A Fast Searching for Similar Text using Genomic Read Mapping Method

Submitted to CIT 2013

Abstract The most important consideration when detecting plagiarism is precision. Thus, the precise determination of the similarity of two documents is critical for the authors of documents. However, the problem complexity is increased by considering precision alone. Typically, the semantic detection of plagiarism has very high complexity, so a syntactic method for detecting plagiarism is used widely. The two main syntactic methods are sequence alignment and fingerprinting. Sequence alignment has powerful characteristics such as very high precision, because it is based on character-by-character comparisons. However, native sequence alignment has a high space complexity (O(n^2)). Fingerprinting is another syntactic method that uses the similarity of vectors extracted from documents. This method has a lower space complexity (O(n)) compared with sequence alignment. However, it also has lower precision because this method does not consider the structural similarity of documents. The method we propose for detecting plagiarized texts can detect plagiarism precisely, even with a low spatiotemporal complexity, by applying the short-read mapping method used for next-generation sequencing (NGS). In addition, we propose a distance measure for documents, which is based on the detection method used to construct phylogenetic tree by calculating the similarities of documents. The proposed method has a maximum precision of 0.95 and a maximum recall of 0.94. The construction of phylogenetic trees for linearly plagiarized documents using the distance measure had an average precision of 0.99. In the future, we will study the phylogeny of naturally plagiarized documents.

I. MOTIVATION

Recently, the interest in plagiarism has increased. Various studies of plagiarism detection are ongoing at present. These studies frequently apply sequence alignment and fingerprinting to small documents. However, these methods have drawbacks in terms of the document size that can be processed, the processing time, and detection precision. With fingerprinting, the processing time required to compare a document based on statistical comparisons is faster, but it is less precise. Sequence alignment makes character-by-character comparisons, so this method has an O(n^2) space complexity and an O(n^2) time complexity. Therefore, it is difficult to apply this method to large documents.

To overcome these drawbacks, we propose a method based on the concept of short-read mapping used for next-generation sequencing (NGS). To detect plagiarism in a space-efficient manner, this method is based on the Burrows-Wheeler Transform (BWT, also known as block-sorting compression) [16]. The output is easier to compress because it contains many repeated characters. For example, if the text T = "BANANAS" is input, T^0 is generated by T^0 = BWT(T) = "ANNBSAA" which includes the index information but the length of T^0 is almost the same as that of the original text T. This means that the space efficiency of BWT is very high. In addition, it is possible to restore T using T^0 by inverse BWT T = BWT^{-1}(T^0). The proposed method uses a data structure called the FM-index to reduce the space complexity during the search process. Next, to detect partially plagiarized texts, the query documents are divided into many reads using the short-read mapping method. Fig. 1 shows the process used by the proposed detection method.

The overall process flow of the proposed detection method is shown in Fig. 1. After preprocessing, the method builds the indexes using BWT and the FM-index as the data are inputted. After the query document has been input, the queries are divided using the short-read mapping method. The fragmented queries (reads) are used to search for partially similar texts in the reference document. Finally, the locations of the queries are computed to detect any plagiarized sections.

In addition to detecting the plagiarism of documents, we propose a method for measuring the distance between two documents in order to construct a phylogenetic tree for a plagiarized document set. This approach is based on the plagiarism detection method. For the distance measure, we...