Heterogeneous Queries for Synoptic and Phrasal Search

Šimon Suchomel, and Michal Brandejs

Faculty of Informatics, Masaryk University, Brno, Czech Republic

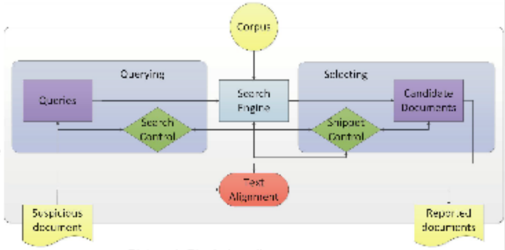
Introduction

Main presumptions

- Queries are the most expensive.
- Downloads are cheap.
- Single-themed documents.
- What is the minimum copied unit?
  - words, sentences, paragraphs

Methodology

- Retrieve theme-similar documents.
- Retrieve documents with the same text according to chosen chunking method.
- Proceed iteratively over queries.



Picture 1: Plagiarism discovery process.

Building of the Queries

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

Paragraphs

- 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.

- 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

Snippet

More than 1 per result.

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

Downloading and post-processing

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.

Tuned for text alignment task, which is surprisingly not ideal for source retrieval task.

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

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

Conclusion

- There is no optimal chunking method – without computation of text characteristics.
- The keywords-based queries are possibly the most profitable.

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

Heterogeneous Queries for Synoptic and Phrasal Search

Šimon Suchomel, and Michal Brandejs

Faculty of Informatics, Masaryk University, Brno, Czech Republic

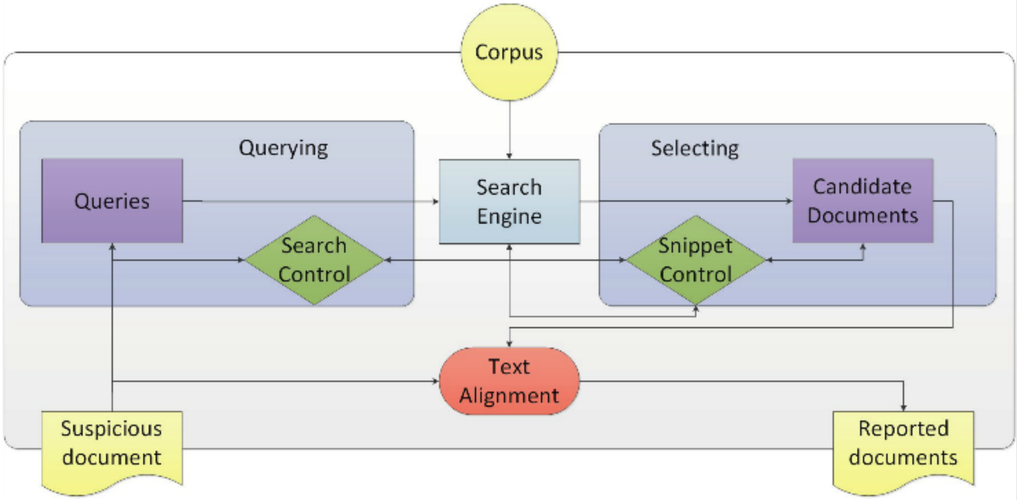
Introduction

Main presumptions

- Queries are the most expensive.
- Downloads are cheap.
- Single-themed documents.
- What is the minimum copied unit?
  - words, sentences, paragraphs

