism characteristics convey that several generally observed approaches for unauthentic document copy that can be described as below. Sentence-to-sentence copying type depicts the mode of plagiarism where specific segments or the complete document belonging to another user. Hidden copying methods involve intentionally masked sections. Paraphrasing concept is the willful representation of others notions, in the context of plagiarism, by ignoring the original author reference by hiding the citation [5]. Language translation-based copying is another human or automatic translation of works presented in some language to other language with aim to disguise the author. Concept plagiarism involves the utilization of larger alien idea through unauthentic means. An illustration of the same is the reflection of other study's notions, concepts, frameworks, simulation phases and other source information [6].

Plagiarism Discovery Tools

Plagiarism discovery (PD) denotes processes assisting the detection of copying conditions. Currently available PDS is often cate-

gorized as extrinsic and internal. Extrinsic PDS assesses the originality of test file over an authentic dataset. The internal PDS evaluates different language attributes of the test file through styometry approach and does not assess it in terms of similarity with extrinsic sources. As can be understood, the former approach attempts to discover literally similar word sections, the later approach emphasizes more on modifications in presenting style [7].

Diverse assessment approaches were put forward for extrinsic PDS. Of this, some of the prominent studies are presented below. Sentence similarity processes attempt to discover long sentences with similar phrases. These phrases are perused as markers for prospect plagiarism until the complete document crosses the preset threshold level. Mostly implemented suffix text approaches like suffix trees were implemented for the detection task [8].

Fingerprinting models are the largely incorporated PD model, attempting at establishing the summary of the entire research by choosing a pair of subsections from the research study. The pair implies the fingerprint and its components are referred to as minutiae. Different models of computational, hash-like models are deployed on these sub-components to convert them into effective byte-strings [9].

Above thousand style-markers are defined in the concept of stylometry [10]. These markers include different lexical attributes like how long a phrase is, to syntax attributes like parts of speech repetitions, to structural attributes like recurrence of punctuations. On the other hand, the internal PDS often consist of independent mixture of different language attributes [11].

CbPD model is another model contrary to the aforementioned models based on textual similarity. This is capable of assessing research works in science field as this needs citation. In the earlier research [12], we first put forward deploying reference study for PDS and assessed the model efficiency through a manually developed corpus.

3. Pros and Cons of PDS

Subjective and objective evaluations of discovery efficiency of PDS are challenging because the incorporated compilations and assessment approaches vary drastically. Two research works attempted to handle such deficiency in comparison. These approaches aim to compare PDS against state-of-art compilations and controlled assessment scenarios. PAN-PC was initialized in 2009, where participants provide basic study models [13]. A cyclic assessment of PDS is evaluated through a research study at the HTW college from 2004 [14].

The PAN-PC assessment dataset primarily consists of manually copied segments generated as well as partly disguised methods. Further, four thousand sections, which have been artificially disguised by individuals were deployed to experiment the characteristics of a plagiarist are incorporated [15]. The HTW university studies, a dataset consisting of 42 text files were intentionally plagiarized and actual reports ranging from one to one and half papers were utilized. The authentic pages are taken from websites over the internet [14], [16].

A few outcomes of these events are provided to present and overview of the inherent pros and cons of the available PDS scheme. The chart below depicts the plagdet numbers for best 5 performers incorporating extrinsic PDS and two performers of internal PDS, who were among the competitors of the event [13]. These assessment outcome values are ordered with respect to the disguising approaches incorporated to copied document sections. The total plagdet numbers for total classes is presented beside each label. The label in the chart "-I" is assigned to differentiate the model of MUHR ET AL., competing in the category of internal PDS while others competed in the extrinsic PDS.

