

LEENA MAHMOUD LULU

Department of Computer Science and Software Engineering
College of Information Technology



Dissertation

Title	<i>A ROBUST SYSTEM FOR LOCAL REUSE DETECTION OF ARABIC TEXT ON THE WEB</i>
Faculty Advisor	Prof. Boumediene Belkhouche
Defense Date	5 December 2016
Abstract	We developed techniques and algorithms for finding text reuse on the Web, with an emphasis of the Arabic language. That is, our objective is to develop text reuse detection methods that can detect alternative versions of the same information and focus on exploring the feasibility of employing text reuse detection methods on the Web. The results of this research can be thought of as rich tools that may become essential parts in validating and assessing information coming from uncertain origins. These tools will prove useful for detecting reuse in scientific literature too. It is also the time for ordinary Web users to become Fact Inspectors by providing a tool that allows people to quickly check the validity and originality of statements and their sources, so they will be given the opportunity to perform their own assessment of information quality. For this purpose, we develop a novel technique to address the challenging problem of local text reuse detection from the Web. Given an input document d, the problem of local text reuse detection is to detect from a given documents collection, all the possible reused passages between d and the other documents. Selecting a subset of the documents that potentially contains reused text with d becomes a major step in the detection problem. In the setting of the Web, the search for such candidate source documents is usually performed through limited query interface. We developed a new efficient approach of query formulation to retrieve Arabic-based candidate source documents from the Web. The candidate documents are then fed to a local text reuse detection system for detailed similarity evaluation with d. Several techniques have been previously proposed for detecting text reuse, however, these techniques have been designed for relatively small and homogeneous collections. Furthermore, we are not aware of any actual previous work on Arabic text reuse detection on the Web. This is due to complexity of the Arabic language as well as the heterogeneity of the information contained on the Web and its large scale that makes the task of text reuse detection on the Web much more difficult than in relatively small and homogeneous collections. Our work to a certain degree is exploratory rather than definitive, in that this problem has not been investigated before for Arabic documents at the Web scale. However, our results show that the methods we described are applicable for Arabic-based reuse detection in practice.