Ph.D. Thesis Opponent’s Review

Thesis title: Data Compression Approach for Plagiarism Detection

Author: Hussein Khaled Hussein Soori

Reviewer: Jiří Dvorský

Overall Evaluation

This thesis focuses on creating textual plagiarism detection tools for detecting plagiarism of Arabic and Czech texts by implementing initial parts of a compression algorithm with its modifications where text similarity can be measured by compression-based similarity metrics. Next, it expands on this work by integrating this technique with a Czech synonyms thesaurus and a Czech stemmer to detect semantic plagiarism, including, paraphrasing and restructuring of Czech texts.

The demand of good language processing supported plagiarism detection tool goes even higher with languages spoken by wider population, such as Arabic language. The thesis presents a novel method for syllabification of Arabic text based on Arabic vowel letters. The thesis also presents a light stemming method for Arabic language. To fine-tune the results of this method, an online parser is used, before stemming, to better categorize the different parts of speech and, later, the output words are matched with an electronic dictionary.

Formal Structure and Organization of the Thesis

The thesis is conceptually divided into four chapters. The first chapter outlines the framework of the thesis. The second chapter covers two topics: syllabification method for Arabic language based on vowel letters and light stemming method for data compression. Chapter three is dedicated to plagiarism detection tools for both Arabic and Czech language. The last chapter concludes the thesis and presents conclusions about the author’s research. The thesis is well written with clear structure and using English at a high level.

Completion of the Thesis Objectives

Objectives of the thesis are:

- finding of applicable methods for plagiarism detection tools for Arabic and Czech texts using a similarity measurement based on LZ compression algorithm,
- semantic plagiarism detection method of texts including rewording, restructuring and using synonyms, and
- rules for syllabification of Arabic text using vowel letters, and a light stemming method, for data compression.

The objectives of the dissertation were fully met.
Remarks and Queries

One of the author’s experiments deals with student diploma works. Is it possible to compare results presented in the thesis with, for example, plagiarism detection at http://theses.cz/?

Publication Activities

The author provides, according to Scopus, a list of 8 publications, including one paper in Scopus journal. I think that the number of publications and their quality constitute a good standard for Ph.D. student.

Conclusion

I am convinced that the presented Ph.D. thesis represents good study providing valuable contribution to the state of scientific knowledge in the area. The author of the thesis proved the ability to conduct research and achieve scientific results. In accordance with par. 47, letter (a) of the Law Nr. 111/1998 (The Higher Education Act) I do recommended the thesis for the presentation and defense with the aim of receiving the Ph.D. degree.

Olomouc, August 31st 2016

Jiří Dvořák
Department of Computer Science
VŠB – Technical University of Ostrava