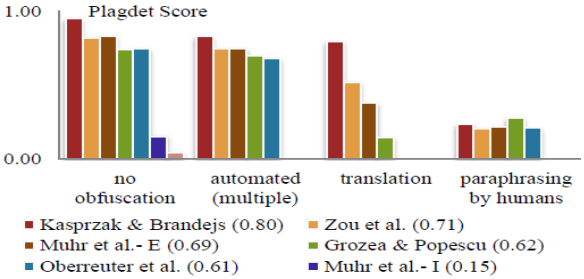

Figure 2: Outcomes of best five performers in PDS approaches PAN-PC'10 [13]

The outcomes depict that c&p type of copying is easy to discover with better precision rates by benchmark PDS. Nevertheless, discovery ratio for obfuscated sections, in specific, for manually disguised sections, is significantly smaller for all models. The event conductors assess the outcomes obtained in discovering cross-language plagiarism to be challenging. The most efficient models utilized machine support for converting languages of the original documents in the standard dataset. These services are observed to be same as the ones deployed for building plagiarized segments. However, the final results showed that manual intervened language conversions were highly disguising practical plagiarism due to large complicatedness and variations involved, thereby, resulting in poor detection rates of the PDS [13].

The results put forward by the HTW University researchers are in synchronization with the results published in the PAN-PC competition. Interestingly, both these studies were unable to detect language conversion-based plagiarism [14]. This strengthens the belief that non-practical identification rates in language conversion context in the competitive event were recorded due to the simulation setup provided in the event.

Moreover, the efficiency of an extrinsic PDS varies largely with respect to the feed datasets given to the model. Accordingly, models and approaches relying on wide range of search indexes of google, bing, etc tend to record the most efficient outcomes [17]. Similarly, the human oriented queries of doubtful words and phrases also achieve efficient outcomes in the same context.

4. PDS on the basis of References

A quick review of the contemporary literature revealed that no major studies have focused on addressing the plagiarism issue through analyzing references provided in the bibliography in any research work. Only our earlier study [12] witnessed some progress in this context. In the paper, we put forward the below description-

CbPD considers approaches, which utilize references for establishing the common sections between the test document and reference documents to establish the extent of plagiarism involved in the test document.

In schools and colleges, the bibliography section comprising of author or source references of research works is perceived as important information for both further understanding and establishing value of the work with respect to related research studies [18].

Detecting the similarity in the references mentioned in both test document and standard document remains a robust marker for detecting semantic commonness between the documents. These reference prototypes are regarded as subsequences in the reference tuples and of both the documents which contain partly shared citations and accordingly, are considered as being alike.

The extent of this alikeness between the reference sections is largely dependent on other features like the chapter's coupling extent. Further, the degree to which the pattern of associated references exists and their proximity are similar. The underlying concept is to compute the possibility of reference orders to be match-

ing by chance. To gain further understanding on the functioning of the associated programs, the study in [12] can be referred.

5. Similarity Check by Context Relevance

In this section, the focus is on unique set of attempts for indicating the redefinition of text data using comprehensive presentation or the paraphrasing as profoundly the plagiarism is based on the semantic and context relevance. Usually the contexts are defined as circumstances that lead to statement or concept which shall be impacting the paraphrasing pertaining to signifying the plagiarized content for the text data.

Critical considerations of the model can be defined as paraphrasing adapted for exploring the background, theorems that are least significant in terms of denoting the scope of plagiarism. But the paraphrasing that is discussed in description of an idea or event or statement of contribution reflects upon most significant aspects pertaining to plagiarism scope affirmation.

