 J Soc Sci, 49(3): 362-370 (2016)

Study Document Validation and Mapping with User-profile for Collaboration

Ruth Oluwatosin Adeyemo[1], Olusesan Adeyemi Adelabu[2] and Akeem Adewale Oyelana[3]

[1]Department of Computer Science, University of Ibadan, Oyo State, Nigeria
[2]Faculty of Science and Agriculture, University of Fort Hare, Alice, P.B. X1314, Eastern Cape, South Africa
[3]Department of Public Administration, University of Fort Hare, Alice, P.B. X1314, Eastern Cape, 5700, South Africa
E-mail: [1]<tosadeyemo@gmail.com>, [2]<201409080@ufh.ac.za>, [3]<201100592@ufh.ac.za>

KEYWORDS Plagiarism Detection. Text Comparison. Study Document. User-profile

ABSTRACT As study and publication remains a yardstick for scientific endeavours, it is not enough for researcher to only publish their papers, it is therefore paramount that the quality of research publications be put in check through validation of research document, ensuring research document submitted in a repertoire does not already exist and providing a possible forum for collaboration amongst researchers. The objective of the study was to use plagiarism detection in study document by comparing a researcher's work with previous publications based on user-profile. Current studies in the field of automatic plagiarism detection for content archives concentrate on algorithms that compare plagiarized documents with potential unique records inside a huge collection of documents. The methodology compared suspicious documents against a set of potential original documents which have been filtered out from the large repertoire of documents based on the user profile. The researchers used two main algorithms which are the study document validation algorithm and text comparison (PlagCheck) algorithms coupled with user-profile to detect plagiarized document hence determine the validity of a study document. The framework was assessed by utilizing a test-set that contained occurrences of verbatim duplications and messages with little or no alteration. The result and performance evaluation showed the researchers' system performed better and faster than existing systems, achieving the accuracy of ninety-eight percent (98%) over splat. The study was able to take care of the challenge of processing time of validation which is usually encountered in other Plagiarism Detection Systems (PDS).

INTRODUCTION

Having noticed the growth in the amount of study papers and publications published yearly by researchers and the availability of free, unrestricted and immediate online access to published scholarly materials, primarily peer-reviewed study articles in academic journals which is beneficial to researchers and the academic world, there is need to consider issues this exposure and rapid growth has brought into the academic world.

Nigeria is yet to have efficient information service system unlike some developed countries like USA and some Latin American countries that are always in firm control of information. The criteria to measure 'world Class College' does not reside in the number of students but in the quality of research. Advanced nations are always evaluated higher because of the quality of information and control over human and capital resources coupled with enhanced living conditions (Sabo 2005). In an increasingly competitive and information driven economy, governments look up to colleges for assistance.

As study and publication remain a yardstick for promotion in academia in Nigeria, it is not enough for researchers to only publish their papers, it is essential that for the social and economic growth and development of Nigeria, many research document and findings from Nigerian tertiary institutions must have an impact on industrial, commercial and administrative processes on all fronts hence, it is therefore paramount that the quality of study publications be put in check. Report from the empirical appraisal made on study documents from Nigeria universities

Address for correspondence:
Oyelana Akeem Adewale
Department of Public Administration,
University of Fort Hare, Alice, P.B. X1314,
Eastern Cape, 5700, South Africa
Cell: +27837286640,
E-mail: 201100592@ufh.ac.za

by Chiemeke et al. (2009) reveals the decadence in the study output especially in the polytechnics.

Cetto (1998) posits that one criterion for measuring study report is the number and nature of distributed works by Nigerians global diaries, and by the quality of reports from the colleges, which add to the generation, dissemination, and use of logical learning for advancement in Nigeria and beyond. The study therefore intends to reveals that there are more than economic and academic benefits to be derived from publishing of papers; one of the crucial issues that cannot be ignored is to determining how to assess the validity and quality of published findings.

