Rule Based Plagiarism Detection using Information Retrieval

Aniruddha Ghosh, Pinaki Bhaskar, Santanu Pal, SivajiBandyopadhyay

Department of Computer Science and Engineering,
Jadavpur University, Kolkata – 700032, India
{arghyaonline, pinaki.bhaskar,santanu.pal.ju}@gmail.com, sivaji_cse_ju@yahoo.com

Abstract. This paper reports about the development of a Plagiarism detection system as a part of the Plagiarism detection task in PAN 2011. The external plagiarism detection problem has been solved with the help of Nutch, an open source Information Retrieval (IR) system. The system contains three phases – knowledge preparation, candidate retrieval and plagiarism detection. From the source documents, knowledge base has been prepared for developing the Nutch index and the queries have been formed from the suspicious documents for submission to the Nutch IR system. The retrieved candidate source sentences are assigned similarity scores by Nutch. Dissimilarity score is assigned for each candidate sentence and the suspicious sentence. Each candidate source sentence is ranked based on these two scores. The top ranked candidate sentence is selected for each suspicious sentence.

Keywords: Plagiarism Detection, Information Retrieval System, Similarity Score, Dissimilarity Score.

1 Introduction

Plagiarism may be defined as the wrongful misuse and close replication of thoughts, ideas, or expressions from the original work of someone in the same language of from another language. From 18th century, plagiarism has been considered as academic dishonesty [1]. For decades, researchers have explored different techniques to detect plagiarism. Plagiarism can occur in different forms – full plagiarism, substantial plagiarism, minimalistic plagiarism, source citation etc. It has become a challenging task in the area of Natural Language Processing. In our approach, we have considered all the forms of plagiarism except minimalistic plagiarism at the sentence level.

Due to absence of controlled evaluation environment to compare results of the algorithms, plagiarism detection is still a challenging task [2]. Researchers have organized various conferences (similar to PAN) to overcome the plagiarism problem. Fingerprint retrieval method [3], candidate retrieval [4] and passage retrieval [5] are the most prominent attempts on plagiarism detection. The system described in [6] works with a natural language parser to find swapped words and phrases to detect intentional plagiarism while n-gram co-occurrence statistic is used to detect verbatim copy. The Longest Common Subsequence technique has been used in [7] to handle text modification. Researchers have used cosine similarity score and n-gram vector

space model at different levels, i.e., word [8] and character [9] levels. In the present work, plagiarism has been treated as an IR problem. An open source search engine, Nutch, has been used to retrieve the plagiarized parts from the suspicious documents.

2 System Framework

The Information Retrieval (Nutch[1]) based Plagiarism Detection system framework is shown in the figure 1. The system is defined in three phases: Knowledge Preparation, Candidate Retrieval, i.e., identification of suspicious sentence and the probable set of source sentence pairs and finally plagiarism detection of each identified suspicious sentence.

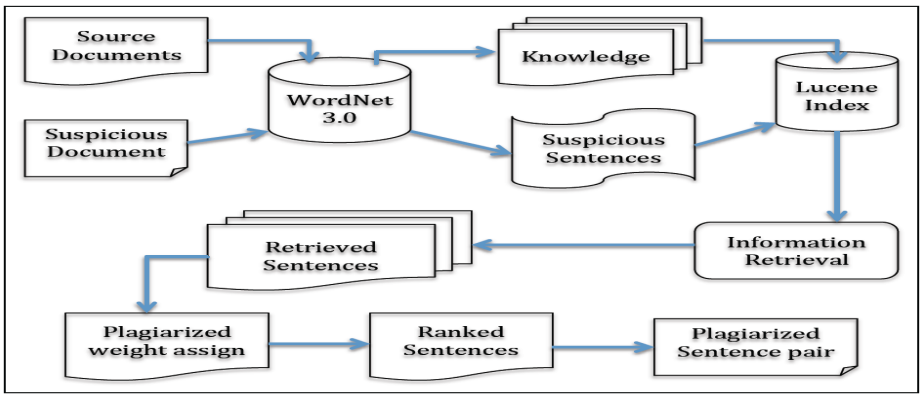

Fig. 1. System Architecture

3 Knowledge Preparation

Each source document is parsed to identify and extract all the sentences in the document. Now Knowledge files are generated for each source sentence. The file names of knowledge files are created in such a manner that the source sentence in the original source document can be tracked.

The knowledge of each sentence in the knowledge file is stored in the form of stems, synonyms, hyponyms, hypernyms and synsets of each word (after removal of the stop words) that are extracted from WordNet 3.0[2]. Duplicate words are removed to get the set of identical sense unique words. These words are used to identify the plagiarized words, the words that are similar in sense to the original words. The original words in the sentence are added to this set of words. Thus, each knowledge file for a sentence consists of a set of words. After all the knowledge files are built, these are indexed using Lucene[3].

