A Copy detection Method for Malayalam Text Documents using N-grams Model

Sindhu.L Bindu Baby Thomas Sumam Mary Idicula

Department of Computer Science

CUSAT

***Abstract*— In this paper a method of copy detection in short Malayalam text passages is proposed. Given two passages one as the source text and another as the copied text it is determined whether the second passage is plagiarized version of the source text. An algorithm for plagiarism detection using the n-gram model for word retrieval is developed and found tri-grams as the best model for comparing the Malayalam text. Based on the probability and the resemblance measures calculated from the n-gram comparison , the text is categorized on a threshold. Texts are compared by variable length n-gram(n={2,3,4}) comparisons. The experiments show that trigram model gives the average acceptable performance with affordable cost in terms of complexity.**

***Keywords*—Copy detection, N-gram Model, Bi-gram, Tri-gram , Malayalam, Plagiarism.**

I. INTRODUCTION

The availability of digital information has made it possible to use digital data which is also the cause of the misuse of the available data Some of the issues associated with the misuse of digital data are Plagiarism detection, ownership identification etc. Plagiarism detection is of particular interest to people in the academia and the publishing sector. Plagiarism means copying thought and text of another person and presenting them as ones's own work [1]. English copy detection systems have been studied since 1990s and some of them can be freely downloaded from the Internet while a copy detection system for Malayalam is not available.

II. MALAYALAM LANGUAGE

Malayalam belongs to the Dravidian family of languages and is one of the four major languages of this family with a rich literary tradition, inflectionally mainly adding of suffixes with the root or the stem word forms rich in morphology. The origin of Malayalam as a distinct language may be traced to the last quarter of 9th Century A.D. Malayalam is a language registering a heavy amount of agglutination. Throughout its gradual evolution malayalam has been influenced by the various circumstances prevailed on different periods. The important influence among these is the influence of Sanskrit and Prakrit brought into Kerala by Brahmins. In modern Malayalam also a good part of vocabulary is of Sanskritic origin. There are different spoken forms in Malayalam even though the literary dialect throughout Kerala is almost uniform.

III. FUNDAMENTALS OF PLAGIARISM

According to the Merriam-Webster Online Dictionary, to "plagiarize" means:

• To steal and pass off (the ideas or words of another) as one's own.

• To use (another's production) without crediting the source.

• To commit literary theft.

• To present as new and original an idea or product derived from an existing source.

• Also according to Turnitin.com and Research Resources this are considered plagiarism:

• Turning in someone else's work as your own.

• Copying words or ideas from someone else without giving credit.

• Failing to put a quotation in quotation marks.

• Giving incorrect information about the source of a quotation.

• Changing words but copying the sentence structure of a source without giving credit.

• Copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not.

Plagiarism can be classified into five categories:

- Copy & Paste Plagiarism.
- Word Switch Plagiarism.
- Style Plagiarism.
- Metaphor Plagiarism.
- Idea Plagiarism.
-

IV. APPROACHES IN PLAGIARISM DETECTION

Plagiarism detection is a way of ensuring quality in academic research. Plagiarism in academics is

considered as academic dishonesty and the responsible are subject to punishment by the university or the research funding organization. There are several approaches to plagiarism detection that have evolved. These approaches include natural language, statistical, ontological, Citation Based [6]. Approaches for detecting plagiarism in natural language include detecting similarity across multiple texts and within a text. The detection of plagiarism across multiple texts include searching for matching common substrings of length *n*, where *n* is chosen based on certain methods [7], [8], [9], and [10]. If n is made fixed then the substrings are said to be *n*-grams. The value of *n* may be different when retrieving subsequences from different parts of the document. The value of *n* cannot be big since not all content is usually copied verbatim from the source document. The detection of plagiarism within a text can be done through tracking the style of the author. Methods of detection originating from file comparison, information retrieval, authorship attribution, compression and copy detection have all been applied to the problem of plagiarism detection [11]. The similarity between texts based on the longest common subsequence , approximate string matching , the overlap of longest common substrings (eg: YAP3 [12],JPLAG [13], the proportion of shared content words, CopyCatch [14], the overlap of consecutive word sequences or word n-grams (e.g. Ferret [15], SCAM [8], COPS [7]).

V. EXISTING PLAGIARISM DETECTION TOOLS AND TECHNIQUES

Turnitin: This is a product from iParadigms [16]. It is a web based service. Detection and processing is done remotely. The user uploads the suspected document to the system database. The system creates a complete fingerprint of the document and stores it. Proprietary algorithms are used to query the three main sources: one is the current and extensively indexed archive of Internet with approximately 4.5 billion pages, books and journals in the ProQuestTM database; and 10 million documents already sub mitted to the Turnitin database [17].

