Detecting Unacknowledged Plagiarism Using String Matching Based Content Framework

[1]Lawal O.O, [2]Ajibode T.S, and [3]Adepoju A.O
[1, 3]Department of Computer Science, D.S Adegbenro I.C.T Polytechnic, Itori , Ogun State
[2]Department of Computer Science, Federal College of Education (Tech) Akoka Lagos State.
E-mail: opepolawal@gmail.com, pojubiodun@gmail.com, ajibodex@gmail.com
Phone: +2348062160129, +2348137223966, +2349056755171

ABSTRACT

Plagiarism is an ugly trend that posed global concern and challenge on information resources and documents spanning the internet repository and academia. Most often, an individual or some authors in team work usually have parallel construction in written text, term paper and research article which connotes similar phrases, vocabulary overlap and collaboration between texts across multiple pages of the same document. In this paper, we present a functional mechanism for higher performance in plagiarism detection especially in collaborative manner. An iterative development technique was adopted involving 'String Matching Based Content' so as to handle text parsing and document analysis from file to file. The system functionality was modeled from users' perspective by providing algorithm for internal control and to check text similarity.

Keywords: Algorithm, Detection system, Plagiarism, String matching

iSTEAMS Proceedings Reference Format
Lawal O.O, Ajibode T.S. & Adepoju A.O (2019): Detecting Unacknowledged Plagiarism Using String Matching Based Content Framework. Proceedings of the 17th iSTEAMS Multidisciplinary Research Nexus Conference, D.S. Adegbenro ICT Polytechnic, Itori-Ewekoro, Ogun State, Nigeria, 21st – 23rd July, 2019. Pp 65-72. www.isteams.net - DOI Affix - https://doi.org/ 10.22624/AIMS/iSTEAMS-2019/V17N1P7

1. INTRODUCTION

Plagiarism is an attempt and/or act of claiming credit for written ideas, published text and intellectual works of others in academia. It is a trait of academic dishonesty involving undergraduate and post graduate students, as well as intermediate researchers in different fields of endeavor. Although, this has enhanced teaching and learning but the havoc done to academia is high, one of which is plagiarism. Digital computer and the advent of internet have made published text plagiarism-prone in every facet of life (Faloore, 2014). However, plagiarizing by use of substitutions to elude detention software has rapidly evolved as students and unethical academics seek to stay ahead of detection software (Grove & Jack, 2014).

Predicated upon an expected level of learning being achieved, all associated academic accreditation becomes seriously undermined if plagiarism become the norm (Cully, 2013). On the other hand, Plagiarism is the stealing of someone's idea and claiming to be one's original work which may be written ideas (Atayero et al., 2011). Plagiarism is the illegal presentation of someone's ideas, words, expressions, data, computer programming, or any other creative endeavor and claiming it to be one's original work without giving the author credit. Since the emergence of digital computer and internet, documents can no longer be handled in paper and ink form as they were decades ago. This has led to "copying and pasting" of textual document since these menu cannot be disabled in application programs so as to curb plagiarism.

Obinna (2012) noted that plagiarism and poor writing skills are the bane of Nigeria's educational system. Plagiarism affects not only the integrity of the individual concerned but also the integrity of the institution associated with the individual. Student's level of knowledge and ability in their respective places of work after graduation is as a result of their disciplinary engagement in achieving integrity of their academic experience. Ignorance and unawareness of legal implication of plagiarism are also responsible for verbatim copying of scientific texts (Jamali et al., 2014). Recently, computer-aided methods of plagiarism detection are in use but not all already existing tools implement completely new comparison algorithms (Atayero et al., 2011).

Nowadays, it is generally believed that attribute counting is inferior to content comparison, since even small modification can greatly affect fingerprints (Atayero et al., 2011). Meanwhile, a clever plagiarist that knows the best way of paraphrasing can easily use word substitution to outsmart detection system. Parsing and document analysis are quite germane to the performance of any plagiarism detection system. The functional mechanism may be desktop-based or web-based application for documents comparison or text similarity. It appears that developing countries suffer more from severe and obvious cases of plagiarism, since scientific foundation of developed countries were long established and fight against plagiarism started earlier (Jamali et al., 2014).