Statement of the Problem

A gradual decline in study document in higher education became noticeable in the late 1980s. The National University Commission (NUC) noted that in terms of quality and quantity, the study document of tertiary institutions in Nigeria was about the best in sub-Saharan Africa up to the late 1980s (Karani 1997). The basic foundations for the study are: good study training and motivation, availability of equipment, and good library facilities. At the onset and acceleration of the decay in the system, these ingredients faded away. By 1996, the quantity and quality of study had declined to an all-time low (Shahabuddin 2009; Singh and Remenyi 2015). These and many more challenges in the academia constituted to the problem of plagiarism in the academic community. The system proffers solution on how to improve study document in the academia is based on the problem statement stated as follows:

- Determine if study document submitted in a repertoire already exists or not based on user profile
- Determine the validity of a study document
- Provide forum for collaboration

Aim of Study

This study is aimed at using plagiarism detection in study document by comparing a researcher's work with previous publications based on user-profile.

Objectives of the Study

- To help learning institutions determine if a publication has been used for promotion before.
- Fair assessment of researchers
- Faster and more efficient means of assessment of researchers

Research Question

- How can academic institutions determine whether a publication has earlier been published elsewhere?

Plagiarism Overview

The advancement of Information Communication Technology has made plagiarism easy for lazy researchers. Before, individual had to struggle in libraries to locate information from books manually. With ICT careless copying from books became very easy. Based on the previous studies conducted, it has been noticed that the act of plagiarism is rampant amongst students worldwide but in recent times it has also become notable among established researchers. In a self-report study performed, among 82,000 students about 40 percent of undergraduates and 25 percent of graduates engaged in plagiarism within 12 months prior to the study (McCabe 2005). Results of other studies range as high as 90 percent of the subjects self-reporting acts of plagiarism (Lim and See 2001).

Plagiarism Detection for Text Documents

In the educated community, instances of plagiarism are detected with relative easy. A computerized literary theft check of ~285,000 logical writings of arXiv.org yielded more than 500 records is liable to plagiarism. Likewise, 30.000 reports (~20% of the gathering) were observed to have contained inordinate self-copyright infringement.

Detecting plagiarism in a mass of researchers is difficult and also time consuming. Detecting plagiarism can be done either manually or by using computer-assisted means. Manual detection requires substantial effort with time, and is impractical in situations whereby there are too many documents to be compared, or in cases whereby the original documents are not available for comparison. Also the manual method of detecting plagiarism requires substantial effort and the use of excellent memory which is impracticable in most especially in situations in which large volume of documents are being involved. For computer-assisted detection, detect-

ing plagiarism is much more realistic and it allows vast collections of documents to be compared to each other. Computer-assisted detection does not totally eliminate the problem with time since it still takes time to compare one document against several others.

Plagiarism Detection Systems

Plagiarism Detection Systems (PDS) can be divided into different types which are hermetic, web, general purpose, natural language, and source code (Mozgovoy et al. 2010). Hermetic PDS performs by scrutinizes only local collection of documents. Plagiarism search by this system is usually within a maintain database of document. Web detection system is aimed at identifying instances of plagiarism that have been collected from internet sources.

Although thorough search for instances of plagiarism is been carried out, we realized that a lot of time is been wasted in carrying out search without any area of specialization on both relating and non-relating documents to the suspicious document. Some existing web system is anti-plagiarism (Kakkonen and Myller 2009) and splat (Collberg et al. 2003). These frameworks are equipped for both web and hermetic identification. The non-specific location framework depend on string matching algorithms which are fit for preparing archives of any nature (whether a content made in a characteristic dialect or project source code). Being widespread, such frameworks experience the ill effects of the absence of specialization, permitting the miscreants to utilize a more extensive scope of allowing the cheaters to use a wider range of effective plagiarism-hiding tricks

This work uses the hermetic detection method, concentrating on document comparison coupled with the user profile which makes searching for plagiarism within a specialized area (user profile space) and hence, the problems related to organization and maintenance of large text databases are taken care of.

AntiPlag System

Because of the fact that manual detection of plagiarism is so tedious and frequently off base, a few programmed identification apparatuses have been created to handle the issue of counterfeiting. Kakkonen and Myller (2009) present AntiPlag; the principle highlight of framework was centered on testing based web unoriginality recognition which is equipped for both hermetic and web location. Their work was done for distinguishing cases of written falsification that have been sourced from the web. The framework utilizes standard web search engines to find records on the web that may have been utilized as sources of plagiarism by the writer of a text. The apparatus was equipped for distinguishing verbatim replicating from the web and from nearby reports.

The suspected sources were downloaded, changed over to American Standard Code for Information Interchange (ASCII) content and spared to nearby the neighbourhood database with the goal that they can be later proposed by utilizing the hermetic detection methods.

