Improved implementation for finding text similarities in large collections of data

Notebook for PAN at CLEF 2011

Ján Grman and Rudolf Ravas
SVOP Ltd., Bratislava, Slovak Republic
{grman,ravas}@svop.sk

Solution background

2008

- Central Register of Thesis and Dissertations of the Slovak Republic – central repository for all academic institutions

2010

- Subsystem for comparison of documents and detection of plagiarism was added

Solution producer

SVOP Ltd.

- a producer and supplier of a library information system

2009 (march)

- first experiments with plagiarism detection solutions

2009 (dec.)

- creation of a commercial system

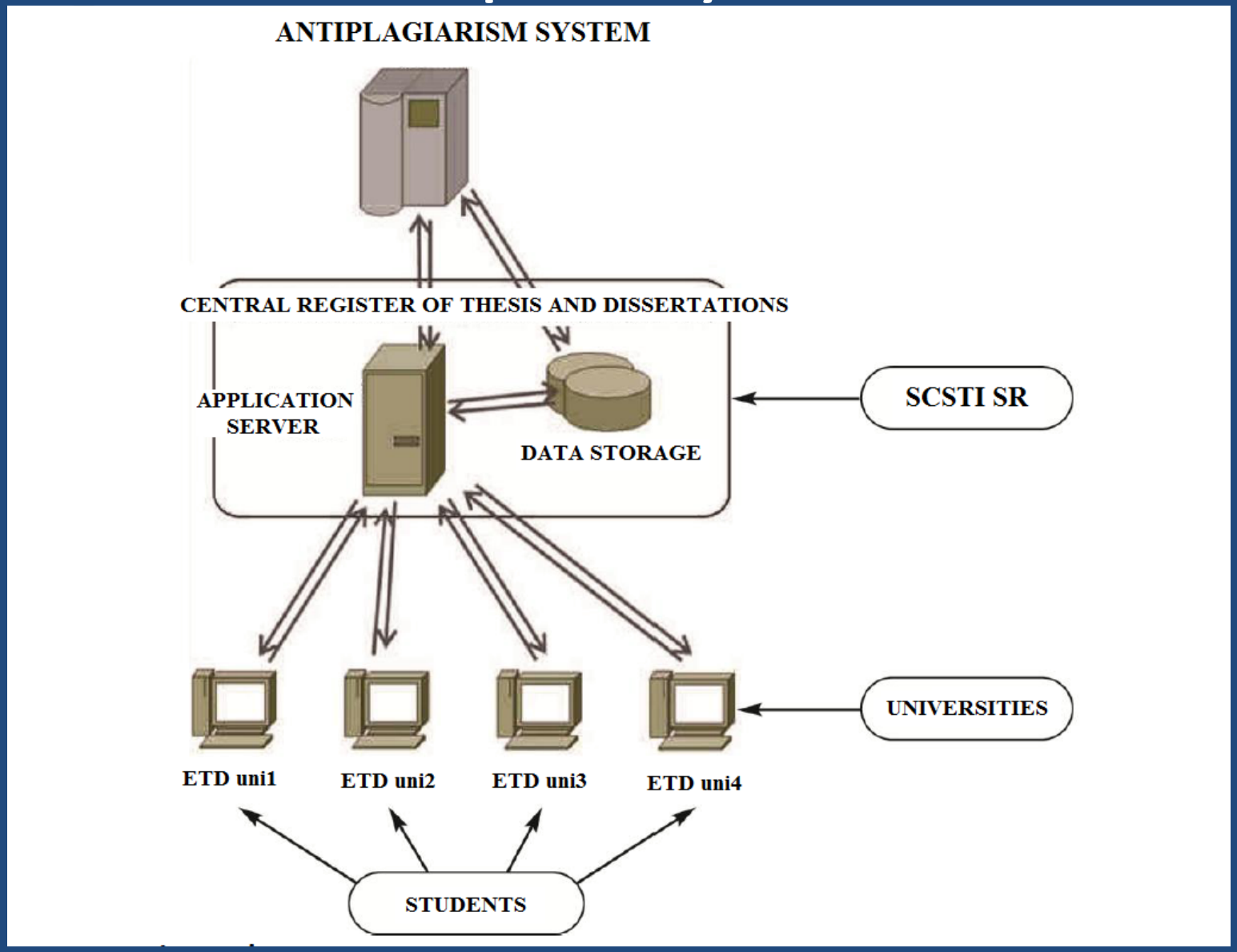
2010 (april)

- introduction of a antiplagiarism system in Slovakia

Basic parameters

- Complex evaluation of thesis and dissertations of all 33 universities in the Slovak Republic
- Language independent solution - documents in Slovak, Czech, Ukrainian, Hungarian and English
- Approx. 80 thousands of thesis and dissertations per year
- 3.4 milion documents from internet (06/2011)
- Core detection algorithm is now running on one server only (but parallel processing available)