Inconsistencies within the written text itself such as changes in vocabulary, style or quality of an article can also be attributed to plagiarism except that such cases are known or described as unacknowledged and collaborative kind of plagiarism. Atayero et al. (2011) described unacknowledged and collaborative plagiarism as the occurrence of similar phrases, quotations, sentences or parallel construction in two or more pages of the same article. Consequently, situations in which spelling mistakes or lexical errors occur between text; overlapping one's phrases and content within a document, and as well as preparing a correctly cited and referenced write-ups from individual research or handling part of that work twice for separate purposes or publications like journal articles, therefore, a framework to provide detection mechanism using string matching based content in two or more documents for text similarity with unlimited text size to check for unnecessary tautology is of necessity.

2. LITERATURE REVIEW

Plagiarism has been defined as an act of reproducing another individual's ideas or any other form of information without giving credit to the source of the information. Plagiarism is a deliberate appropriation of others' ideas and statements without proper referencing (Smith et al., 2007). However, there are currently many measures designed towards combating it, such as the implementation of plagiarism detecting systems, as well as establishing penalties for the act. The advent of the Internet, World Wide Web and Social Media has undoubtedly taken our civilization a step higher, and especially, provided scholars with a useful avenue for the rapid spread and exchange of ideas and information. Centre for the study of higher education said "Plagiarism in higher education can take many forms. Some of the more common forms are listed below, however it should be noted that definitions of plagiarism vary somewhat across the disciplines in accordance with differences in knowledge authorship conventions and traditions.

Accordingly, the 'who is who' among American high school students'- a survey conducted on American best students revealed that eighty percent (80%) of them plagiarized to get to the top of their class, as corroborated by another study in which fifty-five percent (55%) of the respondents admitted to have plagiarized (Wan et al., 2011). Dupree and Sattler (2010) pointed out that seventy-four percent (74%) of students in their report admitted to have plagiarized in previous year. In addition students and faculty admitted that the plagiarizing via the Internet is fairly common. One of the causative factors of plagiarism could be unawareness. Wan et al. (2011) on engineering students studying in Malaysia revealed that most of the students know very little about plagiarism and what makes it a serious offence.

Furthermore, the research revealed that the students had not been given a very formal and concise orientation about it by the school. Although, at the moment, several plagiarism detection tools have been developed which effectively detect plagiarism. Ali et al. (2011) reviewed these tools, pointing out their merits and demerits. Their findings show that none of the tools are hundred percent (100%) efficient, but considering the different features of the existing detection tools, they argue that a hybrid system could raise the capability of detection. In Nigeria, there are disturbing reports about a plague of plagiarism at different levels of the academic sector (Onuoha & Ikonne, 2013). Studies from the western region of Nigeria on university students revealed a varying academic dishonesty among respondents, ranging from those that plagiarize from the Internet, to those that copy from their colleagues (Babalola, 2012).

This fault in the educational background of some Nigerians contributed to their failure to perform optimally abroad where academic rules are taken seriously (Orim et al., 2013). Confusion between plagiarism and paraphrasing among students is another influencing factor of plagiarism. Quite a significant number of students are unaware of the rules guiding paraphrasing.In fact, this is common when students are confronted with paraphrasing paragraphs from unfamiliar subjects or technical jargons. Students fall prey of unintentional plagiarism due to their inability to decipher the thin-line between paraphrasing and plagiarism. Of utmost importance also is the place of poor writing skills of students among the various factors and reasons students plagiarize. It is imperative that faculty members in related courses in a department should help students develop strong writing skills.

3. METHODOLOGY

In view of the problems inherent in the existing system, it is necessary to seek for an improvement because a hash function computes digital fingerprints for each chunk which are inserted into a hash table: a collision of hash codes within the hash table indicate matching chunks. Meanwhile, a clever plagiarist can easily maneuver by smooth paraphrasing; a detection mechanism with string matching algorithm as functional framework is bound to perform better. Most of the works in plagiarism detection are meant for academic purpose where cheating, rewording, rephrasing, or restating without referencing are strictly prohibited. In this regard, numerous plagiarism detection systems have been developed.

