Automated Plagiarism Detection System for Malayalam Text Documents

Sindhu. L
Department of computer science
Cochin University of Science and Technology

Bindu Baby Thomas
Department of computer science
Cochin University of Science and Technology

Sumam Mary Idicula, Ph.D.
Department of computer science
Cochin University of Science and Technology

ABSTRACT

In this paper, a plagiarism detection tool for plagiarism detection in Malayalam documents is presented. Many language-sensitive tools for detecting plagiarism in natural language documents have been developed, particularly for English. Detecting plagiarism in Malayalam documents is particularly a challenging task because of the complex linguistic structure of Malayalam. The plagiarism detection tool presented here has the mechanism of detecting similarity beyond exact words match of Malayalam documents.. The tool is based on a new comparison algorithm that uses some NLP techniques to compare suspect documents which may not be identified using existing methods for Malayalam document plagiarism detection.

General Terms

Plagiarism Detection, Similarity Detection, Malayalam, Natural Language Processing

Keywords

Plagiarism Detection, Malayalam, Natural Language Processing, Lemmatization

1. INTRODUCTION

Plagiarism is defined as the use of someone else’s work, in whole or in part, into one’s own without adequate acknowledgement. Plagiarism has been around for as long as humans have produced work of art and research. The amount of text available in electronic media nowadays has caused cases of plagiarism to increase. So the detection of plagiarism manually is very tedious. Automated plagiarism detection systems are therefore very essential. Their main purpose is to assist people in detecting plagiarism. Most of the plagiarism detection systems do analyze the linguistic patterns for this purpose.. According to Martin (2004) , plagiarism can be classified as based on ideas, references, authorship, word by word, and paraphrase plagiarism. In the case of idea plagiarism, the ideas or thoughts of another person are claimed to be one’s own without proper citation. In the case of plagiarism of references and authorship, citations and entire documents are included without giving the names of their authors. The next case of word by word plagiarism is also known as copy–paste or verbatim copy. It consists of the exact copy of a part or the entire text from a source document into the plagiarized document. In paraphrase plagiarism, content from the source document is paraphrased and used in the plagiarized document. Hence, automatic plagiarism detection has significant importance in identifying the different types of plagiarism caused due to the easy accessibility of text over the internet.

There are two different approaches to automatic plagiarism detection. External or extrinsic plagiarism detection is based on detecting the similarity of the source document with the documents present in the reference text dataset. Intrinsic plagiarism detection approach is based on detecting the plagiarism that exists in a suspicious text itself without having a reference text dataset. The plagiarism detection approach in this work is based on the problem of detecting the plagiarized documents by making use of an existing reference text dataset.. Hence this proposed work is an extrinsic monolingual plagiarism detection approach which identifies whether the suspected documents are plagiarized versions of a given source document.

The rest of this paper is organized as follows: in section 2 the related work on plagiarism detection is presented. In section 3 the methodology chosen and the experimental settings are described. In section 4 the results from the experiments conducted are presented. Section 5 concludes the paper.

2. RELATED WORK

This section describes the existing plagiarism detection methods, which are mainly for English language. No detection tool is readily available for Malayalam document plagiarism detection. Plagiarism detection methods are classified in different ways. According to Lancaster, detection approaches can be classified by their type of detection methodology, availability of the system, number of documents the metrics can process and complexity of the metrics. The general approaches for existing plagiarism detection techniques are mainly non-NLP based.

Plagiarism can be either source code or free text plagiarism. The tools developed for source code plagiarism detection are Plagio Guard, JPlag, Moss, Sherlock etc. Available text based detection tools are Turnitin, Plagiarism Checker X, and Ferret for intra and extra corporal plagiarism detection.

Ceska, classifies existing methods for detection as Relative frequency models, Dotplot visualization models, Similarity measures model, Document fingerprinting, Word pairs metric. The use of a vector space model can detect similarity to determine cosine similarity among vectors of keywords or function-words extracted from the text under consideration. Other methods used for plagiarism detection are structural methods like multilevel text comparison, plagiarism pattern checker and statistical language models.

