|| Volume 6 || Issue 4 || April 2021 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

CONTENT ANALYSIS AND DUPLICATION AVOIDANCE SYSTEM USING MACHINE LEARNING: A REVIEW

PROF. SWATI B. BHONDE[1], RASIKA NILESH BAHETI[2], ALISHA MARUTI ABHALE[3], SWETAMBARI DINKAR ANDHALE[4], ARTI BALU AROTE[5]

AMRUTVAHINI COLLEGE OF ENGINEERING, SANGAMNER[1,2,3,4,5]

---------- ***----------

Abstract: - Colleges have also become an essential subject of title, content checks and repetition avoidance. All is handled manually in the current setup. It is too mild and puts a lot of time into inspection. In customary work, repetition usually occurs. This paper focuses on the reiteration and understudies of undertakings. The framework store proposed has structured the previous projects' repository well.

Keywords- *Title verification, content extraction, content analysis, Machine Learning.*

---------- ***----------

I INTRODUCTION

Machine learning is used to find valuable knowledge from the vast volume of information. Information mining techniques are used to implement and address different types of challenges of discovery. This system would review the title of the project using techniques for knowledge mining and text extraction. The key undertakings when dealing with text are probably deleting text. Watchwords are useful for readers as they can decide quicker how the content deserves to be read. Web programmers benefit from watchwords so their subjects can accumulate comparable substances. Computer calculators benefit from the watchwords because they lower the size of text into the key points of attention. Various text extraction and string matching methods and algorithms. In any event, for text attractions, we use AI measurement.

II.RELATED WORK

The new method of plagiarism detection was found to be too sluggish in this paper [1] and takes too long for monitoring. Matching algorithms rely not only on semántic but also on the lexical structure of the text. The paraphrased text is also difficult to spot. In order to increase the percentage of results found and time management, plagiarism control with acceptable algorithm is the key challenge. The central question in this research is whether new approaches can be implemented, such as Semantic Function Marking, to deal with plagiarism issues with text documents. A lot of records can be accessed and accessible on the internet. Due to this accessibility, copying and pasting from these tools allows users to quickly generate a new document. Often users may overwrite the word with their synonyms to rewrite the plagiarized portion. The explanation for the paper is to find the most plagiaristic material which can be copied efficiently from everywhere. It also lets users or individuals publish their journals in their applications as a plagiarism process.

This paper [2] seeks to explore potential alternatives to the implementation of a methodology within the Indian education scheme that safeguards students from the academic code of ethics or copyright problems, travelling abroad or working in science research. Various factors in the school system have been critically assessed. This paper summarizes a good technique and acceptable ethical methodology for teaching and preparation.

In this article [3], the approach of gathering information is the way to gather information related to a curiosity problem. It sets the related records, which include keywords and sample files, to the user's premise. Search engines that correspond to web search, design search, federated search, Mobile search, enterprise search and social search are perhaps the most familiar feature of the knowledge recovery method. The main focus of this research is on desktop search. Desktop Search is the search variant defined for organizations on which knowledge is based, along with email and websites for the analysis of content, on the files contained on a personal computer. Contextualized document descriptions Content Analysis is a group of manual or computer-based methods. In order to interpret the contents, the various text-pattern corresponding algorithms are used and the contents are found inside an input document. In other implementations, these algorithms are usually used, which include bio-informatics, plagiarism, text mining and the correspondence of documents. String matching is important to find online and offline text patterns. The matching string algorithm is used to match the input text correctly or accurately. This research is mostly aimed at the analysis of the output of current algorithms that fit strings. There are four algorithms used for this comparison: the two-way algorithm, the Colussi algorithm, the optimal failure

algorithm and the overall shifting algorithm. It is seen from this analysis that the matching string algorithm of Colussi gives the best outcome.

The identification of the plagiarized data is most important in this paper [4] for the computing structures of research organizations, businesses, and education institutions to undertake such a job. The consistency of document-to-document base is calculated by existing tools to spot plagiarized documents. We introduced a method for text mining, the use of keywords and semanticized sentence processing to analyze a message. The principal objective is to interpret sentences and keywords and to search for parts of the text that other writers write. This is called a semantical study focused on keyword statements, where paragraphs are distinctive in their style. This strategy works by using keywords and semanticizing sentences, so no language requirements are essential. In this region, we think this function is improving. Compared to current ones, it would yield tremendous results.

Text Mining is an emerging field of study in this paper [5] in which the required user information must be provided from a wide variety of information. In the search box of text details T, the user needs to find a text P. The knowledge has to align whether the quest only succeeds. Many algorithms for this search matched strings. This essay addresses three distinct pattern search algorithms, of which only one pattern occurrence is studied. The Python-based algorithms of Knuth Morris Pattt, Naive and Boyer Moore were compared to each text length and pattern length for their execution time. This article also gives you a brief sense of time complexity, features of other contributors. The paper ends with the correct algorithm for duration and length of text.

