A Survey On Plagiarism Detection In Documents

Anjumol Joseph1 & Rosna P Haroon2
1 M.Tech scholar, 2 Asst. Professor
1,2 Dept. of CSE, Ilahia College of Engineering & Technology Muvattupuzha, India

Abstract— The identification of duplicated and plagiarized passages of text has become an increasingly active area of research. In this paper, investigates methods for plagiarism detection that aim to identify potential sources of plagiarism in documents, particularly when the original text has been modified through the replacement of words or phrases. Here described the detection of plagiarism in Malayalam, English, Arabic, Hindi and Persian languages. The plagiarism detection is necessary in different language documents.

Keywords—Natural Language Processing, Malayalam, Arabic, Persian plagiarism detection

I. INTRODUCTION

Plagiarism is the "wrongful appropriation" and "stealing and publication" of another author's language, thoughts, ideas, or expressions and the representation of them as one's own original work. Plagiarism generally refers to the unacknowledged copying of existing information, such as documents and programs. This can include the reuse of one's own material (known as self-plagiarism), as well as that produced by others. Determining whether plagiarism has occurred is ultimately a human action; however, automated tools can assist with the process. Various factors can signal plagiarism, such as inconsistencies in writing style, unexpected use of advanced vocabulary, incorrect references and shared similarities with existing materials. Broadly speaking, approaches to detecting plagiarism (whether manual or automatic) can be categorised into two main problems. Intrinsic plagiarism detection relates to identifying stylistic inconsistencies within a text that give rise to questions regarding its authorship; extrinsic plagiarism detection relates to identifying the possible sources of a suspicious document.

Malayalam language is one among the four major Dravidian languages in south India and also one among the 22 scheduled languages in India. Malayalam is the official language of state of Kerala, Lakshadweep and Puducherry. According to the number of speakers, Malayalam ranks eight among the fifteen languages in India and is spoken by 37 million people. Malayalam is morphologically rich and is the mother tongue of three Crores of people in Kerala; Lakes of people living in remote area are poor in using English. The major problems of Malayalam are multiple suffixes, high inflections, tendency of adjacent words to concatenate etc. Malayalam is a highly agglutinative language and the morphological variations are more for the language compared to English. The nouns are inflected due to case, gender and number whereas the verbs are inflected due to tense, aspect and mood. Due to the morphological richness and complex nature of the language, thorough preprocessing is needed for Malayalam. Suffix separation is the most important preprocessing technique adopted in many of the NLP projects in Malayalam. The plagiarism detection in Malayalam documents is challenging task.

In English there many methods are available for the plagiarism detection. In this paper presented different methods that are used to detect the plagiarism in English documents. In one paper a particular field is taken, such as medical field. Here the Query expansion technique is applied, so the word sense disambiguation is also carried out.

Arabic language belongs to the Afro-Asian language group. It has much specificity which makes it very different from other Indo-European languages. Arabic language has twenty eight alphabet letters (١, ٢, ٣ ... ٩). Three of them are long vowels (أّ, ّ, ِّ, ٌّ) and the remaining ones are consonant letters. Arabic letters change shape according to their position in the word, and can be elongated by using a special dash between two letters. Arabic writing is right to left, cursive, and does not include capitalization. Diacritization or vocalization in Arabic consists in adding a symbol (a diacritic) above or below letters to indicate the proper pronunciation and meaning of a word. The absence of diacritization in most of Arabic electronic and printed media poses a real challenge for Arabic language understanding. Arabic is a pro-drop language: it allows subject pronouns to drop, like in Italian, Spanish, and Chinese. The language is highly inflectional. An Arabic word may be composed of a stem plus affixes (to refer to tense, gender, and/or number) and clitics (including some prepositions, conjunctions, determiners, and pronouns). Words are obtained by adding affixes to stems which are in turn obtained by adding affixes to roots.
 Persian, also known by its endonym Farsi, is one of the western Iranian language within the Indo-Iranian branch of the Indo-European language family, and the predominant modern descendant of Old Persian. It is primarily spoken in Iran, Afghanistan (officially known as Dari since 1958 for political reasons), and Tajikistan (officially known as Tajiki since the Soviet era for political reasons), and some other regions which historically were Persianate societies. There are approximately 110 million Persian speakers worldwide, with the language holding official status in Iran, Afghanistan, and Tajikistan. For centuries, Persian has also been a prestigious cultural language in other regions of Western Asia, Central Asia, and South Asia by the various empires based in the regions. Persian has had a considerable (mainly lexical) influence on neighboring languages, particularly the Turkic languages in Central Asia, Caucasian, and Anatolia, neighboring Iranian languages, as well as Armenian, Georgian, and Indo-Aryan languages, especially Urdu. It also exerted some influence on Arabic, particularly Bahraini Arabic, while borrowing much vocabulary from it after the Muslim conquest of Persia.

