Detecting Academic Plagiarism with Graphs

Bin-Hui Chou, Einoshin Suzuki

Dept. Informatics, Kyushu University, Japan
chou@i.kyushu-u.ac.jp, suzuki@inf.kyushu-u.ac.jp

Abstract. In this paper, we tackle the problem of detecting academic plagiarism, which is considered as a severe problem owing to the convenience of online publishing. Typical information retrieval methods, stopword-based methods and fingerprinting methods, are commonly used to detect plagiarism by using the sequence of words as they appear in the article. As such, they fail to detect plagiarism when an author reconstructs a source article by re-ordering and recombining phrases. Because graph structure fits for representing relationships between entities, we propose a novel plagiarism detection method, in which we use graphs to represent documents by modeling grammatical relationships between words. Experimental results show that our proposed method outperforms two n-gram methods and increases recall values by 10 to 20%.

1 Introduction

Online publishing provides a platform for researchers to share their research results while it also brings a severe side effect, the academic plagiarism problem. That is, students or researchers copy all the content or a part of passages from others' papers without appropriate citation (Howard, 1995). It is difficult for editors of proceedings and journals to discover all the plagiarism behaviors due to the time limitation and the quantity of publications. An automatic detection method can be used to help editors' jobs and to mitigate the problem.

Existing methods of plagiarism detection evaluate document similarities by using content words (Gustafson et al., 2008; Hoad and Zobel, 2003), stopwords (Stamatatos, 2011) or document fingerprints (Seo and Croft, 2008; Schleimer, 2003). As common in information retrieval (IR), methods (Grman and Ravas, 2011; Gustafson et al., 2008; Hoad and Zobel, 2003) discard stopwords, e.g., "the", "is", and regard the remaining content words as meaningful words. This kind of methods use sequences of the content words to represent a document. Stamatatos (2011) considers that a plagiarist may replace content words to avoid detection, and proposed to represent documents by removing content words but retaining stopwords. Seo and Croft (2008); Schleimer (2003) use hashes of fixed-length chunks as document fingerprints.

Both directly copying and paraphrasing passages without citation are considered as academic plagiarism (Howard, 1995; Rosamond, 2002). Here, we aim to detect plagiarized documents where one paraphrases text from other documents by re-ordering phrases or altering modifiers. Most of the existing approaches use sequences of words as they appear in the document to represent the document so they fail to detect this kind of plagiarism.