But, the functional mechanism of proposed framework is stand-alone with the following operational modules and algorithm:

1. Load files to be checked
2. Ignore white spaces and special characters
3. Compare the files line by line
4. Detect similar phrases
5. Extract and store in memory the similar phrases
6. Check online server for similar phrases
7. Displays the links having similar phrases.

```
Int match         //value to return
        int I, j, k;
        match = -1;
        j=1; k=1; i=j;
while (endText(T,j)==false)
        if (k>m)
        match = I         //match found
        break;
        if ($t_j$ == $p_k$)
        j++;  k ++;
else              //back up over matched characters
        int backup = k-1;
        j=j-backup;
        k=k-backup;       //slide pattern forward, start over
        j++; i=j;
return match;
```

System design

Iterative development technique was adopted in providing functional mechanism of the proposed framework which was code named “plagdetect”; blueprints were presented through the use of visual aid like procedure, data flow, activity and use case diagram. Meanwhile, top-down approach was chosen for the design in order to build the system from main module and later incorporate all other sub-programs has shown in figure 1 and figure 2 below.

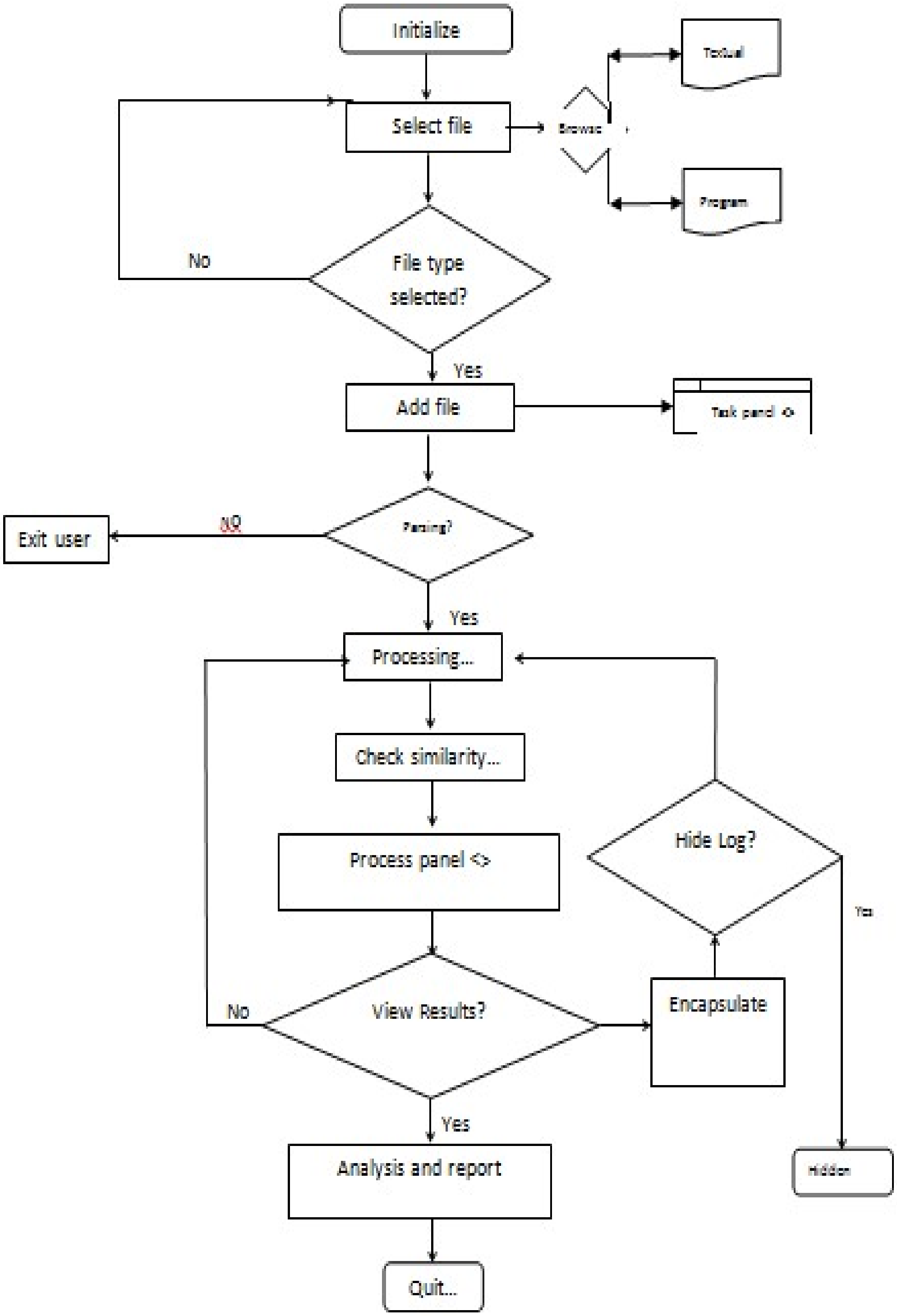

Fig. 1 procedure and data flow diagram

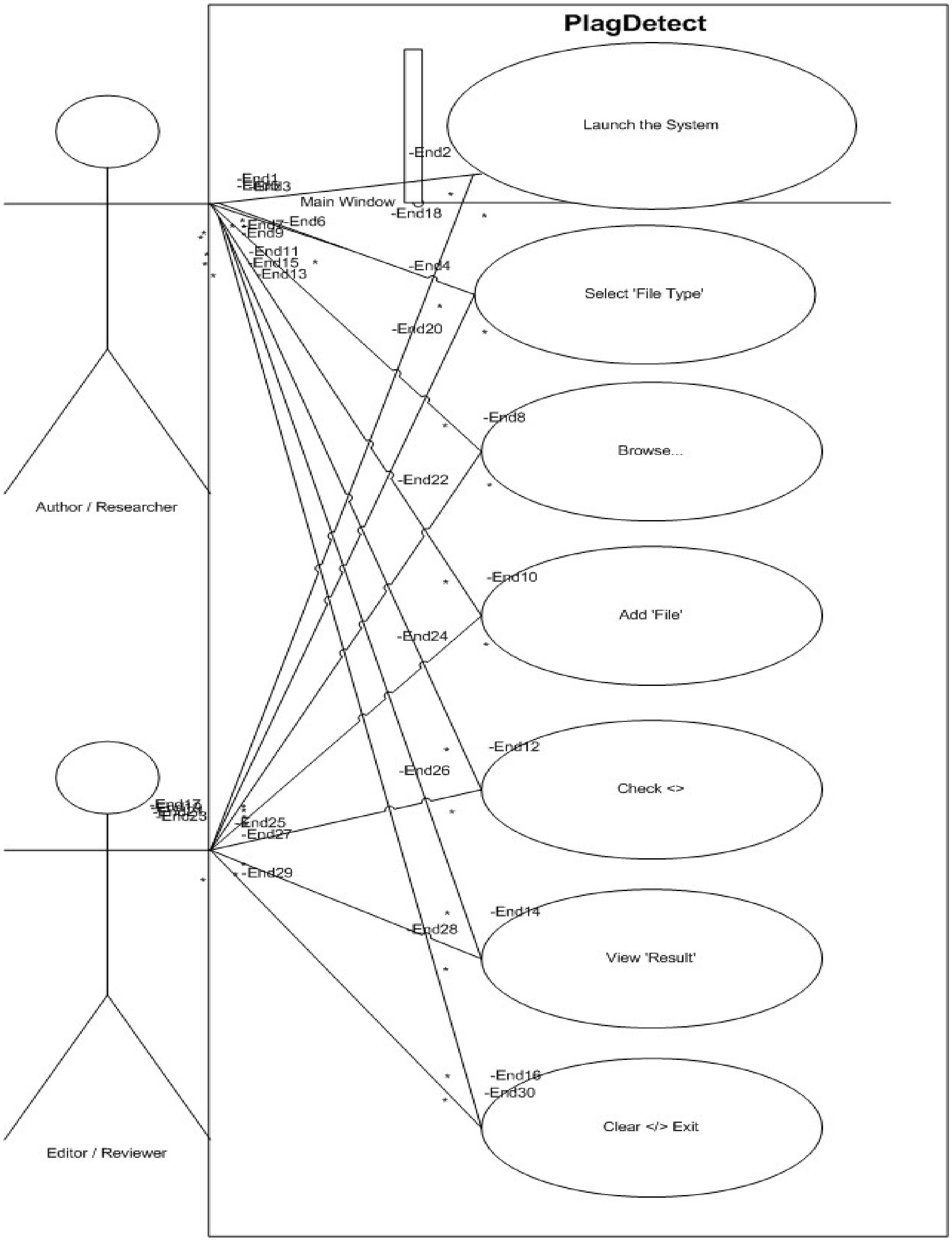

Fig. 2 use case diagram

4. CONCLUSION

Natural language processing is of great importance to the functionality and modern construction of plagiarism detection system as detecting plagiarized phrases in the given document becomes easier after the semantic analysis of the document. Algorithm has been provided