Chapter 15
Latent Dirichlet Allocation and POS Tags Based Method for External Plagiarism Detection: LDA and POS Tags Based Plagiarism Detection

Ali Daud
King Abdulaziz University, Saudi Arabia & International Islamic University Islamabad, Pakistan

Jamal Ahmad Khan
International Islamic University Islamabad, Pakistan

Jamal Abdul Nasir
International Islamic University Islamabad, Pakistan

Rabeeh Ayaz Abbasi
King Abdulaziz University, Saudi Arabia & Quaid-i-Azam University, Pakistan

Naif Radi Aljohani
King Abdulaziz University, Saudi Arabia

Jalal S Alowibdi
University of Jeddah, Saudi Arabia

ABSTRACT

In this article we present a new semantic and syntactic-based method for external plagiarism detection. In the proposed approach, latent dirichlet allocation (LDA) and parts of speech (POS) tags are used together to detect plagiarism between the sample and a number of source documents. The basic hypothesis is that considering semantic and syntactic information between two text documents may improve the performance of the plagiarism detection task. Our method is based on two steps, naming, which is a pre-processing where we detect the topics from the sentences in documents using the LDA and convert each sentence in POS tags array; then a post processing step where the suspicious cases are verified purely on the basis of semantic rules. For two types of external plagiarism (copy and random obfuscation), we empirically compare our approach to the state-of-the-art N-gram based and stop-word N-gram based methods and observe significant improvements.

DOI: 10.4018/978-1-5225-8057-7.ch015

1. INTRODUCTION

The rapid growth of the internet has made it the largest publicly accessible information source of the world. Easy availability and access of documents have created a problem of plagiarism: copying others work to show others that the copied work is related to them without giving a reference to the original work. The problem of plagiarism is evident in academia. A large-scale study on 18,000 students shows that about 50% of the students plagiarized their work (McCabe et al., 2001). From exact document copy-paste (aka the verbatim), to paraphrasing or even translations from other languages, different forms of plagiarism happen in text documents (Stein et al., 2007). Developing an effective and automated tool for detecting plagiarism is a fascinating, practically useful, and challenging task.

External and intrinsic plagiarism detection are two main strategies for plagiarism detection (Stamatatos, 2011). External plagiarism detection is the approach to find passages in the suspicious documents against a set of possible source documents, whereas Intrinsic plagiarism detection aims at discovering plagiarism by inspecting only the input document without comparing it with possible source documents. We can define external plagiarism detection more formally as follows: Given a suspicious document, d, and set of source documents, SD, our goal is to find a set of passage pairs, P, such that,

$$P = < p_{di}, p_{SDj} > | \forall_{pdi}, \forall_{pSDj}: p_{di} \in \mathrm{d} \wedge p_{SDj} \in SD \wedge | p_{di} \cap p_{SDj}| > \delta \quad (1)$$

where, p_{di} is a passage from d, p_{SDj} is a passage from SD, and $p_{di} \cap p_{SDj}$ shows that similarity between p_{di} and P_{SDj} is greater than a threshold, δ, to consider as a plagiarism case. Similarity measure can be defined in many ways.

Usually the task of plagiarism detection comprises of three stages: text representation, similarity estimation (between a suspicious document and source documents), and extraction of sentences (plagiarized and the original). In the task of plagiarism detection, documents are typically represented by sequences of words or characters. Sliding windows of N-grams is the most popular method, which can be defined by the number of characters or the number of words (Schleimer, Wilkerson, & Aiken, 2003). Normally, windows of overlapping N-grams are generated. Overlapping N-grams requires more comparisons and hence gives better accuracy. Methods usually differ in the value of N. Representation can be based on content information (giving importance to important content terms) or structural information (giving importance to stop words). In principle, research works in document representation for plagiarism, can be classified in two categories, depending on the type of information or the features used to index the document terms: (1) content based information, and (2) structural information.

Works in the first category, e.g., (Gupta et al., 2010), give importance to important content terms, whereas works in the second category, e.g. (Stamatatos, 2011) take full advantage of stop word occurrences. So in second category, instead of eliminating stop words, they eliminate all the other tokens. Therefore, it is a method based exclusively on structural information rather than content information. In most of the cases, plagiarized passages are highly modified by changing word order, and words are replaced by synonyms. Because changing the basic syntactic structure is very difficult as compared to replacing synonyms. Hence, the need of using both syntactic and semantic information arises. The idea of using syntactic and semantic information to compute text similarity is well studied in many text mining tasks.

In this paper, we propose a novel external plagiarism detection method that hybrids the semantic and syntactic information by using Natural Language Processing (NLP) techniques. Latent Dirichlet Allocation (LDA) is used to capture semantic similarities, while Parts of Speech (POS) is used to capture