Diverse Queries and Feature Type Selection for Plagiarism Discovery

Šimon Suchomel, Jan Kasprzak, and Michal Brandejs

Faculty of Informatics, Masaryk University, Brno, Czech Republic

1. Introduction

A program for helping detering real-world plagiarism needs to accomplish many tasks. Original documents which served for creation of plagiarism must be retrieved and also suspicious passages according to input document must be highlighted. This poster presents methodology used during PAN2013 competition on uncovering plagiarism.

The whole process is depicted at picture 1. The source retrieval task is divided into 2 subtasks: Quering and Selecting, during which the software utilizes a given search engine. The retrieved sources must be examined in detail in order to highlight as many plagiarism cases as possible. This process is depicted as Text Alignment. Results of this process are called *detections*, i.e. passages of *source document* and *suspicious document*, which are similar enough to each other, and can serve as a basis for further manual examination for possible plagiarism.

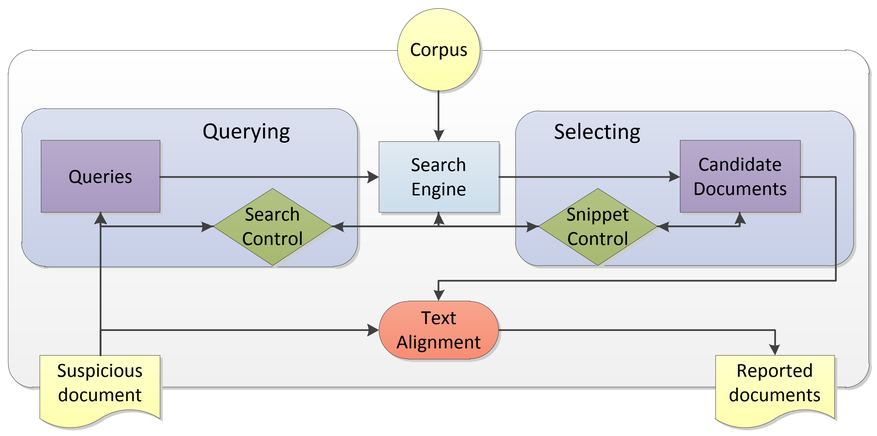


Figure 1: *Plagiarism discovery process.*

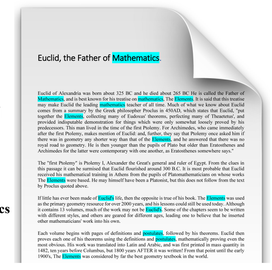
2. Querying

Querying means to effectively utilize a search engine in order to retrieve as many relevant documents as possible with the minimum amount of queries. In real-world, queries as such represent appreciable cost, therefore their quantity minimization should be one of the top priorities. During initial phase, there were three diverse types of queries extracted from each suspicious document.

2.1 Keywords Based Queries

- TF–IDF base automated keywords extraction;
- 5-token long;
- Deterministic;
- Non-positional;
- Non-phrasal.

$$score(w) = TF_{w,d} \cdot IDF_w$$
$$TF_w = \frac{f(w)}{\sum_k f(w_k)}$$
$$IDF_w = log\frac{|D|}{1 + |d : w \in d|}$$
$$D = reference\ corpus$$

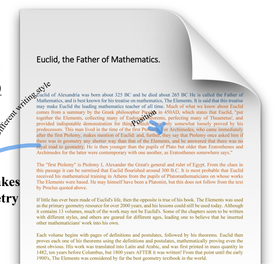


Query:
euclid's postulate elements parallel mathematics

2.2 Intrinsic Plagiarism Based Queries

- Averaged Word Frequency Class based chunking [1];
- Random sentence selection from the chunk;
- Non-deterministic;
- Positional;
- Phrasal.

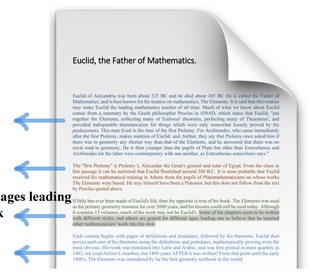
$$c(w) = \left\lfloor \frac{log_2(f(w^*))}{f(w)} \right\rfloor$$
$$w^* = the\ most\ frequent\ word$$
$$AWFC(chunk) = \frac{\sum_{i=1}^{|chunk|} w(c_i)}{|w(c)|}$$



Query:
Archimedes immediately Ptolemy makes mention Euclid Ptolemy asked geometry

2.3 Paragraph Based Queries

- Longest sentences from miscellaneous paragraphs;
- Deterministic;
- Positional;
- Phrasal.