3. PROPOSED METHOD

In this section, the Malayalam language is described, give details about the corpus used, text processing techniques and the proposed detection method in this work.

3.1 Malayalam Language Characteristics

Malayalam language is one among the four major Dravidian languages in south India and also one among the 22 scheduled languages in India. It is mainly spoken by the people of Kerala state and the Union territories of Lakshadweep and Mahe. Malayalam is a morphologically rich and agglutinative Indian language. So it is very difficult to develop a computer system for Malayalam. The major problems are multiple suffixes, high inflecctions, tendancy 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.

3.2 Corpus Used

The attempt here is to detect external plagiarism in Malayalam text documents. External plagiarism detection requires both the suspicious plagiarized texts and the potential original source texts. Till date, no standard collection of texts for plagiarism detection in Malayalam language exists. Since no corpus of Malayalam documents is readily available, a corpus for this purpose was created. Plagiarized documents based on 5 original short passages were created. 50 plagiarized copies with various degrees of copy-and-paste, substituting words in the original with synonyms, making deletions and insertions to the original and 30 non-plagiarized versions were created.

3.3 Experimental Setup

It is found that syntactic structures do not always give optimal output in detecting plagiarism. So to detect plagiarism, a framework to evaluate text similarity with some language processing applied to both the original and plagiarized document propose is proposed.

Due to the complexities and certain drawbacks of existing plagiarism detection systems, the proposed model to detect plagiarism is by fetching word by word comparison. Then for the mismatched words, the system checks the synonym similarity between the mismatched words. This semantic matching technique is able to detect the use of synonym terms in the sentences.

Given the original and suspicious documents, different techniques have been applied to compare their similarities. Ferret performs trigram comparisons between original and suspicious document pairs and computes the similarity for each document pair based on the number of matching sequences of three words. Ferret has been applied for plagiarism detection in Malayalam documents (Sindhu et al.)

3.4 System Architecture

The system architecture used is described according to the block diagram. (Figure 1)

Normalization: In the Normalization phase the following preprocessing steps are done.

- In the Tokenization phase the 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.
- In the Stopwords 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. It reduces the size of the document. A list of stopwords in Malayalam were identified. These stopwords are removed from the text.
- 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. Lemmatization reduces the variants of the same root word to a common concept. No algorithm is readily available for lemmatization of Malayalam words. The main difficulty of word lemmatization of Malayalam is that Malayalam is a highly inflected natural language, having up to 56 different word forms for the same normalized verb. The root form of each word is identified as follows: -Check for suffix starting from the right end of the word using a suffix table. -remove the longest suffix found and make necessary changes to obtain the correct root form.

Word Matching: In this phase each normalised word is checked with the normalised word from text input box two. And a potential match is counted as weight of one

Synonym match: When a non-matching word is found, it is replaced with its synonym to see whether a match is found.

Calculate 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.

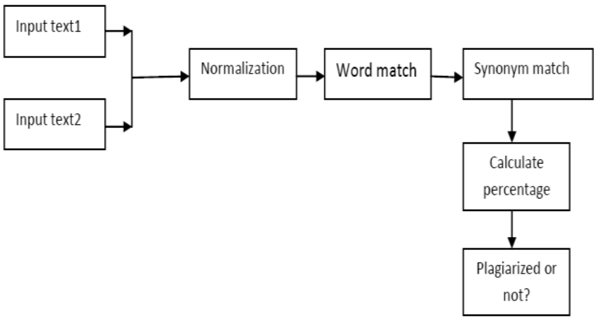


Figure 1: System architecture

3.5 Comparison Methodologies

The system takes trigrams in the experiments, as trigrams are found to be the balance between efficiency and effectiveness (Clough & Stevenson 2009)

Overlapping trigrams in sentences can be show as follows:

Original sentence

“Make hay while the sun shines.”

