IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Special Issue || Dec. 2019 || PP 27-30
International Conference on Management Practices, Innovations & Research 2019

A Review on Various Techniques for Duplication Detection in News Articles

[1]Komal Badge, [2]Jayant Adhikari
[1]*M.Tech Student, Department of Computer Science & Engineering, Tulsiramji Gaikwad Patil College of Engineering and Technology, Nagpur, Maharashtra, India.*
[2]*Assistant Professor, Department of Computer Science & Engineering, Tulsiramji Gaikwad Patil College of Engineering and Technology, Nagpur, Maharashtra, India.*

Abstract—*Identifying near duplicate documents is a challenge often faced in the field of information discovery. Unfortunately many algorithms that find near duplicate pairs of plain text documents perform poorly when used on web pages, where metadata and other extraneous information make that process much more difficult. If the content of the page (e.g., the body of a news article) can be extracted from the page, then the accuracy of the duplicate detection algorithms is greatly increased. Using machine learning techniques to identify the content portion of web pages, we achieve accuracy that is nearly identical to plain text and significantly better than simple heuristic approaches to content extraction. We performed these experiments on a small, but fully annotated corpus. Existing studies on news articles duplication detection mainly focus on newspaper articles, and we further explore the duplication detection in news articles from the newest online We Media data.*
Keywords—*Duplication Detection; News Article; PlagiarismDetection, Big Data*

I. Introduction

News articles published on the Internet typically appear on many different websites in either identical or revised form. For users, identical and nearly identical duplicates are an annoyance. Duplicates slow down the process of finding new information on a topic, and potentially cause missed information if the user mistakenly identifies two documents as identical duplicates when in fact one contains new information. For automated processing such as named entity recognition and visualization, redundant data can cause incorrectly weighted results, markedly skewing search engine results and automated text processing applications. While it is straightforward to find identical news stories in plain text documents, finding identical news stories embedded in web pages is considerably more complex. This is

due to the large amount of "extraneous" information, such as navigation links, ads, Javascript, and other miscellaneous content contained in these pages. While the actual news story text on two separate web pages may be identical, the extraneous content on the pages will not be. Thus standard approaches for determining identical duplicates will fail.

Previously, only press agencies could publish news articles, but now the advent of We Media makes everyone can publish and share news online. There lacks duplication detection and analysis based on the newest We Media data. Hence, we focus on the study of duplication detection in news articles from We Media which contain larger number of data.

II. Related Work

Duplicates are undesirable for many types of data. These include databases, mailing lists, file systems, email and image data. It is common practice to locate identical pieces of data using hashing strategies. Each piece of data is hashed using one of the standard algorithms, such as MD5. Any data represented by the same hash value is considered to be an identical duplicate.

Near duplicate documents are typically determined by computing a similarity score for each pair of documentsin a collection.

A. Cosine Similarity

To compute cosine similarity, documents are mapped into a vector space, typically based on term weights. The weight for each term is computed by the number of occurrences of the term in the document and an inverse measure of its frequency across a document collection. Document similarity is then measured by the cosine distance between the vectors.

B. Shingling

To compute the resemblance of two documents, each is broken into overlapping fragments called shingles. To do this, a shingle length,**a**, is specified. The first shingle is comprised of the first a words of the document. The second shingle consists of the second word in the document through the word located at **a+1**, and so on. Resemblance for the two documents is computed as the intersection size of the two documents' shingle sets divided by the size of the union of these sets. Let **A** and **B** denote the two sets of shingles for two distinct documents, then their resemblance is defined as:

$$r(A,B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|}$$

One of the main drawbacks to shingling is the massive number of shingles generated, especially for large documents. Several strategies are used to reduce the number of shingles, while only slightly reducing the effectiveness of the algorithm.

III. Literature Review

Lu Lu, and Pengcheng Wang in [1] introduced an article duplication recognizing strategy and executed as an apparatus, NDFinder. creators find that the best three subjects with the most elevated extent of duplication are Sports news, Military news, and Technology news. Besides, since articles duplication is straightforwardly identified with articles plagiarism, they effectively apply their examination way to deal with recognize 64 sets of plagiarism articles dependent on duplication results, and afterward, show observationally the plagiarism designs that found. Writers likewise approach their examination deal with article duplication and plagiarism indicator, which is valuable for duplication of the executives and plagiarism counteractive action of news articles.