Methodology

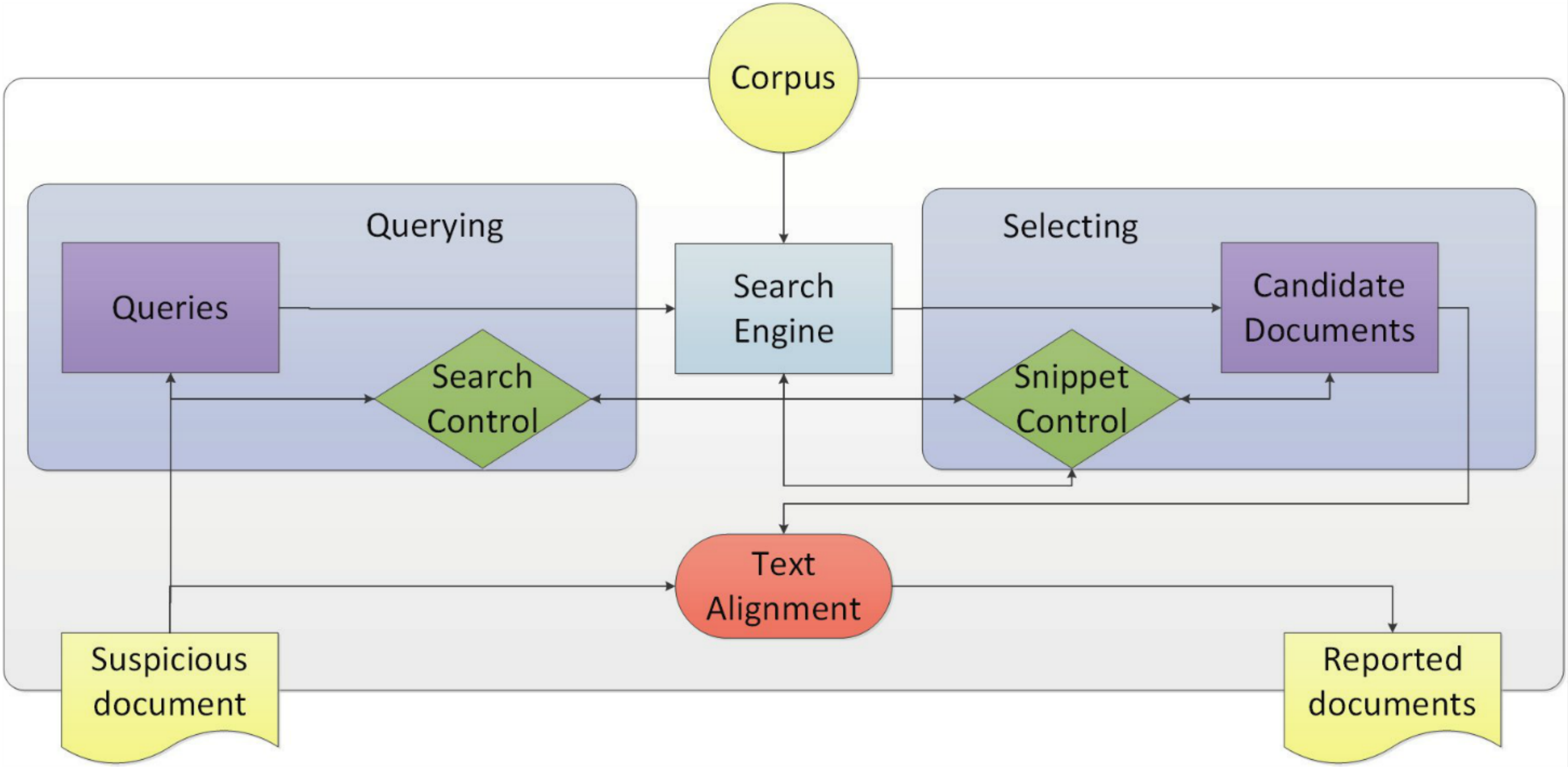
- Retrieve theme-similar documents.
- Retrieve documents with the same text according to chosen chunking method.
- Proceed iteratively over queries.



Picture 1: Plagiarism discovery process.

ryk University, Brno, Czech Republic

RSITAS MASARY



Picture 1: Plagiarism discovery process.

ne
g

Prezi

the Queries

Main presumptions

- Queries are the most expensive.
- Downloads are cheap.
- Single-themed documents.
- What is the minimum copied unit?
    - words, sentences, paragraphs

e.