Modern standard or simply Hindi (Hindi: हिंदी hindī), is a standardised and sanskritised register of the Hindustani language. Hindi is one of the official languages of the Union of India, and the lingua franca of the Hindi belt languages. Hindi is the fourth-most spoken first language in the world, after Standard Mandarin Chinese, Spanish and English.

The next section deals with the related works done in different languages.

II. RELATED WORKS

A. An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE

In this paper it deals with the methods for plagiarism detection that aim to identify potential sources of plagiarism from MEDLINE, particularly when the original text has been modified through the replacement of words or phrases. MEDLINE (Medical Literature Analysis and Retrieval System Online) contains a large number of publications in the area of medicine and related fields. Here it is possible that people may reproduce the same research carried out by others without the connection being noticed and resulting in duplication. So to find out the plagiarism in the MEDLINE, it’s a two stage process. First step is the candidate document selection, in this step involves pre-processing, query formulation, retrieving and merging. Then these are followed by the query expansion technique. The query expansion is applied to deal with paraphrased cases of plagiarism, such as synonym changes, word sense disambiguation etc. Various approaches to Word Sense Disambiguation are investigated to deal with cases where there are multiple Concept Unique Identifiers (CUIs) for a given term.

B. Automated Plagiarism Detection System for Malayalam Text Documents

In this paper it deals with the document to document comparison. Here providing two documents as the input then normalizing the two documents by the following steps, Tokenization, stop words removal and followed by the Lemmatization phase. In Tokenization the input text from both the suspicious document and the original document are broken up into tokens. Tokens are usually words and is taken as a continuous string of characters which are separated by a space, line break, or punctuation characters. Stop words removal phase the commonly occurring words in documents like some verbs, adverbs and adjectives are treated as stop-words. They are removed in order to get more significant results. In the Lemmatization phase, the normalized form of a word is found. Lemmatization is similar to word stemming but it does not require to produce a stem of the word but to replace the suffix of a word, appearing in free text, with a different word suffix to get the normalized word form. After the normalisation phase then it is followed by the Word matching, here each normalized word is checked with the normalized word from text input box two. And a potential match is counted as weight of one. Next step is the Synonym match: When a non-matching word is found, it is replaced with its synonym to see whether a match is found. Then calculating the percentage of matching: In this phase a ratio between the two documents over the similarity and number of words is made. The comparison between documents can be then performed on the basis of standard similarity measures such as the Jaccard coefficient and the containment measure.

C. Design a Persian Automated Plagiarism Detector

In this paper the work is designed to detect the plagiarism in the Persian language documents. It involves Normalizing, stop words removal, sentence segmentation, Tokenization, stemming, Lemmatization, number replacement, these are the pre-processing steps. Then it is followed by the similarity comparison. In the number replacement step it will replaces any number with a dummy character. (“#” for example) The reason for doing this is that in some scientific reports, dishonest person can just change the numbers carefully to cheat the system.
In similarity comparison stage they have two sequences of words. One of the original text and one of the suspicious text. Then using the similarity measure comparing the texts, such as N-grams + Jaccard similarity coefficient or N-grams + Clough & Stevenson metric.

D. Detection of Plagiarism in Arabic Documents

In this paper, we present a plagiarism detection tool for comparison of Arabic documents to identify potential similarities. The tool is based on a new comparison algorithm that uses heuristics to compare suspect documents at different hierarchical levels to avoid unnecessary comparisons. It involves the following steps: Pre-processing, here the tokenization, stop-word removal, rooting, and synonym replacement is carried out. Second step is Fingerprinting: make use of k-grams, where k is a parameter chosen by the user. Third step is Document representation: for each document, create a document tree structure that describes its internal representation. And the last and fourth step is Selection of a similarity metric: use of a similarity metric to find the longest match of two hash strings. Each of these hash values of the pattern string is compared with the hash values of the text string. If the pattern and text hash values are equal, then there are matches between the corresponding pattern and text substrings.

E. Maulik: A Plagiarism Detection Tool for Hindi Documents

The objective of this paper is to present an automated plagiarism detection software tool called Maulik. There are many plagiarism detection tools available for English text. Maulik detects plagiarism in Hindi documents. Method: Maulik divides the text into n-grams and then matches it with the text present in repository as well as with documents present online. Before that the pre-processing taking, it involves the stop word removal and the stemming. Then converting the text into N grams and applying the synonyms, then next step is the Calculation of tf-idf and Similarity Score. Based on the similarity score we can identify the plagiarism.