Complex system



Method principle

- The presented new algorithm method removes numerous limitations of our older method, which has been used as part of a complex Information system for the detection of plagiarism
- In principle, the anti-plagiarism system we developed can be divided into three main parts, namely:
  - pre-processing of input data (in the case of PAN - plain-text pre-processing),
  - detection of passage pairs (plagiarism candidates) and
  - post-processing (removal of overlapping passages and exclusion of uncertain passage pairs)

Pre-processing (for PAN)

- Text translation into English (if needed - using Google API, language detector)
- Working with the text on the level of individual words / word extraction (3 parameters: chars, offset and length)
- Word normalization (stemming, synonym normalization – synonyms acquired from WordNet)
- Result: original text consisting of words is transformed into binary file of word invariants (codes) – one way transformation and reduction

Suspicious passage detection

Our objective was to create a method for detecting similar or matching passages in a suspicious and reference texts so that the detection is invariant

- against a change of word order,
- against the occurrence of changed words,
- against omissions or additions of words in the passage in a suspicious document, whereby no passage length limits will be set (neither minimum nor maximum length).

We assume that passage lengths don't have to be the same

Similarity

The method is based on quantification of the degree of concordance between tested passages. The **degree of concordance or similarity** is defined as the number of elements N_{MW} in an intersection of sets of words from passages in a suspicious and reference text.

$$N_{MW} = |I_S \cap I_R|$$

where N_{MW} is the number of matching words, I_S and I_R are the passages of the suspicious and reference text. The detector selects the area in which the value of N_{MW} exceeds the threshold N_{MWT} .

For all pairs of representations of suspicious and references documents, which were divided into non-overlapping passages (subintervals) with constant number of words were calculated number of matching words and were thresholded so that it can detect at least 15 words consistently.

Selection

In the first stage, if the detected areas are adjacent, then they are merged into a single area. After that, the areas are divided into disjunct areas (pair of passages) so that the resulting passages have the following property.

Let's mark the sub-passages l_{Si} and l_{Rj} of passages l_S and l_R, which either start or end in a word belonging to the set (intersect words).

If the ratios

exceed the selected threshold q_{min}, then the pair l_{Si} , l_{Rj} becomes plagiarism candidate passages for the validity of the assumption where N_{MWT1} is the minimum matching words of the detected passage.

We used q_{min}=0.5 and N_{MWT1}=15.

Post-processing

In our case, two tasks were solved removal of overlapping passages in suspicious document, if source text was the same and increasing of global score by reducing granularity and by increasing precision using our methods (results after removal of overlapping passages were thresholded to three monitored quantities), such as

- T1 - threshold to average ratio of matched word number to number all words of passages in the suspicious and reference text
- T2 - threshold to average ratio of length sum matched word to length sum all words of passages in the suspicious and reference text
- T3 - threshold to minimum length of passages expressed by the number of characters

PAN-10 testing

Plagiarism detection score in PAN-10 (with synonyms and without stop-words) for different threshold settings for parameters T1, T2 and T3.

Row one shows the score for results without post-processing (marked **).

PlagDet	Recall	Precision	Granularity	T_1	T_2	T_3
0.433957	0.737183	0.312248	1.015155	**		
0.811796	0.733454	0.910356	1.001009	50	50	150
0.812908	0.733206	0.913456	1.000951	60	50	150
0.82334	0.730341	0.944667	1.000761	70	50	150
0.823852	0.729678	0.947132	1.000762	70	60	150
0.824488	0.726819	0.953666	1.000746	70	60	200

PAN-11 testing

Plagiarism detection score in PAN-11 (using synonyms without stop-words) for different threshold settings for parameters T1, T2, T3.

The plagiarism detection results in the PAN-11 corpus can be described using two statements: satisfaction with the achieved rank and dissatisfaction with the achieved score.

PlagDet	Recall	Precision	Granularity	T_1	T_2	T_3	Cases
0.5569	0.39692	0.93802	1.002249	70	60	200	22108
0.61539	0.47313	0.89274	1.006975	50	50	150	28781

Optimal parameters

With known correct results it is easy to set suitable system parameters.

In our case we have used the post-processing settings for PAN-11 which produced the best results for PAN-10.

Conclusions

Each plagiarism detection method faces one basic problem – a huge amount of data. That means only methods that are capable of processing a certain amount of data within a reasonable time limit are usable. The PAN-11 was equally processed using a single server and even several times within the given short period of time (about 12 hours - one run).

The main advantages of the new method are better opportunity of detecting paraphrased text, extended support for different word forms, significantly improved detection reliability for texts translated from foreign languages (translation through individual paragraphs, offset alignment of paragraphs – original and translated).

Conclusions

Of course, there are some issues that all creators of complex systems face. The basic one is the definition of plagiarism. How much identical and/or similar text can already be considered plagiarism. Should the computer decide, or should it just be a tool that helps decide?

We would like to thank the competition organizers and the authors of the test corpus for the excellent opportunity to obtain a relatively objective and independent view of the detection capabilities of our solution.

Future

- Adaptive parameters
- Preprocessing improvements
- Specific hardware configurations
  - grids of small and connected computers

Thank you

Ján Grman and Rudolf Ravas
SVOP Ltd., Bratislava, Slovak Republic
{grman,ravas}@svop.sk