Journal of King Saud University – Computer and Information Sciences (2015) 27, 248–268

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model

Salha M. Alzahrani [a,*], Naomie Salim [b], Vasile Palade [c]

[a] College of Computers and Information Technology (CIT), Taif University, Taif, Saudi Arabia
[b] Faculty of Computer Science and Information Systems, University of Technology Malaysia, Johor, Malaysia
[c] Department of Computer Science, University of Oxford, UK

Received 13 August 2014; revised 24 October 2014; accepted 9 December 2014
Available online 27 June 2015

KEYWORDS

Feature extraction;
Fuzzy similarity;
Obfuscation;
Plagiarism detection;
Semantic similarity

Abstract Highly obfuscated plagiarism cases contain unseen and obfuscated texts, which pose difficulties when using existing plagiarism detection methods. A fuzzy semantic-based similarity model for uncovering obfuscated plagiarism is presented and compared with five state-of-the-art baselines. Semantic relatedness between words is studied based on the part-of-speech (POS) tags and WordNet-based similarity measures. Fuzzy-based rules are introduced to assess the semantic distance between source and suspicious texts of short lengths, which implement the semantic relatedness between words as a membership function to a fuzzy set. In order to minimize the number of false positives and false negatives, a learning method that combines a permission threshold and a variation threshold is used to decide true plagiarism cases. The proposed model and the baselines are evaluated on 99,033 ground-truth annotated cases extracted from different datasets, including 11,621 (11.7%) handmade paraphrases, 54,815 (55.4%) artificial plagiarism cases, and 32,578 (32.9%) plagiarism-free cases. We conduct extensive experimental verifications, including the study of the effects of different segmentations schemes and parameter settings. Results are assessed using precision, recall, *F*-measure and granularity on stratified 10-fold cross-validation data. The statistical analysis using paired *t*-tests shows that the proposed approach is statistically significant in comparison with the baselines, which demonstrates the competence of fuzzy semantic-based model to detect plagiarism cases beyond the literal plagiarism. Additionally, the analysis of variance (ANOVA) statistical test shows the effectiveness of different segmentation schemes used with the proposed approach.

1. Introduction

Plagiarism detection (PD) in natural language texts is one example of NLP applications that are linked with approaches from related fields, such as information retrieval (IR), data mining (DM), and soft computing (SC). PD research has focused on finding patterns of text that are illegally copied

* Corresponding author.
E-mail address: s.zahrani@tu.edu.sa (S.M. Alzahrani).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.12.001

from others. The easiest and common way to commit plagiarism is to copy and paste texts from digital resources. This is called literal plagiarism and is easy to spot by current PD methods. Unlike literal plagiarism, obfuscated plagiarism can be hardly seen because plagiarized texts are changed into different words and structure, or maybe into a different language.

Obfuscated plagiarism cases can be in the form of paraphrasing the original texts using different syntactical structures and lexical variations such as synonyms, antonyms, hypernyms, etc., but with no citation given to the original text. Plagiarism can be also hidden when the text is translated from one language to another with no credit to the original version, which is called cross-language plagiarism. Another form is summarized plagiarism, wherein long texts are briefed into shorter forms, which exclude details and keep the most important ideas in the source text, but with no accreditation given to the original source. In these exemplar forms of plagiarism, the texts are changed but ideas in the original texts remain unchanged. Appropriating an idea in whole or in part, with superficial modifications and obfuscations, in order to hide their sources without giving credit to its originator, is called idea plagiarism (Roig, 2006; Bouville, 2008).

Traditional techniques for PD depend on document similarity models such as duplicate detection (Elhadi and Al-Tobi, 2008, 2009) and bag-of-words related models (Barrón-Cedeño et al., 2009, 2010, 2009). Applications of document similarity, however, achieve the retrieval of a set of documents which have global similarity (at the document-level) with the query document from some source archive. The purpose of PD is not achieved yet via the document similarity, and a further detailed comparison between the query document and its candidate list should be carried out to report the local similarity (at the sentence-level, for instance). Exact and approximate string matching has been commonly used to compare two documents in-detail and find plagiarism. The documents are segmented into small comparison units such as character *n*-grams (Grozea et al., 2009), word *n*-grams (Barrón-Cedeño et al., 2009), or sentences (Alzahrani, 2009; Yerra and Ng, 2005; Zechner et al., 2009). An exhaustive matching is carried out, whereby matched *n*-grams (or sentences) that are adjacent to each other are combined into passages. Such methods are effective with verbatim plagiarism, yet not working with plagiarized texts that are literally different.