Trigrams {Make hay while} , {hay while the } , {while the sun} , {the sun shines}

Similarity between texts is measured by computing sets of trigrams for the suspicious and original texts and comparing these to determine the degree of overlap. A similarity function is used to measure the degree of overlap between the two texts represented by the set of trigrams and a threshold is chosen above which the texts are considered as plagiarised. The similarity measures used the detection system are the following:

Jaccard similarity coefficient:

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

where S(A) and S(B) represent the sets of trigrams in the suspicious and original documents respectively. The measure calculates the intersecting trigrams, and normalises it by the trigrams in the union which is the set of all trigrams in those documents.

The containment measure was used by Clough & Stevenson (2009)

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

where S(A) and S(B) represent the sets of trigrams in the suspicious and original documents respectively. The containment measure calculates the intersecting trigrams, but normalises by the trigrams in the suspicious document only. This measure is more suitable when document pairs are of different document lengths. (Broder, 1997) The original documents are usually longer than the plagiarized ones.

3.6 Evaluation Metrics

A set of standard performance measures that are widely used in information retrieval and text mining including confusion matrix and measures calculated from it, such as the percentage detection accuracy, precision, recall and F-measure are calculated. The confusion matrix shows the number of expected versus the obtained cases for each category. Accuracy is the percentage of correctly identified cases but it is not a good measure when the dataset is unbalanced.. Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. F-measure combines precision and recall using their geometric mean.

$$\text{Pre}cision = \frac{\text{Number of correct results}}{\text{Number of all returned results}}$$

$$Recall = \frac{\text{Number of correct results}}{\text{Total number of actual results}}$$

$$F-measure = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}}$$

4. DISCUSSIONS

The results of the experiments presented in this paper are considered satisfactory for classifying the documents in the corpus into two categories: plagiarised and non-plagiarised:

Table 1. Confusion matrix

obtained expected	**Similar**	**Not similar**
Similar	46	4
Not similar	5	25

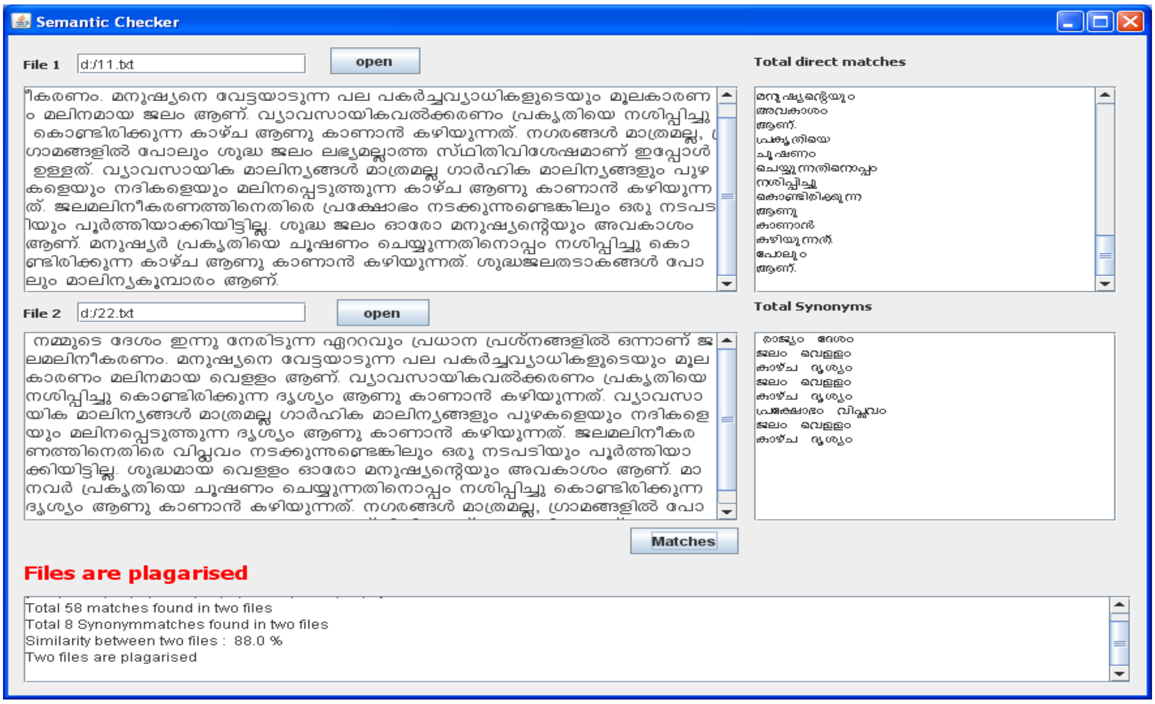
46 out of 50 plagiarised documents were correctly classified and only 5 nonplagiarized documents were classified as plagiarised. The system obtained 92% precision, 90% recall and 90% F-measure.