The assessment of the AntiPlag was completed on test information that comprises cases of verbatim duplicating and text in which plagiarism was concealed by minor editing, supplanting words with equivalent words and by summarizing. On the off chance that the rate of appropriated content in an archive is equivalent or higher than 25 percent, AntiPlag considers it as an extreme instance of plagiarism and alerts the client. The framework supposedly was contrasted with four other web location framework performed better, accomplishing the precision 95.8 percent over the whole test terms (Kakkonen and Mozgovoy 2010). Part of the methodology proposed in the AntiPlag system was improved upon in the system proposed in chapter three of this work.

SPLAT System

Another example of plagiarism detection system is the splat system which uses a web spider to crawl through the web collecting research papers. The programming language webL (Kistler and Marais 1998) was used in the development of the web spider. The web spider works by starting from the homepage of the author and then transverse all links downloading all the research paper for the author. A major problem detected is the issue of the web spider downloading irrelevant papers from the web. So, there are several constraints put in place to limit where the spider could go. Constraints like checking for links containing the web publication, papers or research, searching through author's link us-

ing the breadth first search algorithm, all files downloaded are converted to text, filtering program is run to remove all files that did not convert properly and those that are not research papers. These processes were also used in Cora to find research papers. In detecting the plagiarised documents, text comparison was performed on all documents in the database which took a lot of time. The final algorithm used consists of three main parts which are:

- Parsing the text documents into paragraph and sentences in a canonical from
- Performing a highly optimized, brute-force, pair wise comparison of the parsed documents;
- Producing an Html report of the results.

It was discovered that though this system produce effective results, it utilized a high processing time in comparing a suspicious document against several other documents which are either relevant or irrelevant. This system focused on addressing the challenge of processing time by reducing search space to relevant documents through the user-profile in order to categorise research documents and validation that is based on the user-profile.

Publication System Overview

By looking at submission and decision of publishers, problems and difficulties confronting them can be identified. One of such is the report on examination from districts.

User Profiling

Client profiling is the way of gathering data about a client with the end goal of profile development in the interest of the client. The client's profile, for the most part contains data which are specific to the client. For example, information like age, sex, dates of birth and areas of interest. Client profiles are used by an assortment of electronic administrations for various purposes. One of the essential use of client's profile is that it can be used for proposal. A few works have been done on client profiling in connection with interpersonal organizations and informal communities, for example, Twitter, Facebook.com utilize client profile to discover potential companions in light of the current connections and applicable gathering participations of the client. Additionally, client profiling is being used by expert interpersonal organizations, for example, Linkedln.com. Client profiling, in this work, the researchers propose client profiling as a strategy for making bunches of productions in a database.

User Profile Matching

Matching of client profile is typical in light of the contents. The data contained in a client profile can be given by the client either unequivocally or mined by the application benefit that deals with the profile (Hassan et al. 2013). The absolute most basic client profile substance includes: client interest; client information, client foundation and aptitudes; client objectives; client singular qualities; client conduct and client connection (Schiaffino and Amandi 2009).

RESEARCH METHODOLOGY

The methodology and software proposed in this paper is aimed on external plagiarism detection method with hermetic detection coupled with predefined similarity criteria (user-profile).

Detection Process

The detection of plagiarism in the document (publication) to be validated was performed based on a Three-Stage Plagiarism detection process with the incorporation of automated plagiarism detection algorithm. The stages are the Collection stage, Analysis stage and the Validation stage.

Collection Stage

This phase of the detection process involves the collection of corpus for detecting plagiarism; the corpus was collected from various researchers in diverse discipline of computer science and from downloaded materials from the internet.

Analysis Phase

The analysis phase performs the pair wise comparison of the documents. Firstly, the document to be validated is compared against the set of document retrieved from the reduced search space after which an in-depth search for plagiarism is carried out based on hermetic detection which perform naïve pairwise file-to-file compari-

son, this results in O (f(n)N2) complexity, where N is the number of files in the collection and f(n) is the time to make the comparison between one pair of files of length n. To reduce the search space the document is matched against the local database based on the criteria of the user's area of discipline to reduce the search space, hence generating a datamart of relevant documents. This allows researchers to quickly find instances of plagiarism in a relevant search space rather than validating the document against a large database consisting of both relevant and irrelevant documents.