S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, in [2] proposes the calculation to discover about Jaccard similitude coefficient by estimating the closeness in the right sentence structure linguistic structure and the trial of comparability as far as a blunder by building up the tests with Prolog programming language. Their exploratory outcomes demonstrated that the test strategy by the Jaccard coefficient can perform well in estimating the closeness of words when contrasting and each letter of the word. Especially, each letter can switch positions and considered similar words. By and by, there proposed technique can't distinguish the over-type words in the informational collections. Taking everything into account, the Jaccard comparability coefficient is appropriate adequately to be utilized in the word likeness estimation. In proficiency estimation, the program execution can manage high security when disappointment and mix-up spelling happened.

J. Agarwal in [3] clarifies the plagiarism of computerized archives, which appears to be a difficult issue in the present period. As per writers, plagiarism alludes to the utilization of somebody's information, language and composing without the appropriate affirmation of the first source. plagiarism of another creator's unique work is perhaps the most serious issue in distributing, science, and training. Plagiarism can be of various sorts. To determine these issue creators introduces an alternate methodology for estimating semantic similitude among words and their implications. Creators proposed new methodologies for identifying the plagiarism in the client record utilizing the semantic web. In paper creators have proposed engineering and calculations to better detection of duplicate case utilizing semantic hunt, it can improve the exhibition of duplicate case detection framework. It investigates the client archive.

H. Zhang and T. W. S. Chow in [4] propose not at all like prior strategies, STRUC techniques utilize logical data (i.e., topical squares, segments, and sections), which conveys the distinctive significance of content, and portrays various thoughts circulated all through the record. STRCU incorporated with VEC techniques have been utilized to identify reorder plagiarism, yet further research ought to be completed to examine the benefits of relating STRUC with SEM and FUZZY strategies for thought plagiarism detection.

M. Bautin, C. B. Ward, A. Patil, and S. S. Skiena as indicated by them in [5], the sociologies endeavor to comprehend the political, social, and social world around us, yet have been impeded by constrained access to the quantitative information sources appreciated by the hard sciences. Cautious examination of web report streams holds tremendous potential to take care of longstanding issues in an assortment of sociology teaches through monstrous information investigation. Their work presents the TextMap Access framework, which gives prepared access to an abundance of fascinating measurements on a great many individuals, spots, and things over various intriguing web corpora. Controlled by a flexible and versatile circulated measurement calculation structure utilizing Hadoop, ceaselessly refreshed corpora incorporate newspapers, sites, patent records, authoritative archives, and scientific abstracts; well over a terabyte of crude content and developing every day. Creators depict the design of the TextMap Access framework, and its effect on ebb and flow explore in political theory, humanism, and business/showcasing.

T. W. S. Chow and M. K. M. Rahman in [6], proposes printed highlights, which are basic to catch various sorts of plagiarism. Actualizing rich element structures should prompt the detection of more sorts of plagiarism if a legitimate technique and similitude measure are utilized too. level element extraction

incorporates lexical, syntactic, and semantic highlights, however, it doesn't account for relevant data of the archive. Auxiliary element extraction, then again, considers the manner in which words are disseminated all through the report. Creators order auxiliary highlights into square specific, which encodes the record as progressive squares, and substance specific, which encodes the substance as semantic-related structure. The last mentioned, joined with flat highlights, is reasonable to catch a record's semantics and get the essence of its ideas.

S. M. Alzahrani and N. Salim in [7], proposes the idea of Arabic language structure uncovered the requirement for fluffy or dubious idea to uncover deceptive practices in Arabic records. Creators present an announcement based plagiarism detection approach in Arabic contents utilizing fluffy set IR model. The level of closeness is determined and contrasted with a limit an incentive to pass judgment on whether two articulations are the equivalent or extraordinary. Our corpus assortment has been worked in which all stopwords were expelled and constant words were stemmed from average Arabic IR. Be that as it may, their Arabic fluffy set model methodology doesn't deal with the instance of rephrasing with various equivalent words/antonyms, an insufficiency that will prompt future work of displaying the framework utilizing Arabic thesaurus.

M. Roig in [8] proposes plagiarism can be jumbled by controlling the content and changing the majority of its appearance. Creators take a shot at lexical and linguistic rewording, rewording is the semantic significance requires references around the acquired thoughts and referring to the first creator other than rewording, abridging the content in a shorter structure utilizing sentence decrease, mix, rebuilding, summarizing, idea speculation and idea specification is another type of plagiarism except if it is referred to appropriately.