F. Plagiarism Detection Algorithm Using Natural Language Processing Based On Grammar Analyzing

In this research, propose to use Natural Language Processing (NLP) to create the new way to detect plagiarism. Begin with using syntactic parsing method to parse the suspicious document and find the list of words which have the same meaning with it while considering the Part-Of-Speech (POS) element of that word (semantic parsing). This algorithm also includes creating the new structure of the object before comparing them. First is the suspicious document parsing, here the journal will be parsed into paragraphs then the paragraph will be parsed into sentences. Again the sentences will be parsed into words by using the POS tagging. The words with POS tag will be transformed into metadata, where each metadata represents each paragraph so a document will be represented by several metadata objects. Similarly the database journals are also doing the same steps. Then the detection processing side the similarity measuring by using the Jaccard Similarity Coefficient.

G. Plagiarism Detection using Semantic Analysis

The main objective of this research is to find a suitable way to detect semantic plagiarism which occurs on the meaning and making use of synonyms and replace it instead of the original words. And here the plagiarism detection is done only for documents that fall within the specialty of computer science. So in this paper an additional step is used to detect whether the given paper is fall in the specialty of computer science. The process of plagiarism detection involves mainly 3 steps, first the preprocessing stage, Document Specialization Stage and the semantic plagiarism detection stage. In preprocessing stage the Tokenization and stop words removal are carried out. In Document Disciplinary stage calculating the word frequency; arranging them in descending order, identifying the highest frequency, then identifying the relation between the highest frequency word and the computer science field. If it is under the computer science then checking semantic plagiarism.

H. Copy Detection System for Digital Documents

In this paper presenting a copy detection system for the detection of duplication in the digital documents. And it is done in 3 components. First component is the Document registration component: here given a document for input. Then it divides the document into its component sentences. Hashes each sentences and inserts each hash value into the database. And the next component is the Copy detection component: Divides the test documents into sentences and hashes. Apply algorithms to quickly compare each hash value of the test document with the database. Each matching hash value in the database is retrieved, sorted and grouped according to their document id. Last component is the Database component: It contain several tables and operations support two types of queries for a hash value X. One is return all(document id, sentence based offset relative to the
document) pairs whose hash value is X and return all sentences whose hash value is X.

I. SRL based Plagiarism Detection System for Malayalam Documents

This paper presents a method for plagiarism detection in Malayalam documents based on extracting the semantic roles and computing their similarity to detect plagiarism. The technique can detect documents created by direct copy methods, replacement of words with similar ones, changing the order of words or restructuring the sentences and also converting the sentence from active/passive to passive/active. First step is the preprocessing, it involves the text segmentation, stop words removal, Lemmatization and POS tagging. Second step is the vibhakti generation, this step classifies the words according to their vibhakti (case). A noun may belong to one of the seven cases namely, nirdesika (nominative), prathigrahika (accusative), samyojika (sociative), uddesika (dative), prayojika (instrumental), sambandika (genitive) and aadhaarika (locative). Third step is the semantic role labeling: Based on the vibhakti – Kaaraka relation, the word is tagged as kartaav (subject), karma (object), Karanam (instrumental), swami (dative), sakshi (sociative), and adhikaranam (locative). And final step is the similarity detection: sentence-based similarity analyses between the suspected and original documents are performed. Sentences in suspected documents are compared with each sentence in the candidate documents according to the verbs of the sentences.

II. CONCLUSION AND FUTURE WORK

This paper focuses on different plagiarism detection techniques for identify the unacknowledged copying of existing information, such as documents and programs. The literature survey made awareness about the different approaches that are used to detect the plagiarism, advantages and disadvantages of each approaches etc. This paper not arguing that here is a perfect collection of all the methods in plagiarism detection. Anyway this is a reference to the people who are working in the plagiarism detection or natural language processing area. The future work is: An IR-based Approach Utilising Query Expansion for Plagiarism Detection in Malayalam Documents. In this approach consists of mainly two stages, one Candidate document selection: here the identification of a subset of potential source documents given a suspicious text. For this using the Information Retrieval approach. And the second step is Detailed analysis, here the query expansion technique is used.

IV. ACKNOWLEDGMENT

We would like to thank all the faculty members and students of Ilahia college of Engineering and technology for their immense support. Also we like to express gratitude towards friends and family and all other good hearts for their motivation and support to this work come true.

REFERENCES


Applied Information Technology 10th May 2014, Vol. 63 No.1