with formal representation as detection mechanism using string matching based content for better performance. Thus, handles file to file similarity check within textual search and program codes, like copy detection in large comparison from electronic resources. Enterprise level plagiarism check software would normally compare input documents against databases.

5. FUTURE WORK

Computational intelligence is of great importance to modern construction of plagiarism detection tool and its functionality. Application of natural language processing (NLP) to similarity test and document analysis will be considered as part of the areas for further study.

Its features can leverage on pattern matching to handle ambiguity and unconstrained vocabulary so as to resolve lexical overlap in large files. Multi-lingual capability can also handle language issue in detecting exact copied percentage in part or wholly, translating documents in different languages irrespective of writing style of the author.

REFERENCES

1. Atayero, A.A., Alatishe, A.A., Sanusi, K.O. 2011. Developing iCU: A Plagiarism Detection Software. *International Journal of Electrical & Computer Sciences.* 11(4), 27-32.
2. Ali, A.E.T, Abdulla, H.D., Snasel, V. 2011. "Survey of plagiarism detection methods". Proceedings of 5th Asia Modeling Symposium, Manila, 39-42.
3. Bar, D., Zesch, T., Gurevych, I. 2015. Composing Measures for Computing Text Similarity. Technical report of the language lab., Germany. 1-30.
4. Babalola, Y. T. 2012. Awareness and incidence of plagiarism among undergraduates in a Nigerian private university. Gale-Bult. 22(1), 53.
5. Cully, P. 2013. Plagiarism avoidance in academic submissions. Dublin Institute of Technology, Full PDF available at: http://arrow.dit.ie/bescharcoth/4/*. Retrieved on 15/03/ 2016.*