Validation Phase

This phase of the detection process involves the validation of corpus for detecting plagiarism; this phase is the most important phase in our research document validation system. The phase includes the document comparison process and the threshold detection process.

Analysis of the Proposed System

The challenge of the previous plagiarism detection system was solved in this research by addressing the issue of time. This new system achieves this by using an approach that reduces the search space based on the user-profile. The area of specialization of the proposed user is used to streamline the database and produce a dataset (data mart) that is relevant to the new document which is to be validated. The new system allows users to search and validate their research document within and against a search space of documents which are highly relevant and matches the user-profile of the user rather than searching the whole database.

Loading and Pre-processing of Documents

To perform the comparison on an extensive corpus of archives, the suspected sources were downloaded, changed over to ASCII content and after that spared into the local database. The loading of the document is done manually by inserting the details of a research document with its file content into the local database. The database designed to store documents consists of two tables relating to the document: Category and Publications. Each document is identified with relevant categories that have been created in the database. Once the details of the publications have been entered and validated, the publications are stored into their respective relevant categories. The publications are parsed into various fields common to the research papers. All publications must be in the text file format before being loaded into the database. Any new research document (publication) which is to be validated if not already in the text format is converted to text document before validation.

User Profile

Another client profile is classified by a profile matching part and is subsequently incorporated into the client profile space, utilizing the accessible information put away as a part of the profile.

User Profile Extraction

Author profile's (User profile) are extracted from the details of the validated publication and through user registration. The profile data of the user is entered through the design interface. If the profile extracted matches with any other profile in the profile space, the publication of such author will be updated but if not, the user profile space will be updated with a new user profile.

User Profile Space

For this undertaking a client profile space is required, that is created from a high number of client profiles. Such profiles are gathered over time to yield enough connections and inclinations of the clients. It is additionally conceivable to import client profiles from existing profile databases or to make them from virtual clients or personas.

Finding Potential Matches

To find the set of potential matching document (source documents from which the new document could have possibly plagiarized from), we speculate that for a document to have copied from another source document, there must be a relation between the two documents, that is, they would belong to the same field of study. Based on this, to initiate the matching process, the new document consist details such as the

Publication title, Publication content, Author's name, Keywords, Category etc. The Category field which specifies which field document belongs to, is the basic criteria which is used to find the potential matches is the database.

Once the details of the document are submitted, a search query is issue to extract all documents of authors whose profile matches the category of the documents. Those source documents that match with the area of discipline of the user profile are filtered out from the database to form our potential matches.

The aim of the filtering with the user profile is to lower the number of coincidental matches between new document and the local sources. Also, this is to ensure a faster means of validation of documents within a user profile space relevant to the user profile.

Matching Based on User-profile

After loading the publication details, the publication is matched with existing database, based on an implicit profile of the user. The matching decision has to be made based on the profiles stored in the author profile in the database. Matching of document is by comparing the suspicious document against each document in the set of potential matches. Using user profile to filter out the set of potential matches reduces the search space. This methodology speeds the rate of document validation. Validation of document with this system compared with other PDS is faster and effective.

Validation of Document

Validation of all research documents (document) is done with the researchers' plagiarism detection system (PDS) called PlagCheck which functions by comparing the research document with documents from the resulting matching process performed above.

Document Comparison

Comparison is done based on the text comparison algorithm proposed in this system.

The text comparison algorithm involves, string matching and sentence matching. Sentences in the document are delimited by full stops and then checked for verbatim or copy paste instances. Sentences are looked at from two angles. Sentences that are indistinguishable procure the greatest score conceivable; sentences that are exceptionally like each other gain a score some place between 40-100 percent of the conceivable score.

Comparing sentences for equality is simple and profoundly upgraded. Before taking a gander at the words in a sentence, the "aggregates" of the two sentences–calculated while parsing–are analysed. In the event that they are not the same, it is known instantly that the sentences are not indistinguishable. While periodic cover of the estimations of these entireties occurs between sentences that are distinctive, it is sufficiently uncommon to wipe out every pointless correlation. Just if the wholes and word numbers of two sentences are indistinguishable are the real strings thought about.