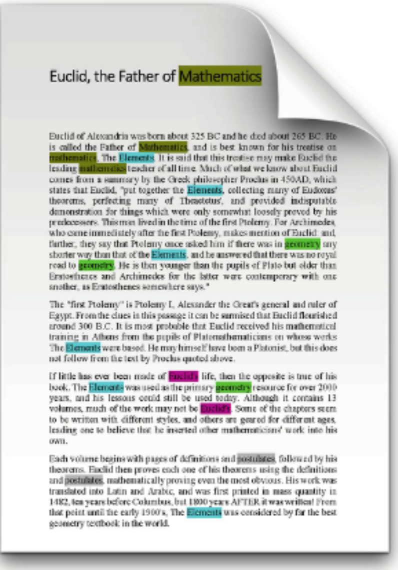
nit?
phs

Methodology

- Retrieve theme-similar documents.
- Retrieve documents with the same text according to chosen chunking method.
- Proceed iteratively over queries.

Building of the Queries

Keywords



Euclid, the Father of Mathematics

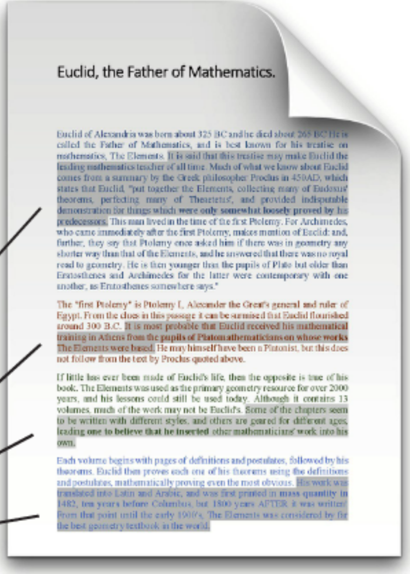
Euclid of Alexandria was born about 325 BC and he died about 265 BC. He is called the Father of Mathematics, and is best known for his treatise on mathematics, The Elements. It is said that this treatise may make Euclid the leading mathematics teacher of all time. Much of what we know about Euclid comes from a summary by the Greek philosopher Proclus in 450AD, which states that Euclid, "put together the Elements, collecting many of Eudoxus' theorems, perfecting many of Theaetetus', and provided indisputable demonstration for things which were only somewhat loosely proved by his predecessors. This man lived in the time of the first Ptolemy. For Archimedes, who came immediately after the first Ptolemy, makes mention of Euclid: and, further, they say that Ptolemy once asked him if there was in geometry any shorter way than that of the Elements, and he answered that there was no royal road to geometry. He is then younger than the pupils of Plato but older than Eratosthenes and Archimedes for the latter were contemporary with one another, as Eratosthenes somewhere says."

The "first Ptolemy" is Ptolemy I, Alexander the Great's general and ruler of Egypt. From the clues in this passage it can be surmised that Euclid flourished around 300 B.C. It is most probable that Euclid received his mathematical training in Athens from the pupils of Platomathematicians on whose works The Elements were based. He may himself have been a Platonist, but this does not follow from the text by Proclus quoted above.

If little has ever been made of Euclid's life, then the opposite is true of his book. The Elements was used as the primary geometry resource for over 2000 years, and his lessons could still be used today. Although it contains 13 volumes, much of the work may not be Euclid's. Some of the chapters seem to be written with different styles, and others are geared for different ages, leading one to believe that he inserted other mathematicians' work into his own.

Each volume begins with pages of definitions and postulates, followed by his theorems. Euclid then proves each one of his theorems using the definitions and postulates, mathematically proving even the most obvious. His work was translated into Latin and Arabic, and was first printed in mass quantity in 1482, ten years before Columbus, but 1800 years AFTER it was written! From that point until the early 1900's, The Elements was considered by far the best geometry textbook in the world.

Paragraphs



Euclid, the Father of Mathematics.

Euclid of Alexandria was born about 325 BC and he died about 265 BC. He is called the Father of Mathematics, and is best known for his treatise on mathematics, The Elements. It is said that this treatise may make Euclid the leading mathematics teacher of all time. Much of what we know about Euclid comes from a summary by the Greek philosopher Proclus in 450AD, which states that Euclid, "put together the Elements, collecting many of Eudoxus' theorems, perfecting many of Theaetetus', and provided indisputable demonstration for things which were only somewhat loosely proved by his predecessors. This man lived in the time of the first Ptolemy. For Archimedes, who came immediately after the first Ptolemy, makes mention of Euclid: and, further, they say that Ptolemy once asked him if there was in geometry any shorter way than that of the Elements, and he answered that there was no royal road to geometry. He is then younger than the pupils of Plato but older than Eratosthenes and Archimedes for the latter were contemporary with one another, as Eratosthenes somewhere says."

