Plagiarims Detection Based On Three Steps

CS533 – Information Retrieval Course Project – Proposal Report

İsmail Uyanık & Deniz Kerimoğlu

Description of the problem:

Plagiarism refers to the use of another's information, language, or writing, when done without proper acknowledgment of the original source [1]. Availability of digital documents (for instance, easy access to the Web) and telecommunications in general provides good chances for plagiarism prosperity turning cheating into extremely easy and engaging process. There are different plagiarism methods. Some of them are [2]:

- copy-paste plagiarism (copying word to word textual information);
- paraphrasing (restating same content in different words);
- translated plagiarism (content translation and use without reference to original work);
- artistic plagiarism (presenting same work using different media: text, images etc.);
- idea plagiarism (using similar ideas which are not common knowledge);
- code plagiarism (using program codes without permission or reference);
- no proper use of quotation marks (failing to identify exact parts of borrowed content);
- misinformation of references (adding reference to incorrect or non existing source).

Plagiarism detection methods minimize plagiarism.

Motivation for plagiarism detection:

Nowadays plagiarism has turned into a serious problem for publishers, researchers and educators. 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.

Methodology:

The algorithm explained in [3] is for PAN competiton and it is successful in many aspects. By making improvement to this algorithm and inserting windowing method to the algorithm we can obtain better performance results.

Expected results:

We will test the algorithm which is run under PAN datasets and finally we expect to reveal most of the low level obfuscations and some of the low level obfuscations in the compared documents.

References:

[1] Wikipedia. Plagiarism. http://en.wikipedia.org/wiki/Plagiarism, 2005.

[2] R. Lukashenko, V. Graudina and J. Grundspenkis "Computer-based plagiarism detection methods and tools: an overview". CompSysTech '07 Proceedings of the 2007 international conference on Computer systems and technologies. ACM New York, NY, USA ©2007.

[3] Chiara Basile, Dario Benedetto, Emanuele Caglioti, Giampaolo Cristadoro, and Mirko Degli Esposti "A Plagiarism Detection Procedure in Three Steps: Selection, Matches and Squares". SEPLN´09 Workshop PAN Competition.