Sentences are viewed as comparative if the convergence of their arrangements of exceptional words is the same size or just marginally smaller than the sets themselves. Since the arrangements of one of kind words are kept up in sorted request, two-fold looking can be utilized to productively figure the extent of the crossing point of the sets. This examination is performed just when noteworthy likenesses exist. Sentences that have a noteworthy disparity in the sizes of their one of a kind word sets are overlooked, just like any sentences that have little arrangements of extraordinary words.

Detection Threshold

The threshold is the maximum percentage of total match that is set to report the instance of plagiarism in compared documents. With the source program for this work, the plagiarism detection system (PDS) determines the rate of overlapped text in the document against any matching document in the local database. A particular given threshold (40%) is stipulated, and any research document with overlapping text above or with equivalent percentage is reported as invalid.

Algorithm and Search Technique

There are two major algorithms purposed for the system. These are: the Research document validation (system) algorithm and the PlagCheck (Text comparison) algorithm.

Research Document Validation Algorithm

- the user enters (Publication, area of Interest)

- user's profile (area of interest) is matched against the local database
- if match exist
- select all publications matching the user's profile
- match the content of 1 above against retrieved documents (datamart)
- plagiarism check (text comparison)
- if the amount of matches exceeds a particular threshold
- then plagiarism exist (document is invalid), the plagiarised portion will be displayed
- else no plagiarism
- add publication to the database
- add authors and their profiles to the database.

Text Comparison (PlagCheck) Algorithm

- extract string from doc1
- extract category (discipline)
- search database and retrieve documents matching category in 2above
- for each document extracted from 3 above,
- Split each string in each document by '.?/!' and store each sentence in a row of array, where documents
  doc1= document to validate,
  doc2= document from database
- compare each row in doc1 with every row in doc2
- increment count on detecting a match
- plagiarism threshold = count/doc2.size
- if plagiarism threshold is above 20percent in every matching document then, plagiarism exist and document is invalid
- else, no plagiarism
- publication valid

Text Comparison of Document

Due to the complexities and certain drawbacks of existing plagiarism detection system, we proposed a model to detect plagiarism by fetching word by word, sentence by sentence comparison of documents against the other on first place regardless of grammars. The proposed algorithm follows the following steps in detecting text similarities: Normalization of document, String and sentence matching based on user profile, calculation of percentage match.

Normalization of Document

In this phase, we split all the documents into sentences which are parsed into an array. Sentences are delimited by special character such as full-stop (., ;?/!)

String and Sentence Matching

After the normalization of the documents we move on to the "Sentence Matching" phase. Here each string or words from the new document are checked with string from the database. A potential match is counted as weight of one.

Calculate Percentage of Matching

In this phase we made a ratio between the two documents over the similarity and number of words. Based on that the program will give an opinion whether the document is valid or not.

Decision

The decision to determine if the document submitted is valid or not depends on the calculated percentage of match. A threshold of 40percent was set, if the percentage match is equivalent or greater than the specified threshold this implies the new document contains a high positive meaning most of the contents has been published in another work. Hence, if threshold exceeds 40percent then the document is invalid but if it not up to 40percent it implies that the document is valid and it is submitted into the database.

Program Coding Platform

In the development of the proposed system, we made use of the Java programming language; the following are some of the software used:

- NETBEANS: Platform for coding and interface design
- MYSQL CONNECTOR
- DATABASE: MYSQL DATABASE

The database which is the backend was built using MySQL which is incorporated in Wampserver.

System Interface

The system interface (PlagCheck) is used for validating a new publication or document. The user is required to enter the details of the

document which are the title of the research output (work), the name of the author(s), keywords and the discipline the work belongs to.

RESULTS

At last, the rundown of examined reports is shown to the client with the data on the aggregate rate of the content that is suspected to have being copied in every document. In the event that the rate of the appropriated content in a record is equivalent or higher than 40 percent, PlagCheck considers it as a serious instance of copyright infringement and alarms the client. The client can examine the archive utilizing a graphical perspective that highlights all the associated examples with counterfeiting in the report.

Performance and Evaluation

The performance of the proposed system in this work was carried out by running a set sample documents that consist instances of plagiarism on the proposed system. The set of test data consists of text that has been designed for evaluating the plagiarism detection system. The data set was also used on other plagiarism detection system so as to compare their result with ours.

Using Splat as a sample study, it shows the result of running the same set of document that was run on the proposed system PlagCheck and Splat.

