Review of the Doctoral Thesis entitled
"Utilization of Entropy in the Text Similarity"
of Ing. Michal Prilepok

The Ph.D. Thesis entitled “Utilization of Entropy in the Text Similarity” was presented by Ing. Michal Prilepok at VSB Technical University of Ostrava.

1. Structure and organization of the dissertation

This Ph.D. Thesis consists of six chapters, references and list of author’s publications.

In the first chapter “Introduction” was described the main aspects of author’s inspiration and fascination of this topic, and organization of this thesis.

In the second chapter “Entropy”, Ing. Michal Prilepok described many important topics as data compression and metric. He presented the aspects of similarity measure, compression-based similarity measures, Kolmogorov complexity, normalized information and compression distance, Lempel-Ziv complexity. This very good and correct development of entropy ideas, are used to analysed text similarity.

In the third chapter “Spam Detection” was described spam analysis, which include the many others kinds of filters. For example filter based on words, filter based on lists – black, white, grey lists, filter based on rules and learning, and the Bayesian Spam filter. Ing. Prilepok also analysed different aspect of spam detection and data compression, spam detection with signatures and data compression, and presented results of tested algorithms.

In the fourth chapter “Plagiarism Detection”, Ing. Michal Prilepok presented analysis of plagiarism detection by compression, text similarity for Czech and Arabic languages. Also he presented semantic and similarity measure methods dedicated to plagiarism detection. He described very good results of applying proposed methods. Those result Ing. Prilepok received in teamwork (vide list of Author’s publications).

In fifth chapter “Electroencephalography Data”, was described the main aspects of electroencephalography, analysis of EEG waves, short history as state-of-the-art of EEG problem, brain computer interface (BCI), asynchronous and synchronous interfaces. The very important aspects is related Ing. Prilepok’s works in EEG analysis. He described converting
EEG data into text data, process of identifying finger movements used by Lempel-Ziv algorithm. He also presented similarity kind of analysis of EEG data based on Self-Organizing Map. He received very good results for proposed methods, that give hope to continue research in future.

In last chapter “Conclusion”, Author summarized his works and achievements.

At the end of Ph.D. Thesis is the list of Author’s publications, which included 24 Ing. Prilepok’s publications. It’s very good result for scientist and researcher, who applying for a Ph.D. title.

The structure and organization of the Ph.D. Thesis by Ing. Michal Prilepok is very good. The dissertation is well-written, systematic and complete, with good English language.

2. Reviewer’s Assessments and Suggestions

The main thesis of this work has been shown at theoretical and practical point of view. Ing. Michal Prilepok described very good results of his works concentrated about entropy utilization in the text similarity. Furthermore, he presented his results at journals and international conferences too.

As remarks for this thesis, I have few questions.

1. In chapter “Entropy”, the formula (6), Author described the relative entropy as:

\[ D(p||q) = D(p(x)||q(y)) \]

I would like to ask, if this formula takes into account the dependence of variables \( x \) and \( y \), or this formula is dedicated only to analysis variable \( x \)?

2. In chapter “Metric”, the definition of metric, Authors describe the axiom of symmetry as:

\[ D(x, y) = D(x, y) \]

I would like to ask, if this formula refers to symmetry axiom?

It should be \( D(x, y) = D(y, x) \)

3. The title of a section “Kolmogorov Complexity”, p.14, should be “Kolmogorov Complexity”.

4. In chapter “Spam Detection and Data Compression”, Author describes the Bayesian Spam Filter and Normalized Compression Distance intervals, as:

\[ \text{BSF} > 0.5 \]
\[ 0.5 > \text{BSF} < 0.75 \]

and
\[ \text{NCD} > 0.5 \]
\[ 0.5 > \text{NCD} < 0.75 \]

I would like to ask, if this definition is appropriate to the indicated ranges?
5. In chapter “Preparing Command List for TG”, figure 25, Author described the angles of inclination, as \( \alpha_1, \alpha_2, \alpha_3 \)
   I would like to ask, if the angles \( \alpha_2 \) and \( \alpha_3 \) are different, or the same. The drawing shows that \( \alpha_2 \) as the same as \( \alpha_3 \), and \( \alpha_3 \) are different than \( \alpha_3 \). Please explain these aspects.

Please also explain, what is your most important motivation for concentrating on topic of your dissertation – entropy Utilization in the text similarity?

3. The overall evaluation of Ph.D. Thesis

The Ph.D. thesis by Michal Prilepok is a high quality dissertation, providing valuable contribution. Michal Prilepok demonstrates the ability to conduct scientific research individually and deserves the Ph.D. degree in Computer Science.

4. Conclusion

I recommend the thesis by Ing. Michal Prilepok for the presentation and defence with the aim of receiving the Ph.D. degree.

Krakow, May 28, 2016

Dr Lidia Ogiela
Department of Applied Computer Science
AGH University of Science and Technology