The present paper [6] will present an assessment of 5 string search algorithms: Boyer-Moore, Knuth-Morris-Pratt, Karp-Rabin and Horspool. It is clarified how they work, when they work and when they are ideally suited to a specific problem. Any time we use our machines, string search algorithms are used. They help us to find our scripts, look for search aggregator strings and correct our mistaken terms. They are an important class of string algorithms that aim to find a place in the greater string or text where one or more string (the so-called pattern) occurs.

The Aho-Corasick algorithm was explored in this paper [7] and is ideally suited for multiple pattern matching and can be used in many application fields. The algorithm's complexity is linear in length plus the search time and the sum of matches to the output. In large numbers of keywords, it is appealing, since all keywords can be matched in one pass at the same time. Aho-Corasick offers solutions to many real world challenges such as intrusion prevention, plagiarism, bioinformatics, forensic information technology, mining text and many more. Aho-Corasick is among the most efficient text mining algorithms.

In this paper [8], a powerful algorithm (called ACM) matching strings with compact memory and high worse performance was proposed. The suggested ACM decreases the memory need without taking in complex processes with a mystical heuristic number based on the Chinese Remainder Theorem. The latency is also significantly decreased for off-chip memory references. In hardware and applications, the proposed ACM is quickly implemented. ACM then makes fast and cost-effective IDSs.

This paper [9] is well-known and significant for identifying the nucleotide or amino acid sequence in a protein sequence database for trends in the pattern discovery process in today's world. While in computer science the model matching is widely used, its implementations cover a wide variety, including knowledge selection in editors. We suggest in this paper a new model matching algorithm with an increased performance compared to the well-known literature algorithms up to now. After the analysis of famous algorithms like Boyer Moore, Horspool and Raito, our proposed algorithm created. When we talk of the overall success of the proposed algorithm, the shift given by the bad search character in Horspool was improved and a fixed order of comparison was established. The algorithm suggested is contrasted with other algorithms well-known.

III.PROPOSED SYSTEM ARCHITECTURE

Many graduate students from India are unaware of academic writing, plagiarism, ethical issues and research methodologies. Due to lack of appropriate or experienced teaching instructors in most of the institutions the standards of education and ethics in the education system are under serious threat. Such education system allows many students to practice unethical ways of pursuing an academic degree. Earlier awareness of research methodology and ethical issues were considered necessary only for research scholars doing PhD and was recently extended for post-graduation students as well. Government of India announced plagiarism as an academic fraud and unethical due to which the student or research scholar may attract punishment.

In proposed system, to store all previous implemented project data with synopsis in the system. The student registers details first. The details are name, email, password, roll no, branch and year. To verify the title student has to give the project name, keywords and abstract as an input. Then based on content analysis and plagiarism system will process a result.

This saves time in the long term because there is no need to re-organize, re-format, or try to remember details about projects. It also increases research efficiency since both the data

collector and other researchers will be able to understand and use well-annotated data in the future.

Advantages:

1.Title of the project and abstract will be the main inputs to the system

2.Based on content analysis and plagiarism tool system will process a result

3.System will be reservoir of old projects as well.