[1]http://nutch.apache.org/
[2]http://wordnet.princeton.edu/
[3]http://lucene.apache.org/

4 Candidates Retrieval

Each suspicious document is parsed to identify and extract all the sentences in the suspicious documents. Each Suspicious sentence is considered from the parsed suspicious document to generate the query. First all the stop words are removed from the sentence and then the remaining words are being stemmed using WordNet 3.0 stemmer to get the root form of each word.

After generating the query from the suspicious sentences, the query is fired to Nutch to retrieve the probable set of source sentences corresponding to each suspicious sentence. As source documents are split into sentences into files and each file contains only one sentence, Nutch performs a sentence-sentence mapping for a proximal match between the query and indexed source files. A set of probable candidate source sentences is identified by Nutch in ranked order for each suspicious sentence. Nutch provides the similarity score between a suspicious sentence and the corresponding candidate source sentence.

5 Plagiarism Detection

An algorithm for dissimilarity measurement, proposed in [10], has been used to calculate the dissimilarity score between the suspicious sentence and its corresponding retrieved candidate sentences. For identical sentences that have most number of identical n-grams, the dissimilarity score is 0. Using this measure we have calculated the dissimilarity scores of each source sentence corresponding to the suspicious sentences.

The dissimilarity score are subtracted from the similarity score for each candidate source sentence and a final fine-grained score has been generated. All the retrieved candidate source sentences for each suspicious sentence are ranked according to this fine-grained score. The top ranked candidate source sentence is identified as the source sentence for the plagiarized sentence in the suspicious document.

6 Evaluation

The plagiarism detection system was evaluated using the evaluation framework described in [2]. The evaluation scores are shown in Table 1.

Table 1. Evaluation

Measurement	Precision	Recall	Granularity	Pladget
Score	0.0011829	0.0050052	2.0028818	0.0012063

7 Conclusion and Future Works

The present task is our first attempt in plagiarism detection. We have tested the plagiarism at the sentence level but phrase level experimentation is still left for investigate. In future, an algorithm has to be developed to test the relevance of the candidate source sentences retrieved by Nutch and choose the most relevant plagiarized part. The knowledge files for the source documents will also have to be updated.

Acknowledgment

The work has been carried out with support from Department of Information Technology (DIT), Govt. of India funded Project Development of "Cross Lingual Information Access (CLIA)" System Phase II.

References

1. Wikipedia article on Plagiarism: http://en.wikipedia.org/wiki/Plagiarism
2. Potthast M. et al.: An Evaluation Framework for Plagiarism Detection. In Proceedings of the COLING 2010, Beijing, China, August 2010.
3. Yurii Palkovskii, Alexei Belov and Irina Muzika.: Exploring Fingerprinting as External Plagiarism Detection Method: Lab Report for PAN at CLEF 2010. In Braschler et al.[2]. ISBN 978-88-904810-0-0.
4. Viviane P. Moreira, Rafael C. Pereira and Galante Renata.: UFRGS@PAN2010: Detecting External Plagiarism: Lab Report for Pan at CLEF 2010. In Braschler et al.[2]. ISBN 978-88-904810-0-0.
5. Clara Vania and Mirna Adriani.: External Plagiarism Detection Using Passage Similarities: Lab Report for PAN at CLEF 2010. In Braschler et al.[2]. ISBN 978-88-904810-0-0.
6. M. Mozgovoy, T. Kakkonen and E. Sutinen.: Using Natural Language Parsers in Plagiarism Detection. In Proceeding of SLaTE'07 Workshop, Pennsylvania, USA, October 2007.
7. Chen, Chien-Ying, Jen-Yuan Yeh and Hao-Ren Ke.: Plagiarism Detection using ROUGE and WordNet. Journal of Computing, 2(3), pages 34-44, March 2010. https://sites.google.com/site/journalofcomputing/. ISSN 2151-9617.
8. Cristian Grozea and Marius Popescu.: Encoplot - Performance in the Second International Plagiarism Detection Challenge: Lab Report for PAN at CLEF 2010. In Braschler et al.[2]. ISBN 978-88-904810-0-0.
9. Basile et al.: A Plagiarism Detection Procedure in Three Steps: Selection, Matches and "Squares". In Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 2009), Donostia-San Sebastian, Spain.
10. Vlado Keselj, Fuchun Peng, Nick Cercone and Calvin Thomas.: "N-gram-based Author Profiles for Authorship Attribution". In Proceedings of the PACLING'03, Dalhousie University, Halifax, Nova Scotia, Canada, pp. 255-264, August 2003.