Experiments have proven that using this method can detect plagiarism by direct copying and also plagiarism by synonym replacement. But still final human judgement is essential to correctly judge the results obtained from the plagiarism detection system.

5. CONCLUSION

Information technologies bring the issue of digital plagiarism along with the benefits. This work aimed to develop a computer system to discover plagiarism in Malayalam document submissions. In this paper, a framework for automatic plagiarism detection for Malayalam documents has been proposed. Till date, no tool is available for checking for plagiarism with synonym replacement in Malayalam language. A contribution made as part of this work is the a lemmatization module which can used as part of any Malayalam language processing task. The results of the plagiarism detection system indicate that this system has the capability to detect exact copy and also changes caused due to synonym replacement.

One of the problems the system face is the non availability of possible sources to compare with the suspected documents. This limits the potential of algorithms that compute document-to-document similarity. As future work, it is intend to improve the detection accuracy by including other semantic and contextual features. Also the proposed method can be improved by using more documents.



File 1 d:/11.txt

```
ീകരണം. മനുഷ്യനെ വേട്ടയാടുന്ന പല പകർച്ചവ്യാധികളുടെയും മൂലകാരണ
ം മലിനമായ ജലം ആണ്. വ്യാവസായികവൽക്കരണം പ്രകൃതിയെ നശിപ്പിച്ചു
 കൊണ്ടിരിക്കുന്ന കാഴ്ച ആണു കാണാൻ കഴിയുന്നത്. നഗരങ്ങൾ മാത്രമല്ല, 
ഗാമങ്ങളിൽ പോലും ശുദ്ധ ജലം ലഭ്യമല്ലാത്ത സ്ഥിതിവിശേഷമാണ് ഇപ്പോൾ
 ഉള്ളത്. വ്യാവസായിക മാലിന്യങ്ങൾ മാത്രമല്ല ഗാർഹിക മാലിന്യങ്ങളും പുഴ
കളെയും നദികളെയും മലിനപ്പെടുത്തുന്ന കാഴ്ച ആണു കാണാൻ കഴിയുന്ന
ത്. ജലമലിനീകരണത്തിനെതിരെ പ്രക്ഷോഭം നടക്കുന്നുണ്ടെങ്കിലും ഒരു നടപട
ിയും പൂർത്തിയാക്കിയിട്ടില്ല. ശുദ്ധ ജലം ഓരോ മനുഷ്യന്റെയും അവകാശം
ആണ്. മനുഷ്യർ പ്രകൃതിയെ ചൂഷണം ചെയ്യുന്നതിനൊപ്പം നശിപ്പിച്ചു കൊ
ണ്ടിരിക്കുന്ന കാഴ്ച ആണു കാണാൻ കഴിയുന്നത്. ശുദ്ധജലതടാകങ്ങൾ പോ
ലും മാലിന്യകൂമ്പാരം ആണ്.
```