The "first Ptolemy" is Ptolemy I, Alexander the Great's general and ruler of Egypt. From the clues in this passage it can be surmised that Euclid flourished around 300 B.C. It is most probable that Euclid received his mathematical training in Athens from the pupils of Platomathematicians on whose works The Elements were based. He may himself have been a Platonist, but this does not follow from the text by Proclus quoted above.

If little has ever been made of Euclid's life, then the opposite is true of his book. The Elements was used as the primary geometry resource for over 2000 years, and his lessons could still be used today. Although it contains 13 volumes, much of the work may not be Euclid's. Some of the chapters seem to be written with different styles, and others are geared for different ages, leading one to believe that he inserted other mathematicians' work into his own.

Each volume begins with pages of definitions and postulates, followed by his theorems. Euclid then proves each one of his theorems using the definitions and postulates, mathematically proving even the most obvious. His work was translated into Latin and Arabic, and was first printed in mass quantity in 1482, ten years before Columbus, but 1800 years AFTER it was written! From that point until the early 1900's, The Elements was considered by far the best geometry textbook in the world.

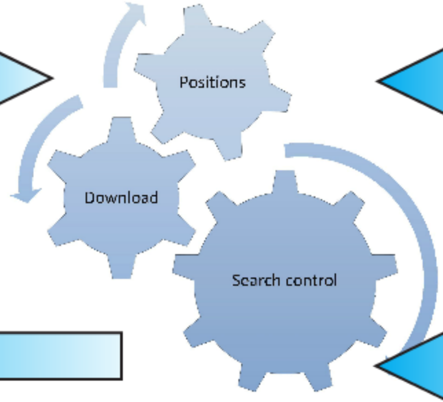
Query	Source	Position interval	
postulate euclid geometry elements axiom al	Pilot		
elements considered geometry	phrase:Collocation		
geometry textbook world	phrase:Collocation		
parallel postulate line angle proof euclidean	Collocation		
geometry textbook world mathematician work obtuse	Collocation		
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.

- 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

Snippet

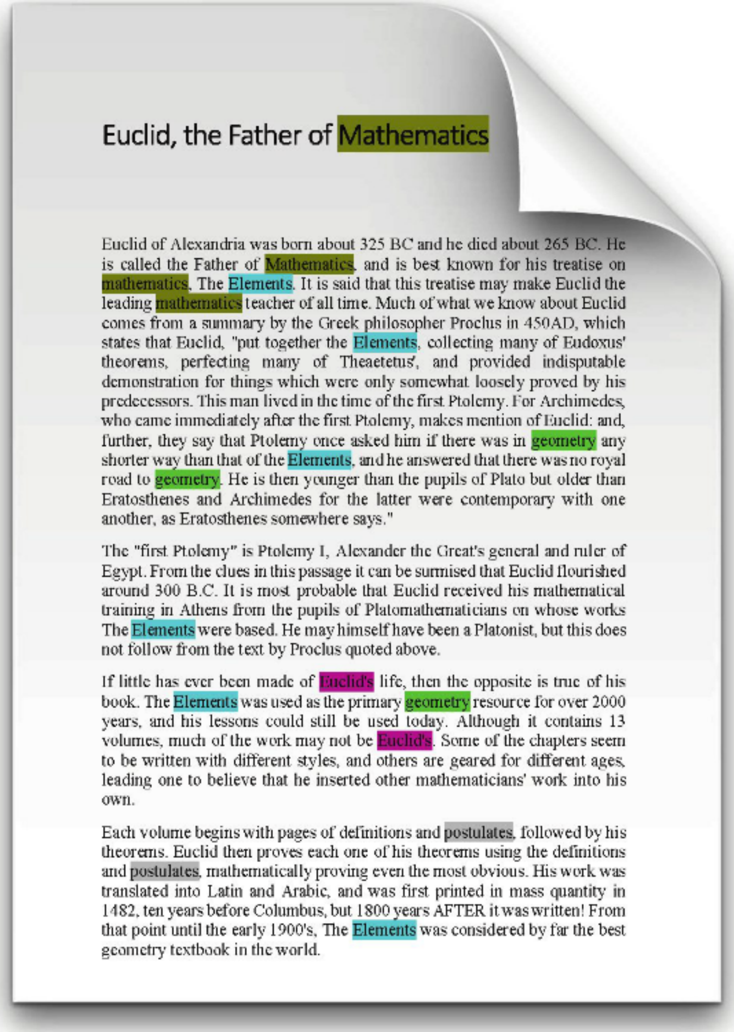
More than 1 per result.

