Ing. Nabil Ouddane, Ph.D.
Independent Researcher
Mateje Kopeckeho 606/22 Ostrava- Poruba,
Czech Republic.

Opponent review of Ph.D. dissertation of Mgr. Hussein Khaled Hussein Soori

Evaluated Ph.D. dissertation work of Mgr. Hussein Khaled Hussein Soori, clearly belongs to the field of his study. The work is divided into three main chapters, the objectives of the work and the conclusion. Load-bearing part of the work are represented in the chapters 2 and 3. The work contains a list of figures and tables clearly organized. There is referred to 87 pages, and titles in a review of the literature.

1. The fulfilment of dissertation objectives

This thesis is mainly about data compression. It is divided into two parts. The first part presents a novel method for syllabification of Arabic text based on Arabic vowel letters. It 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.

The second part 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.

In the first part, the syllabification method is based on Vowel Letters in Arabic. The results obtained are compared to different methods based on four different criteria for short and long files. For example, it is compared to character based, syllable based, word based, 2 grams and 3 grams based methods. The presented method showed good results especially in case of short1 and short4 files as presented in Table 6.

As for the light stemming method, the researcher first introduces his light stemming methods. After that the results are fine-tuned by using an online parser with a wider category with parts of speech. The results obtained show remarkable improvements. Although the results were not perfect due to some errors of categorization by the online parser and also some morphological reasons as described in the thesis, but they showed good results. They also showed the possibility to apply this method in light stemming for Arabic language.

In the second part, the thesis focuses on presenting a method used earlier on Slovak text where an initial part of the Lepel-Ziv compression method is used to detect textual plagiarism detection. But the main divergence in approach has to do with insertion of words and phrases rather than just characters in Lepel-Ziv method. Another difference is that the techniques is using only the first phase of Lempel-Ziv method pertaining to the input process to compare texts. A false suspicious document tool was created to conduct the experiment. The experiment is applied on Arabic texts with their specific features, including normalization of Arabic text. In addition to that, stop words were removed. In this experiment 71.42% of the documents were found to be plagiarized documents in 90 out of 126 documents, 28.58% of the documents were found to be partially plagiarized in 36 out of 126 documents, and 100% of the documents were found to be non-plagiarized in 24 out of 24 documents. The results were presented in Table 32.

The experiment is also applied on Czech texts where stop word removal is used. In this experiment 82.60% of the document were found to be plagiarized documents in 76 out of 92 documents, 17.40% of the documents were found to be partially plagiarized in 16 out of 92
documents, and 100.0% of the documents were found to be non-plagiarized in 8 out of 8 documents. The results were presented in Table 33.

Finally the experiment expanded to include semantic textual plagiarism of Czech students’ assignments where a Czech synonyms thesaurus and stemmer was added to the experiment to detect rewording and restructuring of Czech language texts. This work involves synonyms detection. In this experiment a corpus of 100GB was used and contained BA, MA, and PhD students’ assignments, semester works and theses from different majors. After removal of meta data and sorting the corpus into majors, semesters and academic years, the corpus size became 77GB with 13.671 files of plain text, out of which the computer science major students’ works were used to apply the experiment on. The total size of the extracted testing data used for our initial experiment was 1.98 GB of plain text. The Czech structured thesaurus and stemmer used were Slovnik českých synonym (Vocabulary of Czech synonyms) by Pala & Vsiansky. These thesaurus and stemmer are used currently by Apache OpenOffice for their Czech language pack. The thesaurus contains 386,891 tokens, 166,331 words and it is in UTF-8 format.

In the experiment on the short corpus, the proposed method was able to distinguish well between plagiarized and non-plagiarized texts in six versions. In the second long text corpus (1.98 GB of plain text), out of 883 document, 211 suspicious documents were detected with 23.9% suspicious tuples of documents as shown in Table 35.

2. Evaluation of the dissertation work:
The dissertation is well written with chapters divided in a logical manner according to the required standard. The literature review was extensive. The experiments were conducted and described clearly with illustrating tables and figures. For every experiment, the results were shown clearly and conclusions were drawn from the results in a straightforward manner.

It would have been better if some of the illustrating tables and figure were put in the appendices section at the back. For example, Table 14 on page 40 and Figures 15-19 on pages 68-70.

The evaluated dissertation work of Mgr. Hussein Khaled Hussein Soori meets the condition of separate creative scientific work, and contains the original author’s results and proposals, using different languages and surpassing the semantic differences, providing an academic tool operating in cross linguistic environment.

In general, efforts have been carefully provided to do this work surely in a clear and proper way, with the collaboration of academic publishing with highly reputed researchers from VSB-TUO, fulfilling the scientific standard, and providing an optimal tool, and useful for the academic field.

I recommend submitting this Ph.D. dissertation work for defense.

3. Question about the thesis:
1. What are the future developments you intend to work on pertaining to the semantic plagiarism detection method you proposed in your thesis?
2. Can we have any problems if we use an old Arabic poems or quotes with wide semantic meaning, and will the proposed method be efficient?

Respectfully submitted,

Ostrava on 05.09.2016

Ing. Nabil Ouiddane. Ph.D.