SafeAssignment [http://safeassign.com/]: This web based service by Mydrop box. The system searches 300,000 documents that are known to be offered by Paper Mills. SafeAssignment also utilizes proprietary archives of institutional partners. Password protected and zipped archives can be indexed on demand. The service uses proprietary searching and ranking algorithms for match detection of fingerprints with its resources. The plagiarism detection result is presented to the user after a couple of minutes of submission.

Docoloc [http://www.docoloc.de]: This is a web based service which utilizes the searching and ranking capabilities of the Google API. The user of the service uploads the document that needs to be evaluated to a server. The software provides a simple console to set fingerprint (search fragments) size, date constraints, filtering and other report related options. The analysis report is sent to the browser or user's email identifying the matched fragments and internet sources.

CopyFind: [19]. While its goal is also to detect plagiarism, its search domain is more limited. Rather than finding borrowed text from the Internet, it compares a collection of student papers to each other. If enough similarity is found between two papers, the papers are flagged for further inspection. CopyFind has no ability to determine if sources from the Internet were used.

SCAM and CHECK [20] [21]: These two Plagiarism Tools deal with finding similarities among documents in a common database. Their main focus is on finding similar documents in a file system or other databases of digital media. They look at similarities of documents as a whole, not at individual sentences. Usually if the documents are no more similar than 25% the same documents are not flagged. These systems do not take any contextual similarities into consideration. This makes them easy to defeat by merely changing key words throughout the new document. The above discussed tools are available online in their respective sites.

Similarity between sentences (or more generally objects) can be captured numerically using similarity measures such as Jaccard similarity, Overlap similarity,Cosine similarity.

VI. PROPOSED PLAGIARISM DETECTION SYSTEM FOR MALAYALAM TEXT

A plagiarism detection system for Malayalam text passages based on the n-gram Model is proposed. This model uses n-grams for representing the text. N Gram Model was first used in text categorization based on the statistical information gathered from the usage of sequence of characters [4]. N grams are consecutive overlapping characters formed from an input stream. N-gram means that token of n words are used extracting the words from the passages and these n-grams matched. Then the resemblance measures are computed for text categorization.

In this approach, a document collection $D=\{D_1,D_2,D_3,\ldots\}$is used as reference corpus, and a suspicious(plagiarized) document collection $P=\{P_1,P_2,P_3,\ldots\}$ are used. Each suspicious document P_i will be compared to the documents in D . The resemblance measures are computed as Ferret as follows:

$$R\_score_n = \frac{|S(A)\cap S(B)|}{|S(A)\cup S(B)|} \quad (1)$$

Where S (A) is the set of n-gram from passage A, where $A\in D$, S (B)is the set of n-gram from passage B where $B\in P$. The Matched n-grams are calculated as $M = S(A) \cap S(B)$ (2)

And the total number of n-gram is computed as $N = S(A) \cup S(B)$ (3)

R_score_n is the resemblance metric of two documents when segmented by n, $0 \le R\_score_n \le 1$.
When $R\_score_n = 0$, document A and B have no identical n-gram, and if $R\_score_n = 1$, document A and B are identical. The smaller n is, the more candidates will be detected. T_n is the detect candidates of a given n.

$$Tn(B) = \{A \mid A \in D \text{ and } B \in P \text{ and } \mid Sn(A) \cap Sn(B) \mid \neq 0\} \quad (4)$$

The value of R ranges between 0 and 1 . We have set a threshold of 75% resemblance as the threshold for classifying text as plagiarized.

VII. EXPERIMENTS

1) Experiments with n-gram Model

We have used passages from the standard Malayalam online newspaper articles and also rephrased them. Now only documents in the .txt format is considered. N-grams for both the passages are calculated. To get the n-gram from the text, we process the text by following strategies: Firstly, we divide text into sentences by punctuations; Secondly, the non-malayalam characters in the sentences will be ignored. Finally, all the extracted sentences will be divided into n-grams (n=2, 3, 4, …). For a given sentence A,

നദികളും ധാരാളം കിണറുകളും കുളങ്ങളും ജലാശയങ്ങളുമൊക്കെയുള്ള കേരളത്തിന്റെ കാര്യംതന്നെ നോക്കൂ,

$S_n(A)$ the set of unique n-grams segmented from sentence A:

Table 1. Segmentation of text

n	$S_n(A)$
2	{ നദികളും ധാരാളം , ധാരാളം കിണറുകളും , കിണറുകളും കുളങ്ങളും , കുളങ്ങളും ജലാശയങ്ങളുമൊക്കെയുള്ള , ജലാശയങ്ങളുമൊക്കെയുള്ള കേരളത്തിന്റെ , കേരളത്തിന്റെ കാര്യംതന്നെ , കാര്യംതന്നെ നോക്കൂ }
3	{നദികളും ധാരാളം കിണറുകളും, ധാരാളം കിണറുകളും കുളങ്ങളും, കിണറുകളും കുളങ്ങളും ജലാശയങ്ങളുമൊക്കെയുള്ള , കുളങ്ങളും ജലാശയങ്ങളുമൊക്കെയുള്ള കേരളത്തിന്റെ , ജലാശയങ്ങളുമൊക്കെയുള്ള കേരളത്തിന്റെ കാര്യംതന്നെ , കേരളത്തിന്റെ കാര്യംതന്നെ നോക്കൂ }

A document collection $D=\{D_1,D_2,D_3,\ldots\}$is used as reference corpus, and a plagiarized document collection $P= \{P_1,P_2,P_3,\ldots\}$ are used. It is found that if a document has w words, then the number of bigrams, trigrams and fourgrams will be w-1, w-2 and w-3 respectively. Hence for a trigram search there will be a maximum of (w1-2)*(w2-2) comparisons where w1 and w2 are the number of trigrams generated from the reference and the plagiarized document respectively.

The tri-gram model is compared with other n-gram models to asses our selection of using tri-grams as the extracting word model. Copy detection with bi-gram model is the maximum but the complexity of extracting and comparing bi-gram is also the maximum. The copy detection rate of four-gram model is the smallest; it finds a very small number of matched four-grams as it compare longer sentences. The trigram model gives the average acceptable performance with affordable cost in terms of complexity and false alarms.

2) Comparison with other n-gram Models

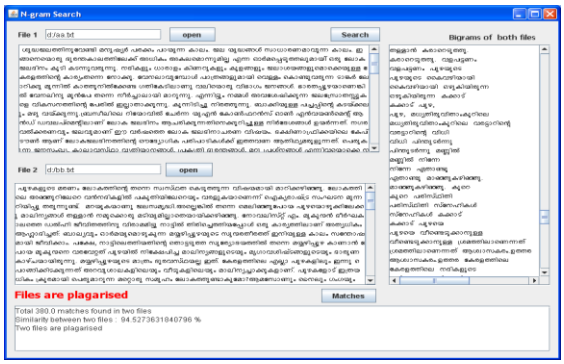

Figure1. A screen shot of Bigrams search

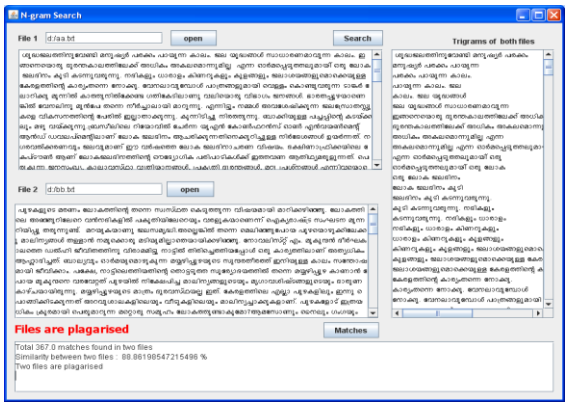

Figure2. A screen shot of Trigrams search

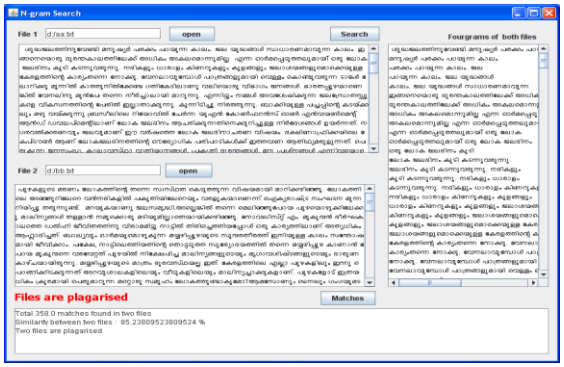

Figure3. A screen shot of Fourgrams search

3)Results and Discussions

We have performed bigram, trigram and fourgram comparisons on the same pairs of reference and plagiarized files and obtained the results as shown in table2.