L. Lloyd, D. Kechagias, and S. Skiena in [9], clarifies the Lydia venture. The Lydia venture looks to construct a social model of individuals, spots, and things through normal language handling of news sources and the factual examination of element frequencies and co-areas. Lydia is still at a generally beginning period of improvement, however, it is as of now creating fascinating examination of significant volumes of content. Creators track the worldly and spatial dispersion of the substances in the news: who is being discussed, by whom, when, and where? Lydia is intended for the rapid investigation of online content. We look to dissect a huge number of curated content feeds every day. Lydia is equipped for recovering a day by day newspaper like The New York Times and afterward dissecting the subsequent stream of content in less than one moment of PC time.

IV. Implementation Details

Fig. 1 shows the system architecture of proposed system. In our proposed approach user inputs a single document for plagiarism checking. Initially pre-processing is performed on document in which unnecessary space within document, special characters, etc. are removed and then stopword removal process is performed in which the keywords such as a, an, the, numbers in documents & other stopword are removed. Then stemming processed is performed in which ing, ed, etc. of each keyword is removed. At the end only dictionary keywords are remain in input document.

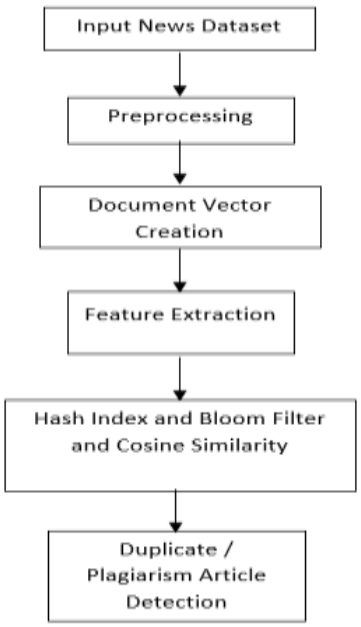

Fig. 1: System Architecture

After getting dictionary keyword from document, important keywords separated out (keywords having count greater than threshold k). These top k keyword set is passed to neural network classifier which performs classification on previously stored documents in database in two classes such as documents containing top k keywords (say class 1) and documents which don't contain top k keywords (say class 0). Then we use documents containing top k keywords (class 1) for further processing.

After this, the document vector of Input document and class 1 document is generated. Then TF-IDF of all document is generated and finally cosine similarity is calculated between input document and class 1 documents. If similarity is found between input document and any other document then input document is mark as plagiarism document and similarity percentage is calculated.

V. Conclusion and Future scope

In this paper we study the detecting plagiarism is very important not only in news article but also in industry, music, artwork etc. In particular, it has been shown in this study how the problem of plagiarism can be handled by using different techniques and tools. In this paper we saw that various software and tools are available for detecting plagiarism The comparison of the software and tools shown that still now their no software and tools that can detect or to prove that the document has been plagiarize 100%, because each software and tool has advantages and limitation, according to the features and performance described in the table. However there limitations in this software, tools which will affect the success of plagiarism detection significantly. We also presented our proposed approach for plagiarism detection.

References

[1]. Lu Lu, Pengcheng, Wang Duplication Detection in News Articles Based on Big Data, 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics
[2]. S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of Jaccard coefficient for keywords similarity," Proc. IMECS, 2017, pp. 380–384
[3]. J. Agarwal et al., "Intelligent plagiarism detection mechanism using semantic technology: A different approach," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, 2013, pp. 779-783.
[4]. doi: 10.1109/ICACCI.2013.6637273.
[5]. H. Zhang and T. W. S. Chow, "A coarse-to-fine framework to efficiently thwart plagiarism," Pattern Recog., vol. 44, pp. 471–487, 2015.
[6]. M. Bauti n, C. B. Ward, A. Patil, and S. S. Skiena, "Access: news and blog analysis for the social sciences," Proc. WWW, 2014, pp. 1229– 1232.
[7]. T. W. S. Chow and M. K. M. Rahman, "Multilayer SOM with tree structured data for efficient document retrieval and plagiarism detection," IEEE Trans. Neural Net w., vol. 20, no. 9, pp. 1385–1402, Sep. 2014.
[8]. S. M. Alzahrani and N. Salim, "Plagiarism detection in Arabic scripts using fuzzy information retrieval," In Student Conf. Res. Develop., Johor Bahru, Malaysia, 2013, pp. 281–285.
[9]. M. Roig, Avoiding Plagiarism, Self-Plagiarism, and Other Questionable Writing Practices: A Guide to Ethical Writing. New York: St. Johns Univ. Press, 2013.
[10]. L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A system for largescale news analysis," In International Symposium on String Processing and Information Retrieval, 2012, pp. 161–166.