AUTOMATE PLAGIARISM DETECTION

Eugeniu STRATULAT, Stanislav STROIANEȚKI, Victoria BOBICEV

[1] *Technical University of Moldova, Ștefan cel Mare, 168, Chișinău, Moldova*

*Victoria Bobicev victoria.bobicev@ia.utm.md

The paper presents a study in which an application for plagiarism detection has been created. It has been evaluated using the set of documents provided by PAN 2009 task on external plagiarism detection [1]. The task has been formulated as follows: Given a set of suspicious documents and a set of source documents the task is to find all text passages in the suspicious documents which have been plagiarized and the corresponding text passages in the source documents.

The organizers provided a training corpus which comprises a set of suspicious documents and a set of source documents. A suspicious document may contain plagiarized passages from one or more source documents.

The main metrics used for document comparison was NCD (Normalized Compression Distance) which is actually a family of functions which take as arguments two objects (some texts) and evaluate a fixed formula expressed in terms of the compressed versions of these objects, separately and combined [3]. The method is the outcome of a mathematical theoretical developments based on Kolmogorov complexity [4]. The smaller is the result, the more similar are the objects.

The application for plagiarism detection has been written in PHP. The similarity of two lines is calculated using the algorithm described in [2].

The selected threshold value has been estimated on the base of training data. This value provides the best plagiarism detection accuracy on the given texts.

In order to evaluate our application we used 400 documents from the set provided by the task organizers. We calculated Precision and Recall on 1/10 part of this set, namely, on 40 documents.

The information of the plagiarism in these 40 documents has been provided by the task organizers, so we knew exactly that only 5 of these 40 documents contained plagiarized fragments. The application returned exactly 5 files in which plagiarism was found. This result demonstrated that the application is good for the task.

Keywords: *plagiarism, automate plagiarism detection, text classification, substring search.*

References

1. POTTHAST M., STEIN B., EISELTA., WEIMAR B., Barrón-cedeño A., ROSSO P. Overview of the 1st International Competition on Plagiarism Detection. In SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), 2009, pp. 1-9.
2. OLIVER I., *Programming Classics: Implementing the World's Best Algorithms*. Prentice Hall, 1994.
3. CILIBRASI R., VITANYI P. M. B. Clustering by compression. In *IEEE Transactions on Information Theory*. 2011, 51 (4): 1523–1545.
4. BENETT C.H., GACS P., LI M., Vitányi P.M.B., ZUREK W., Information Distance, IEEE Trans. Inform. Theory, 1998 IT-44:4 1407–1423.