Query:
chapters written styles geared ages leading inserted mathematicians' work

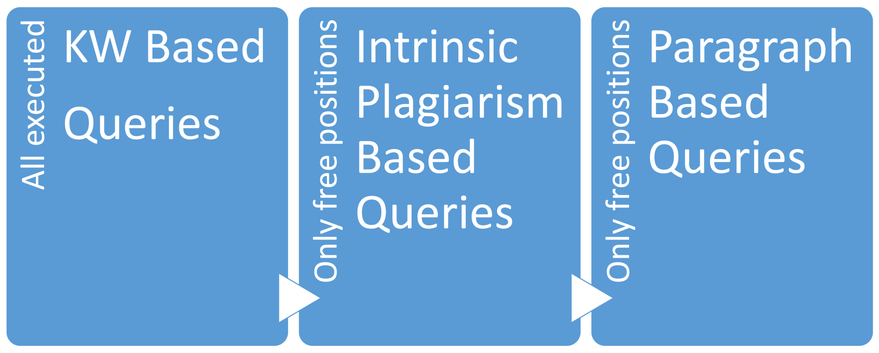


Figure 5: *Stepwise queries execution process.*

3. Selecting

Document snippets were used for deciding whether to download the document for the text alignment. We used 2-tuples measurement, which indicates how many neighbouring word pairs coexist in the snippet and in the suspicious document. Performance of this measure is depicted at Figure 6. Having this measure, a threshold for download decision needs to be set in order to maximize all discovered similarities and minimize total downloads. A profitable threshold is such that matches with the largest distance between those two curves.

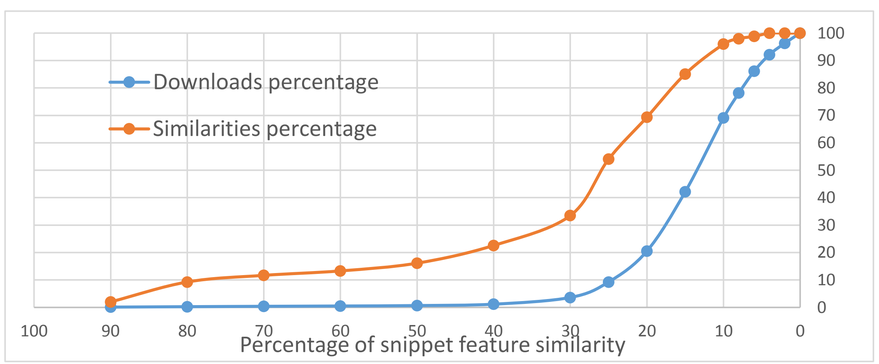


Figure 6: *Downloads and similarities performance.*

4. Text Alignment

The system uses the same basic principles as in [4]:

- ***common features*** between source and suspicious documents
  - word 5-grams
  - stop-word 8-grams [3]
- ***valid intervals*** of characters covered by common features "densely enough"
- ***postprocessing***—remove overlapping detections, join neighbouring detections

4.1 Alternative Features

- ***contextual n-grams*** [5]
  - ***The quick*** brown ***fox jumped*** over the lazy dogs.
  - The ***quick brown*** fox ***jumped over*** the lazy dogs.
- plain word 4-grams
  - ***The quick brown fox*** jumped over the lazy dogs.
  - The ***quick brown fox jumped*** over the lazy dogs.

feature	recall	precision	granularity	plagdet
plain 5-grams	0.6306	0.8484	1.0000	***0.7235***
contextual 4-grams	0.6721	***0.8282***	1.0000	***0.7421***
plain 4-grams	***0.7556***	0.7340	1.0000	***0.7447***

Table 1: *Comparison of contextual 4-grams and plain word 4-grams*

4.2 Global Postprocessing

- Similar to PAN 2010 [2]
- Overlapping detections removal
- ***Result:*** improvement, but not as significant as in 2010

5. Conclusion

5.1 Candidate retrieval

- Second best ratio of recall to the number of queries
- Missing support for phrasal search in ChatNoir is a big stumbling block

5.2 Text alignment

- Significant improvement against PAN 2013
- Word 4-grams are better than contextual 4-grams
- We need a better ranking system than plagdet!

References

[1] Sven Meyer Zu Eissen and Benno Stein. Intrinsic plagiarism detection. In *Proceedings of the European Conference on Information Retrieval (ECIR-06)*, 2006.

[2] J. Kasprzak and M. Brandejs. Improving the reliability of the plagiarism detection system. In *Notebook Papers of CLEF 2010 LABs and Workshops*. Citeseer, 2010.

[3] E. Stamatatos. Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, 2011.

[4] Šimon Suchomel, Jan Kasprzak, and Michal Brandejs. Three way search engine queries with multi-feature document comparison for plagiarism detection. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, pages 1–8, 2012.

[5] Diego A Rodríguez Torrejón and José Manuel Martín Ramos. Detailed comparison module in coremo 1.9 plagiarism detector. In *CLEF (Online Working Notes/Labs/Workshop)*, pages 1–8, 2012.

Contact information:
Šimon Suchomel suchomel@fi.muni.cz
Jan Kasprzak kas@fi.muni.cz
http://www.fi.muni.cz/~kas/pan13/