Anomalies Searching in Text Sequences

Abdulwahed Almarimi, Gabriela Andrejková and Peter Sedmák

Institute of Computer Science
Faculty of Science
Pavol Jozef Šafárik University

abdoalmarimi@gmail.com
gabriela.andrejkova@upjs.sk
peter.sedmak@student.upjs.sk

Each written text in some language has some author or more authors (authors have their individual sublanguage). An analysis of some text if authors are not known could be done using methods of data analysis and data mining, and using structural analysis. In the paper, it is described a system of modified *Self-Organizing Maps* [**?**]working on sequences built from a text. The system is trained to input sequences and after the training it determines text parts with anomalies using a cumulative error and complex analysis.

A given text written in some genre, in some language and grammar presents a sequence of letters, sequence of words, sentences, sections and we find some anomalies in this text (written in a time). For example, anomalies should show that some parts of the text were written (modified) by another author or that somebody manipulated with the text. The problem belongs to problems working with text and studying authorship attribution and plagiarism, but in both problems there exist some groups of comparable authors and comparable texts. It means the results of analysis can be compared according to texts and authors. In our problem any author is known and we analyze each text as one extra text.

Very good description of Self-Organizing Maps (SOM) extensions for temporal structures is in [**?**], some of the extensions are useable for sequences. SOM models of neural networks were applied to time series in [**?**] and it was some inspiration to use it in the text analysis. In the text analysis we used English recommended texts from benchmark [**?**] and Arabic texts from [**?**].

In this contribution, we applied a new system for anomalies detection in Arabic and English texts based on SOM model neural network. The system covers anomalies in texts consisting from two texts written by different authors.

References

[1] G. A. Barreto and L. Aguayo: Time series clustering for anomaly detection: Using competitive neural networks. Proceedings WSOM 2009, LNCS(5629), pp. 28-36, 2009.

[2] CorpusArabic: King saud university corpus of classical arabic. http://ksucorpus.ksu.edu.sa http://ksucorpus.ksu.edu.sa, 2011.

[3] CorpusEnglish: pan-plagiarism-corpus-2011. http://www.uniweimar.de/ en/media/chairs/ webis/corpora/ pan-pc-11/, 2011.

[4] B. Hammer, A. Micheli, N. Neubauer, A. Sperduti, and M. Strickert: Self-organizing maps for time series. WSOM 2005, Paris, pp. 1-8, 2005.

[5] T. Kohonen: Self Organizing Maps. Prentice-Hall, 2 ed., 2007.