Context based similarity assessment function

- The semantic similarity of the test paper dv from the source paper C is computed through the following formulae-
- Overlapping of all words and hyponyms of test paper and source paper as $i_{sd}(dv,c)$
- Calculate the proportion of the count of such hyponyms and words in $i(dv,c)$ for total number of these in source paper, represented mathematically as $\frac{|i_{sd}(dv,c)|}{|c|}$
- Calculate the semantic similarity $dst_{sd}(dv,c)$ as $\frac{|i_{sd}(dv,c)|}{|c|}$
- The idea similarity of the test paper dv from the source paper C is computed through the following formulae-
- Overlap the concept and functionality sets of test paper and source paper as $i_{cd}(dv,c)$
- Calculate the proportion of the count of such sets in $i_{cd}(dv,c)$ for total count of these sets in source paper, represented mathematically as $\frac{|i_{cd}(dv,c)|}{|c|}$
- Calculate the concept similarity $dst_{cd}(dv,c)$ as $\frac{|i_{cd}(dv,c)|}{|c|}$
- The context similarity of the test paper dv from the source paper C is computed through the following formulae-
- Overlapping of the keywords $\{kw(dv)\}$ of any passage or chapter of the test document dv and keywords $\{kw(c)\}$ of corresponding section in source paper C as $i_{kw}(dv,c)$
- Calculate the proportion of the count of keywords in $i_{kw}(dv,c)$ for total count of keywords in $\{kw(c)\}$ as $\frac{|i_{kw}(dv,c)|}{|\{kw(c)\}|}$
- Calculate the similarity $dst_{kw}(dv,c)$ as $\frac{|i_{kw}(dv,c)|}{|\{kw(c)\}|}$
- Overlap the citation titles of suspect document dv and reference titles of source document C as $i_{vt}(dv,c)$
- Calculate the proportion of count of titles in $i_{vt}(dv,c)$ for the volume of titles in $\{vt(c)\}$ as $\frac{|i_{vt}(dv,c)|}{|\{vt(c)\}|}$
- Calculate the similarity $dst_{vt}(dv,c)$ as $\frac{|i_{vt}(dv,c)|}{|\{vt(c)\}|}$
- Overlap the source publishers list $\{al(dv)\}$ obtained from references of test file dv and publishers list $\{al(c)\}$ obtained from the source paper C as $i_{al}(dv,c)$
- Calculate the proportion of count of publishers in $i_{al}(dv,c)$ for count of publishers in $\{al(c)\}$ as $\frac{|i_{al}(dv,c)|}{|\{al(c)\}|}$
- Calculate the similarity $dst_{al}(dv,c)$ as $\frac{|i_{al}(dv,c)|}{|\{al(c)\}|}$
- Later, the context similarity can be calculated $dst_{cod}(dv,c)=\left(dst_{kw}(dv,c)+dst_{al}(dv,c)+dst_{vt}(dv,c)\right)^{-1}$

- The total similarity of test file dv from the source file C is calculated as below-

- $dst(dv,c)=(dst_{cd}(dv,c)+dst_{cod}(dv,c)+dst_{vt}(dv,c))^{-1}$

//The total similarity is determined on the basis of the proportion of concept, context and semantic similarities respectively.

6. Empirical Study

Categorically, the fundamental outcomes highlight the intersection between the test files and source files, which is using percentage of the common words, percentage of certain common words that are extracted using consecutive common sequences. It is considerable that both the models are hard-baselines.

In the Table I, results of the corpus for the experiments are denoted. The outcome from the experimental study reflects that the proposed method has attained a higher level of accuracy and F_1 measure considered to the other models, thus resulting in better set of performance than the best baseline configuration (i.e., 1−gram) with an overall growth of 5.24% with respect to precision rates.

The fundamental outcomes are significantly larger as can be observed in the table below. The table successfully depicts that reflecting the relation of phrase overlapping as the prime reason for identifying plagiarism. In addition, the suggested approach could successfully accomplish superior classification function, depicting that there exist specific functions wherein the individual phrases often intersect approaches that the other models were unable to target.

Table 1: Comparison of The Suggested Approach with The Standard Approaches Over the Same Dataset

	Acc in %	**f-measure**
proposed	0.78	0.683
n-grams	0.63±0.62	0.6144±0.044
seq lengths	0.67±0.042	0.64±0.032

Table II denotes the performance using the proposed method vs. divergent rewriting actions. Considering the scope for observation, as the verbatim sequences permits correctly classified near copy and the non- plagiarized cases, it is imperative that heavy class revision is usually considered as non- plagiarized.

Also, the transposition actions do not reflect any progressive signs unless the automation identifies more precisely the heavy revision cases rather than the verbatim automation. In the proposed case scenario, the automation detects any kind of insertion or deletion

of certain actions which reflect more accurate elements across the plagiarism classes.

