ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

Paper Duplication Avoidance System using Machine Learning Framework

Pranav Bhat

Department of Electronics and Telecommunication, Maharashtra Institute of Technology, Pune. Savitribai Phule Pune University, Pune, India

ABSTRACT: Detecting the duplicate data is of most concern to research organizations, industries and educational institutions on the computational systems for such a task has become important. Existing methods for detection of plagiarized document compute the similarity of document-to-document basis. Title and duplication avoidance has become an important issue in Colleges. In existing system everything done manually. It's too slow process and takes too much time for checking. Normally, Duplication happens in traditional work. This paper focuses on preventing repetition of project and students implement new concepts. Proposed system store well organized repository of the previous projects.

KEYWORDS: Data mining, Machine Learning, duplication avoidance, LDA

I. INTRODUCTION

Data mining is used for finding the useful information from the large amount of data. Data mining techniques are used to implement and solve different types of research problems.This system wants to verify project title using data mining and Text Extraction techniques. Extracting Text is one of the most important tasks when working with text. Readers benefit from keywords because they can judge more quickly whether the text is worth reading. Website creators benefit from keywords because they can group similar content by its topics. Algorithm programmers benefit from keywords because they reduce the dimensionality of text to the most important features.
Multiple Methods and Algorithms for Text extraction and String Matching. But we use LDA algorithm for Text Attractions.

II. LITERATURE REVIEW

The current plagiarism detection system was found to be too slow and takes too much time for checking. The matching algorithms are also dependent on the text's lexical structure rather than semantic structure. Therefore, it becomes difficult to detect the text paraphrased semantically. The big challenge is to provide plagiarism checking with appropriate algorithm in order to improve the percentage of finding result and time checking. The important question for the plagiarism detection problem in this study is whether it is possible to apply new techniques such as Semantic Role Labeling to handle plagiarism problems for text documents many documents are available on the internet and are easy to access. Due to this availability, users can easily create a new document by copying and pasting from these resources. Sometimes users can reword the plagiarized part by replacing the word with their synonyms. Motivation of the paper is to find the most plagiarism content that should be copied from anywhere identified in the efficient manner. Further it helps to as plagiarism detection process in applications to user or individual publish their journals [1].

A study has been attempted to evaluate possible approaches to introduce a methodology in Indian education system whereby students going abroad or working in scientific research are protected from the academic code of conduct or

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

copyright issues. A critical assessment was carried out on various imparting factors of education system. A decent procedure and adoptable ethical ways of teaching and training methodologies are being summarized in this paper [2].
Information retrieval is the method of retrieving the knowledge relevant to an issue of curiosity. It locates the relevant documents, on the premise of user's question which includes keywords or example files. Probably the most acquainted application of information retrieval system is search engine corresponding to Web search, Desktop search, Federated search, Mobile search, Enterprise search and Social search. This research work mainly focused on the desktop search. Desktop search is the specified variant of enterprise search, where the information foundations are the files stored on a personal computer, together with email and websites established on content analysis. Content Analysis is a group of manual or computer based approaches for contextualized explanations of documents. To analyze the content the different text pattern matching algorithms are used and it is used to discover all the existences of a limited set of patterns inside an input document. Commonly these algorithms are used in several applications that include information security bio-informatics, plagiarism detection, text mining and document matching. String matching is essential for finding text patterns that are in online and offline. String matching algorithm is used to matches the pattern precisely or about in the input document. The main objective of this research work is to analyze the performance of existing string matching algorithms. For this comparison there are four algorithms are used namely, two way algorithm, Colussi algorithm, optimal mismatch algorithm and Maximal shift algorithm. From this analysis it is observed that the Colussi string matching algorithm gives the better result [3].

The plagiarized data is of most concern to research organizations, industries and educational institutions on the computational systems for such a task has become important. Existing methods for detection of plagiarized document compute the similarity of document-to-document basis. We proposed system that will do text mining, exploring the use of keywords and semantic analysis of statements as a feature for analyzing a document. The main goal is to analyze the statements and keywords, looking for segments of the document that is written by other authors. This is considered as a semantic analysis using keyword based statement analysis where paragraphs with unique in style. This approach does work on the use of keywords and semantic analysis of statements, so it is do not require any language specifications. We feel that this feature shows improvement in this area. It will achieve tremendous results compared to existing models [4].