CONCLUSION

The study has been able to introduce the PlagCheck as a plagiarism detection system that is aimed at hermetic detection in documents to validate and assess the quality of a study document. Hence, the plagiarism detection tool was used for assessment of researchers work and also in the process creates a venue of researchers to view details of other researchers working in their area of specialization hereby promoting collaboration amongst researchers.

Contribution to Knowledge

The exponential increase of publications published yearly by diverse authors requires effective and efficient method of validation to prevent and detect the submission of duplicate (already existing) document into the repertoire.

To this end, with the design approach proposed in this paper and the results obtained, the researchers have been able to successfully present a faster and more efficient means of study document validation.

ACKNOWLEDGEMENTS

First and foremost, the researchers would like to give thanks to Almighty God. It would not have been possible to complete the paper without His guidance and protection.

The researchers would like to thank the University of Fort Hare. This research would not be possible without her funding.

REFERENCES

Cetto A 1998. *Scientific Journal Publishing in the Developing World*. India: Lesotho Coasted, Chennas.

Chiemeke S, Longe OB, Longe FA, Shaib IO 2009. Research Outputs from Nigerian Tertiary Institutions: An Empirical Appraisal. Library Philosophy and Practice. From <http://www. webpages. uidaho. edu/~mbolin/chiemeke-longe-shaib.pdf> (Retrieved on 1 November 2016).

Collberg C, Kobourov S, Louie J, Slattery T 2003. SPLAT: A System for Self-Plagiarism Detection. From <http://splat.cs.arizona.edu/icwi_plag.pdf> (Retrieved on 2 November 2016).

Hassan R, Scholes R, Ash N 2013. *Ecosystems and Human Well-being: Findings of the Conditions and Trends Working Group of the Millenium Ecosystem Assessment.* USA: Island Press.

Kakkonen T, Mozgovoy M 2010. Hermetic and web plagiarism detection systems for student essays - An evaluation of the state-of-the-art. *Journal of Educational Computing Research*, 42(2): 135-139.

Kakkonen T, Myller N 2009. AntiPlag-A Sampling-Based Tool for Plagiarism Detection in Student Texts. *Proceedings of the 8th European Conference on e-Learning*, Bari, Italy, October 28-29.

Karani F 1997. Higher Education in Africa in the 21st Century. *Paper Presented at the Africa Regional Consultation Preparatory to the World Conference on Higher Education,* Dakar, Senegal, 1st -4th April, 1997.

Kistler T, Marais H 1998. WebL- a programming language for the Web. *Computer Networks and ISDN Systems*, 30(1): 259-270.

Lim VKG, See SKB 2001. Attitudes toward, and intentions to report, academic cheating among students in Singapore. *Ethics & Behaviour,* 11(3): 261–274.

McCabe DL 2005. Cheating among college and university students: A North American perspective. *International Journal for Academic Integrity,* 1(1): 1–11.

Mozgovoy M, Kakkonen T, Cosma G 2010**.** Automatic student plagiarism detection: Future perspectives. *J Educational Computing Research,* 43(4): 511-531.

Okebukola P, Solowu OM 2001. Survey of university education in Nigeria. *Journal of Curriculum Studies,* 223(2).

Sabo B 2005. Universities, Research and Development in Nigeria: Time for a Paradigmatic Shift. Paper Pre-

pared for 11th General Assembly of CODESRIA, on Rethinking African Development. Beyond Impasse: Towards Alternatives, Maputo, Mozambique. From <http://www.codesria.org/Links/conferences/general_assembly11/papers/bako.pdf.> (Retrieved on 31 October 2016).

Schiaffino S, Amandi A 2009. Intelligent user profiling. In: M Bramer (Ed.): *Artificial Intelligence*: *An International Perspective, Lecture Notes in Computer Science 5640*. Berlin: Springer, pp. 193–216.

Shahabuddin S 2009. Plagiarism in academia. *International Journal of Teaching and Learning in Higher Education*, 23(3): 353-359.

Singh S, Remenyi D 2015. Plagiarism and Ghost-writing: The Rise in Academic Misconduct. From <http://www.scielo.org.za/pdf/sajs/v112n5-6/12.pdf> (Retrieved on 30 October 2016).

Paper received for publication on October 2015
Paper accepted for publication on November 2016