A recent literature review on the field of PD research (Alzahrani et al., 2012) has shown that there is a need for effective and efficient algorithms to find patterns of plagiarism that are semantically, but not literally, the same with original texts. Most of the current PD methods fail to detect obfuscated plagiarism cases because the similarity metrics of compared texts are computed without any knowledge of the linguistic and semantic structure of the texts (Ceska, 2007). Just a few methods have been developed based on a partial understanding of texts, e.g., when the words are replaced by synonyms, antonyms and hypernyms (Yerra and Ng, 2005). For example, Alzahrani and Salim (2010) presented a method to compute the similarity score between sentences based on the words and their synonyms. The method may be helpful to detect semantically similar texts, but should be further enhanced because not all synonyms relate to every meaning.

Recently, sentence similarity measures based on the semantic relatedness of their words have attracted researchers in different areas and for different applications, such as knowledge-based systems (Lee, 2011), text clustering (Shehata et al., 2010), text categorization (Luo et al., 2011), and text summarization (Binwahlan et al., 2010). A study by Lee (2011) proposed a semantic-based sentence similarity measure wherein two sentences can be compared based on a semantic space composed of a noun vector and a verb vector. A cosine similarity was computed between the noun vectors of two sentences and between the verb vectors of the sentences, which is further combined into a single similarity score. In Li et al. (2006), a sentence similarity measurement was presented based on the syntactic structures, semantic ontology and corpus statistics. Fernando and Stevenson (2008) presented a method to detect paraphrases of short lengths. A joint similarity matrix was constructed based on joint words from compared texts, wherein the similarity values between word pairs were calculated using different semantic similarity metrics.

In this paper, we propose a deep word analysis, in accordance with the WordNet lexical database (Miller, 1995), to detect similar, but not necessarily the same, passages. We focus on highly obfuscated plagiarism cases which are rephrased into another text without proper attribution to the original text. Unlike existing PD methods, which extract bag-of-words features (such as *n*-grams) without use of semantic features, we implemented a feature extraction method (FEM) which maintains the part-of-speech (POS) semantic spaces of the texts before further chunking of the text. Text segmentation is thereafter done using different schemes including word 3-gram, word 5-gram, word 8-gram with 3-word overlapping, and sentences. The purpose of using different segmentation schemes is to investigate which one works better along with the semantic features in the text. A fuzzy semantic-based approach is presented based on the assumption that words (from two compared texts) have a fuzzy (approximate or vague) similarity with fuzzy sets that contain words of the same meaning from a certain language. To fuzzify the relationship of word pairs (from text pairs), we proposed a WordNet-based semantic similarity metric as a fuzzy membership function. The fuzzy relationship between two words ranges between 1, for words that are identical or have the same meaning (i.e. synonyms), and 0 for words that are totally different (i.e., do not have any semantic relationship). A fuzzy inference system was constructed to evaluate the similarity of two texts and infer about plagiarism.

Experimental work was conducted on 99,033 various cases composed of handmade/simulated plagiarism cases, artificial plagiarism cases constructed automatically from some text documents and inserted into another, and plagiarism-free cases. Results of PD on those cases were assessed using precision, recall, F-measure and granularity averaged over 10-fold cross-validation data. The proposed approach was evaluated statistically against different state-of-the-art baselines using paired t-tests, which demonstrate the effectiveness of this approach to detect highly obfuscated plagiarism cases.

The remainder of this paper is organized as follows. Section 2 presents related work on semantic similarity measures based on lexical taxonomies such as WordNet, and overviews of related PD methods. Section 3 describes the feature extraction methods used in this study. Section 4 presents the proposed model for PD based on a fuzzy semantic model. In section 5, we discuss the experimental design including the datasets, baselines, parameters setting, evaluation metrics, the 10-fold cross-validation approach, and statistical analysis.