Table 2. Percentage of plagiarism detected

	Ref	Plag-iarized	Bi-gram	Tri-gram	Four-gram
1	R1	P1	73.33	71.42	69.2
2	R2	P2	94.5	88.8	85.2
3	R3	P3	73.82	57.0	40.0
4	R4	P4	63.29	43.8	33.0

VIII. CONCLUSION & FUTURE WORK

In this paper we have presented a copy detection mechanism for Malayalam text using the trigrams as our word extraction model. Syntactic-based methods do not consider the meaning of words, phrases, or sentence. This is a major limitation of these methods in detecting some kinds of plagiarism. Nevertheless they can provide significant speedup gain comparing to semantic-based methods especially for large data sets since the comparison does not involve deeper analysis of the structure and/or the semantics of terms. We have used Resemblance measure R for computing the probability of the matching text. Based on a threshold the given text is categorized as plagiarized. To assess the validity of trigram model selection, it is compared with the bi-gram and four-gram models. In future, we plan to extend this work by (i) checking for plagiarism where words have been replaced by similar words (ii) checking text in the .doc format also.

IX. REFERENCES

[1]Muhammad A Khan, Abdul Aleem, Abdul Wahab, Nasir Khan M., "Copy detection in Urdu language documents using n-grams model",IEEE 2011

[2] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, Vaclav Snasel, "Survey of Plagiarism Detection Methods", Fifth Asia Modelling Symposium 2011

[3] A Barr'on-Cede-no,P. Rosso, "On Automatic Plagiarism Detection Based on n-Grams Comparison", ECRIC 2009, LNCS 5478, pp

[4] Cavnar, W. B., Trenkle, J. M., N-gram-Based Text Categorization, Symposium on Document Analysis and Information Retrieval, April 1994

[5] Ariho ohsato, Izumi Suzuki , Yoshi mikami., A language and character set determination method based on N-gram statistics, ACM Transactions on Asian language information processing,Sep 2002

[6] Bela Gipp and Jöran Beel, "Citation Based Plagiarism Detection - A New Approach to Identify Plagiarized Work Language Independently ", ACM 978-1-4503-0041-4/10/06, HT'10, pp 273-274, June 13–16, 2010,Toronto, Ontario, Canada.

[7] Brin, S., Davis, J. And Garcia-Molina, H. (1995), Copy Detection Mechanisms for Digital Documents, Proc. of the ACM SIGMOD International Conference on Management of Data, 398-409.

[8] Shivakumar, N. and Garcia-Molina, H. (1996), Building a Scalable and Accurate Copy Detection Mechanism, Proceedings of 1st ACM Conference on Digital Libraries DL'96.

[9] Lyon, C., Malcolm, J. and Dickerson, B. (2001), Detecting Short Passages of Similar Text in Large Document Collections, In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, 118-125.

[10] Broder, A. Z. (1998), On the resemblance and containment of documents, Compression and Complexity of Sequences, IEEE Computer Society.

[11] Clough, P.D. (2003), Measuring Text Reuse, PhD thesis, University of Sheffield CopyCatch product website,http//www.copycatchgold.com/

[12] Wise, M. (1996), YAP3 Improved Detection of Similarities in Computer Programs and Other Texts, Presented at SIGCSE'96, 130-134.

[13] Prechelt, L., Malpohl, G. and Philippsen, M. (2000), JPlag Finding plagiarisms among a set of programs, faculty of Informatics, University of Karlsruhe, Technical Report 2000-1.

[14] Woolls, D. and Coulthard, M. (1998), Tools for the Trade, Forensic Linguistics, Vol. 5(1), 33-57.

[15] Lyon, C., Malcolm, J. and Dickerson, B. (2001), Detecting Short Passages of Similar Text in Large Document Collections, In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing,118-125.

[16] iparadigms anti plagiarism product website, http//www.plagiarism.org/

[17] Turnitin, iParadigms, LLC. Turnitin- Plagiarism prevention engine.
Available online at http//www.turnitin.com

[18] Mydropbox, SafeAssignment Product Brochure, http//www.mydropbox.com/info/SafeAssignment_Standalone.pdf

[19] Plagiarism Finder1.2.2, m4-software.com Plagiarism.org- Research resources at plagiarism.org, http//www.plagiarism.org/research_site/e_what_is_plagiarism.html

[20] Robert Thomsen, D. M. Akbar Hussain, "Plagiarism Detection Based on SCAM Algorithm", Proceedings of the International multiconference of Engineers and computer scientists, vol 1,March 16-18, IMECS 2011,Honkong.

[21] Si, A., Leong, Lau, H., & Rynson W. (1997). CHECK A Document Plagiarism Detection System. ACM Symposium for Applied Computing, pp.70-77, Feb. 1997.