Copy Detection Using Latent Semantic Similarity

De Cao Tran, Tri Cao Tran

College of Information and Communication Technology
Cantho – Vietnam
tcde@cit.ctu.edu.vn, tctri@cit.ctu.edu.vn

Abstract: Text copying is easy to do but not easy to be detected. Copy detection is a delicate ground. It determines whether a document is copied from others or not. It helps to verify and to detect paper redistribution, thesis plagiarism or copyright violation.

There have been some researches on different methods to determine whether a text segment of a document or a whole document is copied from other documents. These methods are mainly based on string matching. However, such methods work well on word for word copying. They are not efficient in terms of text-modification, i.e. by synonym replacement, paraphrasing or sentence restructuring.

This paper proposes a method to verify whether a document is copied from others by measuring the semantic similarity between the document under consideration and the registered documents. The technique used in this paper is the latent semantic indexing (LSI) which is based on the Vector Space Model (VSM) and the Singular Value Decomposition (SVD) algorithm. LSI technique is used to measure the semantic similarities between two paragraphs of two documents, then, the later is used to compute the similarities between two documents.

The experimental investigation shows that the proposed method works efficiently. A document copied from others with a few modifications - such as synonym replacement, word order changing and paraphrasing - can be identified.

Key words: copying detection, Latent Semantic Indexing, plagiarism, semantic search, similarity measure, text indexing.

I. Introduction

The growing of the Internet has brought with it many available documents. People can easily search for their desired documents and make copies instead of writing the documents themselves. The “copying” in the bad sense, such as plagiarism, redistribution of a document, copying papers, theses, etc. becomes widely. It is a serious problem in education as well as business.

Copying a text from a few documents in a digital library is very easy. Is it easy to detect a document copied from others registered in a digital library? How to determine whether a text is copied from others or not? If the copy is not absolutely the same (word for word), how to determine the degree of the copying?

This paper focuses on the detection of copying text document. The subject has been studied for about 10 years. There are methods and tools for detecting a text document that is copied from other documents. The documents may be in a closed source such as a digital library or an open source including all documents on the Internet.

Some well-known tools are COPS [1], SCAM [2,3], KOALA [4], and recently, Ferret’s solution for English [5], and Ferret’s solution for Chinese [6]. Basically, these methods are based on capturing the similarities between a chunk in a verified text (text under consideration for copying verification) and a chunk in a registered document. If the number of common (or similar) chunks in the verified text and in the source document over a predefined threshold, the verified text is considered as a copy from the source document. Otherwise, the verified text is not copied from the source documents. The “chunk” concept is different from method to method. For example, in SCAM, a chunk is a word; in COPS, it is a sentence (the definition of sentence is similar to English sentence concept) whereas in Ferret’s solution, a chunk is a trigram.

The existing methods work efficiently on detecting a copy of verbatim kind. However, they have difficulty in verifying and detecting plagiarism in which some modifications are made on the copied text by synonym replacement, paraphrasing, etc. In reality, plagiarism appears widely, and it is very difficult to be verified and detected. The difficulty is not only in detecting algorithm but also in the consensus about the “plagiarism” term. How can we define a plagiarism, and if it is, how can we determine the degree of copying or plagiarism?

This paper proposes a solution for detecting a plagiarism in a narrow sense. We do not intend to define plagiarism. The semantic copying used here means a copy with a bit modification such as synonym replacement, sentence restructuring (changing word order or rewriting a compound sentence instead of two simple sentences). The proposed method is based on latent semantic indexing (LSI). A document is modeled as a set of text segments, each segment is modeled as a vector, and then, the semantic similarities between two text segments are the cosine of two corresponding vectors. In other words, the vector space model is used for document modeling and the LSI technique is used to calculate the semantic similarities between two text segments.

II. Copying Detection Problem

Copying detection problem can be formalized as following:

Let T is a text document, called *verified text* or *verified*