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

Downloading and post-processing

Building of the

Keywords



Euclid, the Father of Mathematics

Euclid of Alexandria was born about 325 BC and he died about 265 BC. He is called the Father of Mathematics, and is best known for his treatise on mathematics, The Elements. It is said that this treatise may make Euclid the leading mathematics teacher of all time. Much of what we know about Euclid comes from a summary by the Greek philosopher Proclus in 450AD, which states that Euclid, "put together the Elements, collecting many of Eudoxus' theorems, perfecting many of Theaetetus', and provided indisputable demonstration for things which were only somewhat loosely proved by his predecessors. This man lived in the time of the first Ptolemy. For Archimedes, who came immediately after the first Ptolemy, makes mention of Euclid: and, further, they say that Ptolemy once asked him if there was in geometry any shorter way than that of the Elements, and he answered that there was no royal road to geometry. He is then younger than the pupils of Plato but older than Eratosthenes and Archimedes for the latter were contemporary with one another, as Eratosthenes somewhere says."

The "first Ptolemy" is Ptolemy I, Alexander the Great's general and ruler of Egypt. From the clues in this passage it can be surmised that Euclid flourished around 300 B.C. It is most probable that Euclid received his mathematical training in Athens from the pupils of Platomathematicians on whose works The Elements were based. He may himself have been a Platonist, but this does not follow from the text by Proclus quoted above.

If little has ever been made of Euclid's life, then the opposite is true of his book. The Elements was used as the primary geometry resource for over 2000 years, and his lessons could still be used today. Although it contains 13 volumes, much of the work may not be Euclid's. Some of the chapters seem to be written with different styles, and others are geared for different ages, leading one to believe that he inserted other mathematicians' work into his own.

Each volume begins with pages of definitions and postulates, followed by his theorems. Euclid then proves each one of his theorems using the definitions and postulates, mathematically proving even the most obvious. His work was translated into Latin and Arabic, and was first printed in mass quantity in 1482, ten years before Columbus, but 1800 years AFTER it was written! From that point until the early 1900's, The Elements was considered by far the best geometry textbook in the world.