6. Dupree, D., Sattler, S. 2010. Texas Tech University Academic Integrity Survey Report. Retrieved from <http://www.depts.ttu.edu/provost /q ep/docs/Dupree and Sattler _Academic_Integrity_Report _Cover.pdf
7. Faloore, O.O. 2014. Towards a More Enduring Prevention of Scholarly Plagiarism among University students in Nigeria. *Global Journal of Human-Social Science: Sociology and Culture.* 14(6), 1-7.
8. *Grove, M., Jack, W. 2014.* "Roget would blush at the crafty cheek Middlesex lecturer gets to the bottom of meaningless phrases found while marking essays". URL <http://www.msn.com>.
9. Jamali, R., Ghazinoory, S., Sadeghi, M. 2014. "Plagiarism and Ethics of Knowledge": Evidence from International Scientific Papers. *Journal of Information Ethics*. 23(1), 101-110.
10. Orim, S.I., Davies, J.W., Borg, E. 2013. Exploring Nigerian postgraduate students' experience of plagiarism: A case study. *International Journal for Educational Integrity*. 9(1), 20–34.
11. Obinna, C. 2012. "Plagiarism, bane of Nigeria's educational development". Retrieved from <http://www.vanguardngr.com/2012/09/plagiarism-bane-of-nigerias-educational-devt-provost/> on April 24th, 2016.
12. Onuoha, U.D., Ikonne, C.N. 2013. Dealing with the Plague of Plagiarism in Nigeria. *Journal of Education and Practice.* 4(11), 1-11.
13. Smith, M., Ghazali, N., Minhad, S.F. 2007. Attitudes towards plagiarism among undergraduate accounting students. Malaysian, *Review of ACCA.*
14. Wan, R., Nordin, S. E., Halib, M.B., Ghazali, Z.B. 2011. Plagiarism among Undergraduate Students in an Engineering University: An Exploratory Analysis. *European Journal of Social Sciences.* 25(4), 537-549.