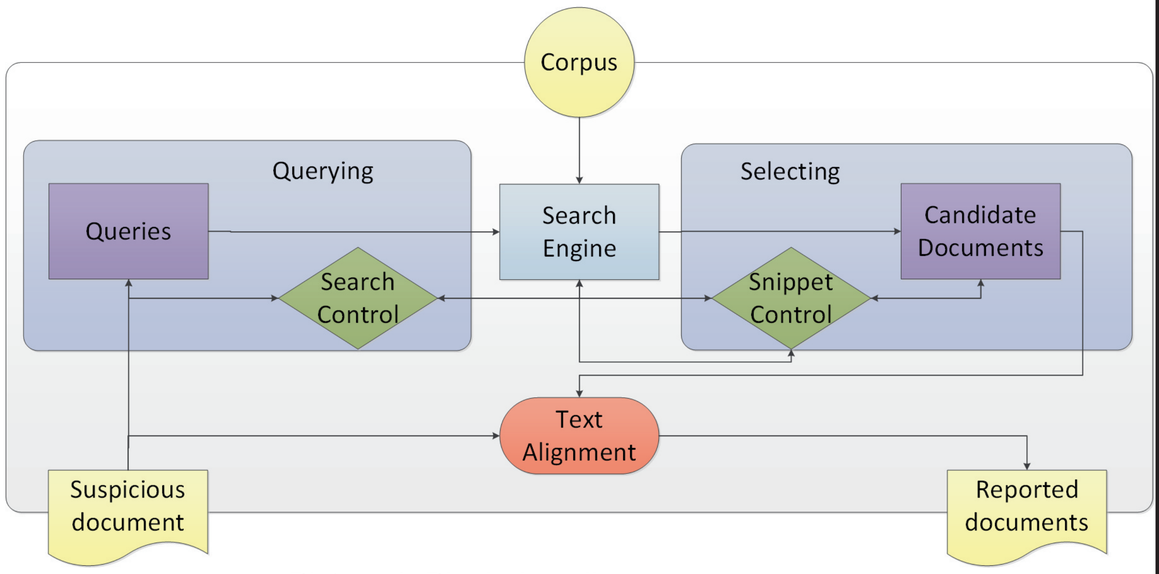
Heterogeneous Queries for Synoptic and Phrasal Search

Šimon Suchomel, and Michal Brandejs

Faculty of Informatics, Masaryk University, Brno, Czech Republic

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 PAN2014 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.

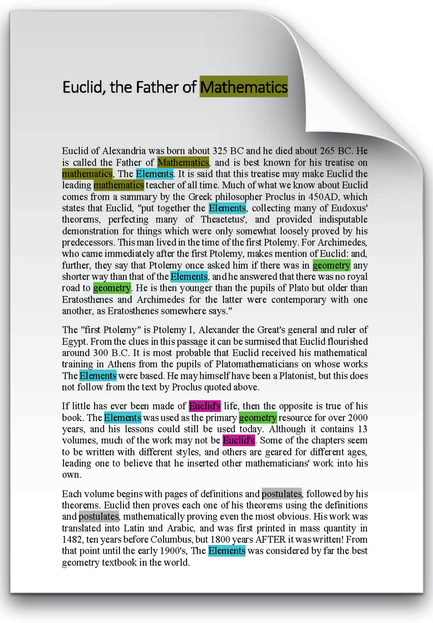


Picture 1: Plagiarism discovery process.

Building of the Queries

Several types of queries were prepared.This year we combined keywords-based queries and paragraph-based queries together. Some of the prepared queries could be discarded from the execution, no query refinement was applied according to the results, but some of the top-scored keywords could appear in the different combination in more than one query.

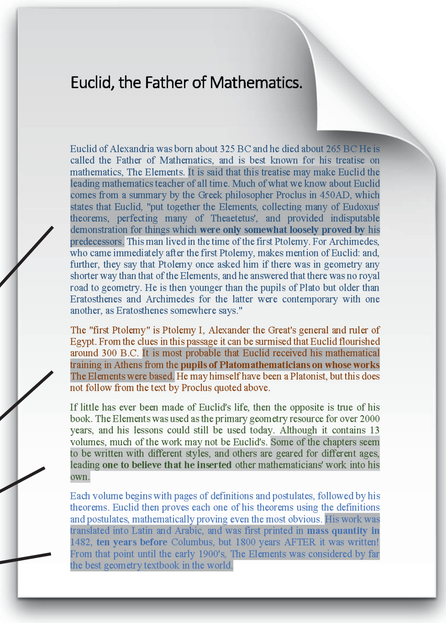
Keywords



Query	Source	Position interval	
postulate euclid geometry elements axiom al	Pilot		
elements considered geometry	phrase:Collocation		
geometry textbook world	phrase:Collocation		
parallel postulate line angle proof euclidean	Collocation		
geometry textbook world mathematician work obtuse	Collocation		
equal point volumes assumption publish haytham	KW		
book girolamo eratosthenes year theory factorization	KW		
equidistant proclus consider praise number father	KW		
proceed volume obvious lambert uniqueness definition	KW		
were only somewhat loosely proved by	phrase:Paragraph	0	1131
pupils of platomathematicians on whose works	phrase:Paragraph	1134	1569
one to believe that he inserted	phrase:Paragraph	1571	2038
mass quantity in ten years before	phrase:Paragraph	2040	2535

- Lemmatization; stop-words removal;
- TF-IDF scoring for keywords;
- from top 3 KW, there were collocations extracted;
- longest collocations form phrasal queries;
- 6 tokens long queries.

Paragraphs



- Longest sentence from each paragraph;
- no stop-words removal;
- 6 tokens long phrasal queries;
- postitional queries.

Chat Noir

- pilot query
- non-phrasal queries

Indri

- pilot query
- phrasal queries

Downloading and post-processing

Snippet

- for each document query
- 2-tuples measurement
- 20% concordance for download

Post-processing

The system uses the same basic principles as in PAN 2013.

- **Common features** between source and suspicious documents;
  - word 5-grams;
  - stop-word 8-grams.
- **Alternative features**;
  - contextual n-grams;
  - plain word 4-grams.
- Overplapping detection removal.

In the post-processing phase a similarity between the suspicious and the source document was calulated. If any similarity was detected, the suspicious document were reported as a potential source of plagiarism.

Conclusion

This poster shows the key aspects and changes from our erstwhile systems for candidate document retrieval used at PAN 14 lab on uncovering plagiarism. The architecture stems from PAN 12 and PAN 13 labs and the real-world anti-plagiarism system which is in use at Masaryk University. The results of the PAN show that this approach is one of the best for a real-life adoption, since it achieved a decent recall with just a fraction of used queries. Such approach is applicable for detection of suspicious texts, which may contain plagiarism, that can then be selected for further investigation.

Contact information:
Šimon Suchomel, suchomel@fi.muni.cz
http://www.fi.muni.cz/~xsuchom1/pan14/

CLEF 2014, 15-18 September 2014, Sheffield, UK