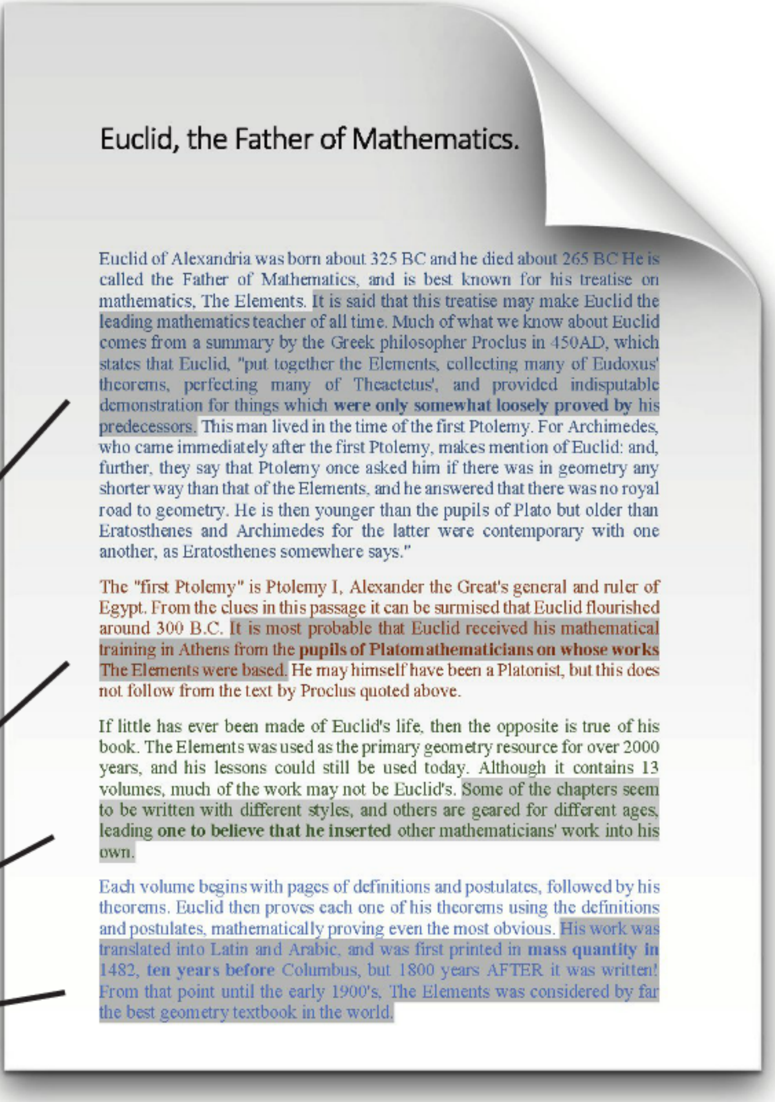
Query

postulate euclid geometry elements axiom al
elements considered geometry
geometry textbook world
parallel postulate line angle proof euclidean
geometry textbook world mathematician work obtus
equal point volumes assumption publish haytham
book girolamo eratosthenes year theory factorizatio
equidistant proclus consider praise number father
proceed volume obvious lambert uniqueness definit
were only somewhat loosely proved by
pupils of platomathematicians on whose works
one to believe that he inserted
mass quantity in ten years before

- 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

Source	Position interval	
e:Collocation		
e:Collocation		
cation		
cation		
e:Paragraph	0	1131
e:Paragraph	1134	1569
e:Paragraph	1571	2038
e:Paragraph	2040	2535



Euclid, the Father of Mathematics.

Euclid of Alexandria was born about 325 BC and he died about 265 BC He is called the Father of Mathematics, and is best known for his treatise on mathematics, The Elements. It is said that this treatise may make Euclid the leading mathematics teacher of all time. Much of what we know about Euclid comes from a summary by the Greek philosopher Proclus in 450AD, which states that Euclid, "put together the Elements, collecting many of Eudoxus' theorems, perfecting many of Theaetetus', and provided indisputable demonstration for things which **were only somewhat loosely proved by** his predecessors. This man lived in the time of the first Ptolemy. For Archimedes, who came immediately after the first Ptolemy, makes mention of Euclid: and, further, they say that Ptolemy once asked him if there was in geometry any shorter way than that of the Elements, and he answered that there was no royal road to geometry. He is then younger than the pupils of Plato but older than Eratosthenes and Archimedes for the latter were contemporary with one another, as Eratosthenes somewhere says."

The "first Ptolemy" is Ptolemy I, Alexander the Great's general and ruler of Egypt. From the clues in this passage it can be surmised that Euclid flourished around 300 B.C. It is most probable that Euclid received his mathematical training in Athens from the **pupils of Platomathematicians on whose works** The Elements were based. He may himself have been a Platonist, but this does not follow from the text by Proclus quoted above.

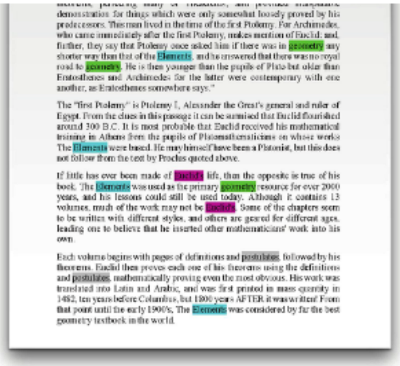
If little has ever been made of Euclid's life, then the opposite is true of his book. The Elements was used as the primary geometry resource for over 2000 years, and his lessons could still be used today. Although it contains 13 volumes, much of the work may not be Euclid's. Some of the chapters seem to be written with different styles, and others are geared for different ages, leading **one to believe that he inserted** other mathematicians' work into his own.