Lastly, the automated process incorporated to identify the addition/removal functions was observed to possess highest precision among all plagiarism groups.

Table 2: Performance Assessment of Divergent Plagiarism Practices Detection.

No plagiarism	0.79 ± 0.11
Consistent rewrite	0.273 ± 0.27
Inconsistent rewrite	0.34 ± 0.12
Minor phrase change	0.67 ± 0.07

7. Conclusion

Model that is discussed in this manuscript reflects upon the paraphrased or comprehensive presentation of some text which is inevitable, as the definition of such renowned concept like the newton's law, the other theorems that do not change. But if it has to be briefed as a need for exploring the actual contributions, it will be posed as a comprehensive of the actual version. In such instances, the content presented for a comprehensive state or the paraphrased state shall not be treated as plagiarism.

Even if the same content is posed as a new contribution of any format like the comprehensive or the paraphrased, they are to be considered as plagiarism. Such an act which notifies comprehensive presentation or the paraphrased version for a content as plagiarism in specific conditions and in certain conditions it might not be considered as context-oriented plagiarism.

Profoundly the domain of journal publication classification for the overall presentation of the content is for distinct set of sections that are majorly considered as context of the text. Based on the context, content similarity is weighed based on estimation for overall plagiarism comprised for the chosen input document sets against the corpus.

References

[1] McArthur, Thomas Burns, and Roshan McArthur, eds. Concise Oxford companion to the English language. Oxford University Press, USA, 2005.

[2] Sun, Zhaohui, et al. "Systematic characterizations of text similarity in full text biomedical publications." PLoS One 5.9 (2010): e12704.

[3] zu Guttenberg, Karl-Theodor. Verfassung und Verfassungsvertrag: Konstitutionelle Entwicklungsstufen in den USA und der EU. Duncker & Humblot, 2009.

[4] Guttenplag wiki. Online Resource, Retrieved Apr. 10, 2011 from http://de.guttenplag.wikia.com, 2011.

[5] Clough, Paul. "Plagiarism in natural and programming languages: an overview of current tools and technologies." (2000).

[6] Fröhlich, Gerhard. "Plagiate und unethische Autorenschaften." (2006): 81-89.

[7] Stein, Benno, Moshe Koppel, and Efstathios Stamatatos. "Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection PAN'07."

[8] Monostori, Krisztián, Arkdy Zaslavsky, and Heinz Schmidt. "Document overlap detection system for distributed digital libraries." Proceedings of the fifth ACM conference on Digital libraries. ACM, 2000.

[9] Hoad, Timothy C., and Justin Zobel. "Methods for identifying versioned and plagiarized documents." J. Am. Soc. Inf. Sci. 54.3 (2003): 203-215.

[10] Rudman, Joseph. "The state of authorship attribution studies: Some problems and solutions." Computers and the Humanities 31.4 (1997): 351-365.

[11] Stein, Benno, Nedim Lipka, and Peter Prettenhofer. "Intrinsic plagiarism analysis." Language Resources and Evaluation 45.1 (2011): 63-82.

[12] Gipp, Bela, and Jöran Beel. "Citation Based Plagiarism Detection: A New Approach to Identify Plagiarized Work Language Independently." HT'10. 2010.

[13] Potthast, Martin, et al. "Overview of the 2nd international competition on plagiarism detection." In Proceedings of the SEPLN'10 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. 2010.

[14] http://plagiat.htw-berlin.de/software/.

[15] 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.

[16] Weber-Wulff, Debora. "Test cases for plagiarism detection software." Proceedings of the 4th International Plagiarism Conference. 2010.

[17] Maurer, Hermann, Frank Kappe, and Bilal Zaka. "Plagiarism-A Survey." Journal of Universal Computer Science 12.8 (2006): 1050-1084.

[18] Garfield, Eugene. "Citation indexes for science. A new dimension in documentation through association of ideas." International journal of epidemiology 35.5 (2006): 1123-1127.