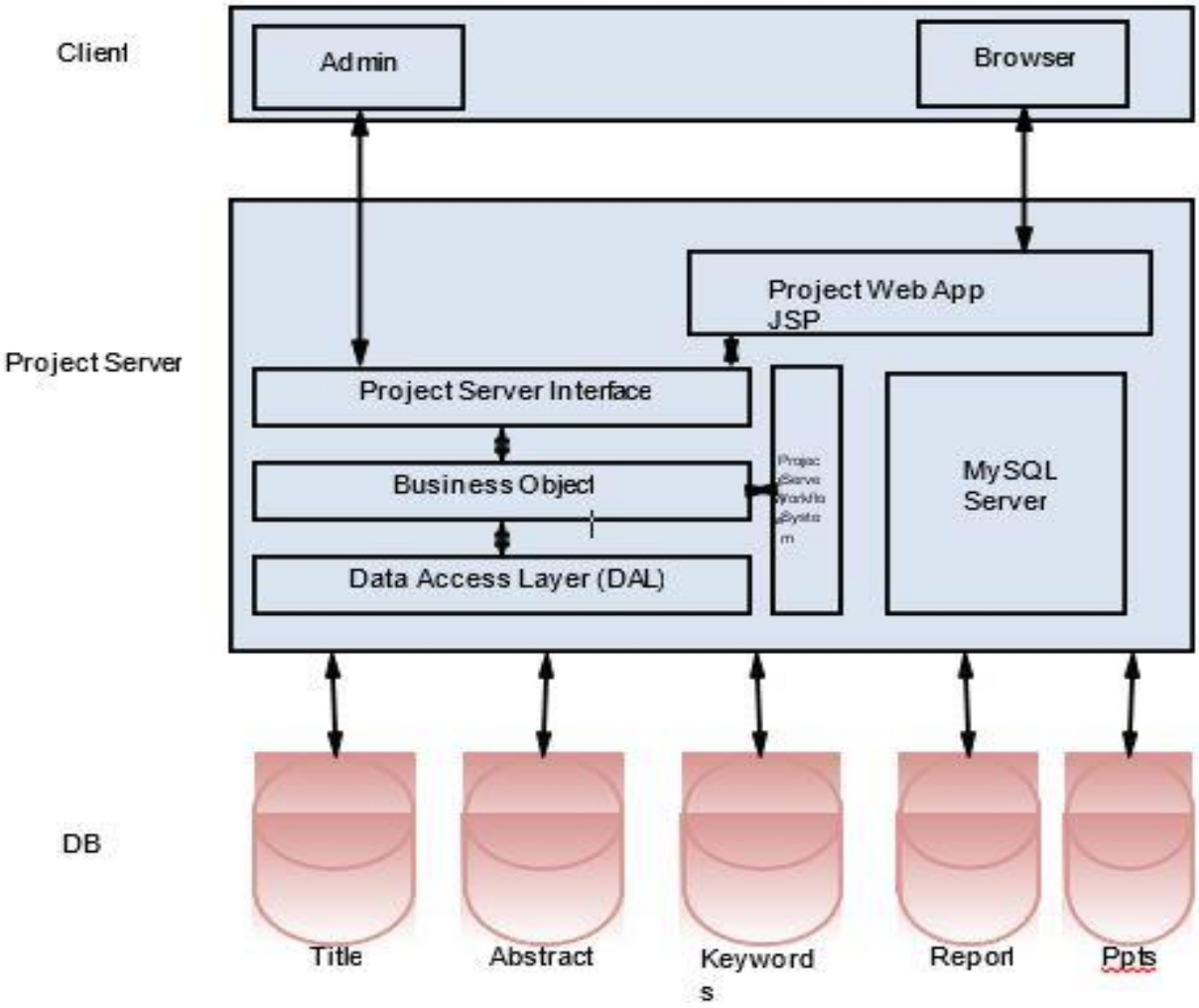

Fig 1. Proposed System Architecture

Algorithm:

Latent Dirichlet Allocation (LDA) Algorithm:

First and foremost, LDA provides a generative model that describes how the documents in a dataset were created. In this context, a dataset is a collection of D documents. Document is a collection of words. So our generative model describes how each document obtains its words. Initially, let's assume we know K topic distributions for our dataset, meaning K multinomial containing V elements each, where V is the number of terms in our corpus. Let βi represent the multinomial for the ith topic, where the size of βi is V: |βi|=V. Given these distributions, the LDA generative process is as follows:

Steps:

1. for each document:

(a) Randomly choose a distribution over topics (a multinomial of length K)

(b) For each word in the document:

(i) Probabilistically draw one of the K topics from the distribution over topics obtained in (a), say topic βj

(ii) Probabilistically draw one of the V words from βj

IV CONCLUSION

Projects would be held in place within the same storage space since the plan was set up ahead of time, allowing for it to reproduce.

REFERENCES

1.Ababneh Mohammad, OqeiliSaleh and Rawan A Abdeen, Occurrences Algorithm for String Searching Based onBrute-Force Algorithm, Journal of Computer Science,2(1): 82-85, 2006.

2.Saima Hasib, Mahak Motwani and Amit Saxena Importance of Aho-corasick string matching algorithm in pdf

3.JormaTarhio and EskoUkkonen, Approximate Boyer-Moore String Matching, SIAM Journal on Computing,Volume 22 Issue 2, 243 – 260, 1993.

4.Ravendra Singh et al, "A Fast StringMatching Algorithm", , Int. J. Comp. Tech. Appl., Vol 2 (6), 1877-1883. ISSN: 2229-6093, December 2011.

5.Hemlatha A.M, M.Subha, "A study on plagiarism checking with appropriate algorithm in data mining", International Journal of Research In Computer Application and Robotics, Nov 2014, ISSN 2320-7345

6.Kamalakar Pallela, Sneha Talari, "Plagiarism : A serious ethical issue for indian students", 2016 IEEE International Symposium on technology and society (ISTAS) 20-22 October 2016

7.Joshi O D, Pudale A H, Ghorpadde R D, "Text extraction technique applied to plagiarism detector : the semantic analysis for analyzing the writing style" International Journal of Computer science engineering(IJCSE) Mach-2016 ISSN 2319-7323

8.Dr. S.Vijayarani Ms. R.Janani: String Matching Algorithms for Retrieving Information from Desktop – Comparative Analysis.

9.Dr. (Ms). Ananthi Sheshasayee1 Ms. G. Thailambal2: A COMPARITIVE ANALYSIS OF SINGLE PATTERN MATCHING ALGORITHMS IN TEXT MINING