File 2 d:/22.txt

```
 നമ്മുടെ ദേശം ഇന്നു നേരിടുന്ന ഏറ്റവും പ്രധാന പ്രശ്നങ്ങളിൽ ഒന്നാണ് ജ
ലമലിനീകരണം. മനുഷ്യനെ വേട്ടയാടുന്ന പല പകർച്ചവ്യാധികളുടെയും മൂല
കാരണം മലിനമായ വെള്ളം ആണ്. വ്യാവസായികവൽക്കരണം പ്രകൃതിയെ
നശിപ്പിച്ചു കൊണ്ടിരിക്കുന്ന ദൃശ്യം ആണു കാണാൻ കഴിയുന്നത്. വ്യാവസാ
യിക മാലിന്യങ്ങൾ മാത്രമല്ല ഗാർഹിക മാലിന്യങ്ങളും പുഴകളെയും നദികളെ
യും മലിനപ്പെടുത്തുന്ന ദൃശ്യം ആണു കാണാൻ കഴിയുന്നത്. ജലമലിനീകര
ണത്തിനെതിരെ വിപ്ലവം നടക്കുന്നുണ്ടെങ്കിലും ഒരു നടപടിയും പൂർത്തിയാ
ക്കിയിട്ടില്ല. ശുദ്ധമായ വെള്ളം ഓരോ മനുഷ്യന്റെയും അവകാശം ആണ്. മാ
നവർ പ്രകൃതിയെ ചൂഷണം ചെയ്യുന്നതിനൊപ്പം നശിപ്പിച്ചു കൊണ്ടിരിക്കുന്ന
ദൃശ്യം ആണു കാണാൻ കഴിയുന്നത്. നഗരങ്ങൾ മാത്രമല്ല, ഗ്രാമങ്ങളിൽ പോ
```

Total direct matches

```
മനുഷ്യന്റെയും
അവകാശം
ആണ്.
പ്രകൃതിയെ
ചൂഷണം
ചെയ്യുന്നതിനൊപ്പം
നശിപ്പിച്ചു
കൊണ്ടിരിക്കുന്ന
ആണു
കാണാൻ
കഴിയുന്നത്.
പോലും
ആണ്.
```

Total Synonyms

```
 രാജ്യം   ദേശം
ജലം   വെള്ളം
കാഴ്ച   ദൃശ്യം
ജലം   വെള്ളം
കാഴ്ച   ദൃശ്യം
പ്രക്ഷോഭം   വിപ്ലവം
ജലം   വെള്ളം
കാഴ്ച   ദൃശ്യം
```

Files are plagarised

```
Total 58 matches found in two files
Total 8 Synonymmatches found in two files
Similarity between two files :  88.0 %
Two files are plagarised
```

Figure 2: Screen shot of similarity between two documents with direct copy and synonym replacement

6. REFERENCES

[1] Paul Clough 2000 Plagiarism in natural and programming languages: an overview of current tools and technologies, Department of Computer Science, University of Sheffield, technical report

[2] Clough, P. and Stevenson, M. 2009. Developing A Corpus of Plagiarised Short Answers, Language Resources and Evaluation: Special Issue on Plagiarism and Authorship Analysis, In Press. Journal Language Resources and Evaluation.

[3] C.Manning and H.Schutze. 1999, Foundation of Statistical Natural Language Processing , The MIT Press, Massachusetts Institute of technology , Cambridge, USA, ISBN 0-262-13360-1.

[4] Lancaster, T. and Culwin, F. 2007. Preserving academic integrityfighting against nonoriginality agencies. British Journal of Educational Technology. 38, 1 , 153-157.

[5] Z.Ceska. 2008. Plagiarism detection based on Singular value decomposition: Advances in Natural Language Processing 5221, 108-119.

[6] Shivakumar, N. and Garcia-Molina, H. 1995. SCAM: A copy detection mechanism for digital documents. Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries.

[7] Hoad, T.C. and Zobel, J. 2003. Methods for identifying versioned and plagiarized documents. Journal of the American Society for Information Science and Technology. 54, 3, 203–215.

[8] L Sindhu,.Bindu Baby Thomas, and Sumam Mary Idicula , 2013,,A Copy detection Method for Malayalam Text Documents using n-grams Model‟, National conference on Indian Language Computing ,Department of Computer Science, CUSAT.

[9] Sindhu. L, Thomas, B.B., and Idicula, S.M. 2011. A Study of Plagiarism Detection Tools and Technologies. International Journal of Advanced Research in Technology, vol. 1. pp. 64-70.