Text Mining is an emerging area of research where the necessary information of user needs to be provided from large amount of information. The user wants to find a text P in the search box from the group of text information T. A match needs to be found in the information then only the search is successful. Many String matching algorithms available for this search. This paper discusses three algorithms in unique pattern searching in which only one occurrence of the pattern is searched. Knuth Morris Pratt, Naive and Boyer Moore algorithms implemented in Python and compared their execution time for different Text length and Pattern length. This paper also gives you a brief idea about time Complexity, Characteristics given by other authors. The paper is concluded with the best algorithm for increase in text length and pattern length [5].
Anevaluation of five string searching algorithms will be presented; Brute Force, Boyer-Moore, Knuth-Morris-Pratt, Karp-Rabin and the Horspool algorithm. How they work, when they work and when they are best suited for a particular problem will be explained.
String searching algorithms are being used every time we use our computers. They help us find our files, search for strings on search aggregators and correct our misspelled words. They are an important class of string algorithms that try to find a place where one or several strings (called pattern) exists in a larger string or in a text [6].

The Aho-Corasick algorithmIs best suited for multiple pattern matching and it can be used in many application areas. The complexity of the algorithm is linear in the length of the patterns plus the time taken of the searched Text plus the amount of output matches. It is found to be attractive in large numbers of keywords, since all keywords can be simultaneously matched in one Pass. Aho-Corasick provides solution to many real World problems like Intrusion detection, Plagiarism Detection, bioinformatics, digital forensic, text Mining and many more. Aho-Corasick is one of the most productive algorithms in text mining [7].

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

An efficient string matching algorithm (named ACM) with compact memory as well as high worst-case performance. Using a magic number heuristic based on the Chinese Remainder Theorem, the proposed ACM significantly reduces. The memory requirement without bringing complex
Processes. Furthermore, the latency of off-chip memory references is drastically reduced. The proposed ACM can be easily implemented in hardware and software. As a result, ACM enables cost-effective and efficient IDSs [8].

The pattern matching is a well-known and important task of the pattern discovery process in today's world for finding the nucleotide or amino acid sequence patterns in protein sequence databases. Although pattern matching is commonly used in computer science, its applications cover a wide range, including in editors, information retrieval. In this paper we propose a new pattern matching algorithm that has an improved performance compare to the well-known algorithms in the literature so far. Our proposed algorithm has been evolved after the comparatively study of the well-known algorithms like Boyer Moore, Horspool and Raito. When we are talking about the overall performance of the proposed algorithm it has been improved using the shift provided by the Horspool search bad-character and by defining a fixed order of comparison. The proposed algorithm has been compared with other well-known algorithm [9].

III. PROPOSED SYSTEM ARCHITECTURE

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.

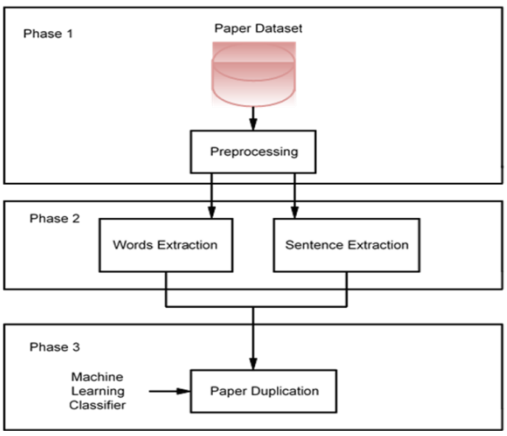

Fig 1. Proposed System Architecture

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

Algorithm:

1. 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. EXPERIMENTAL SETUP

Experiments are done by a personal computer with a configuration: Intel (R) Core (TM) i3-2120 CPU @ 3.30GHz, 4GB memory, Windows 7, Python, Dataset. The application is web application used tool for design code in Eclipse and execute on Tomcat server.

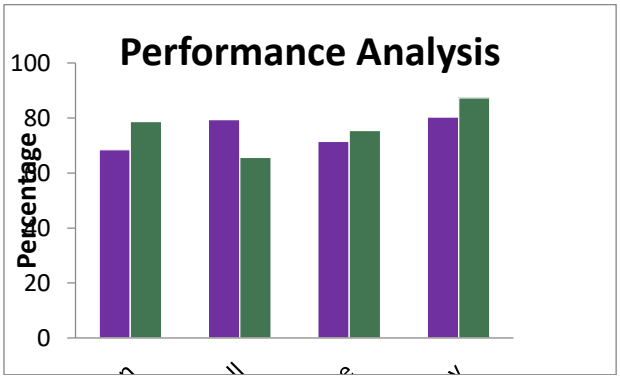

Fig 2. Comparison Graph

	Existing	Proposed
Precision	68.45	78.70
Recall	79.44	65.64
F-Measure	72.11	74.31
Accuracy	80.29	87.26
Execution Time (ms)	435	245

Table 1. Comparative Table

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

V. CONCLUSION

The proposed system is capable enough to handle the reduplication avoidance in projects and hence keep the reservoir of the same.

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 String Matching 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