Each volume begins with pages of definitions and postulates, followed by his theorems. Euclid then proves each one of his theorems using the definitions and postulates, mathematically proving even the most obvious. His work was translated into Latin and Arabic, and was first printed in **mass quantity in** 1482, **ten years before** Columbus, but 1800 years AFTER it was written! From that point until the early 1900's, The Elements was considered by far the best geometry textbook in the world.

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

Query	Source	Position interva	
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

ed;



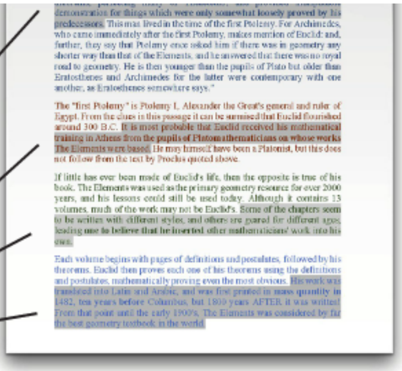
demonstration for things which were only somewhat loosely proved by his predecessors. This man lived in the time of the first Ptolemy. For Archimedes, who came immediately after the first Ptolemy, makes mention of Euclid: and, further, they say that Ptolemy once asked him if there was in geometry any shorter way than that of the Elements, and he answered that there was no royal road to geometry. He is then younger than the pupils of Plato but older than Eratosthenes and Archimedes for the latter were contemporary with one another, as Eratosthenes somewhere says."

The "first Ptolemy" is Ptolemy I, Alexander the Great's general and ruler of Egypt. From the clues in this passage it can be surmised that Euclid flourished around 300 B.C. It is most probable that Euclid received his mathematical training in Athens from the pupils of Platomathematicians on whose works The Elements were based. He may himself have been a Platonist, but this does not follow from the text by Proclus quoted above.

If little has ever been made of Euclid's life, then the opposite is true of his book. The Elements was used as the primary geometry resource for over 2000 years, and his lessons could still be used today. Although it contains 13 volumes, much of the work may not be Euclid's. Some of the chapters seem to be written with different styles, and others are geared for different ages, leading one to believe that he inserted other mathematicians' work into his own.

Each volume begins with pages of definitions and postulates, followed by his theorems. Euclid then proves each one of his theorems using the definitions and postulates, mathematically proving even the most obvious. His work was translated into Latin and Arabic, and was first printed in mass quantity in 1482, ten years before Columbus, but 1800 years AFTER it was written! From that point until the early 1900's, The Elements was considered by far the best geometry textbook in the world.

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



demonstration for things which were only somewhat loosely proved by his predecessors. This man lived in the time of the first Ptolemy. For Archimedes, who came immediately after the first Ptolemy, makes mention of Euclid: and, further, they say that Ptolemy once asked him if there was in geometry any shorter way than that of the Elements, and he answered that there was no royal road to geometry. He is then younger than the pupils of Plato but older than Eratosthenes and Archimedes for the latter were contemporary with one another, as Eratosthenes somewhere says."

The "first Ptolemy" is Ptolemy I, Alexander the Great's general and ruler of Egypt. From the clues in this passage it can be surmised that Euclid flourished around 300 B.C. It is most probable that Euclid received his mathematical training in Athens from the pupils of Platomathematicians on whose works The Elements were based. He may himself have been a Platonist, but this does not follow from the text by Proclus quoted above.

If little has ever been made of Euclid's life, then the opposite is true of his book. The Elements was used as the primary geometry resource for over 2000 years, and his lessons could still be used today. Although it contains 13 volumes, much of the work may not be Euclid's. Some of the chapters seem to be written with different styles, and others are geared for different ages, leading one to believe that he inserted other mathematicians' work into his own.

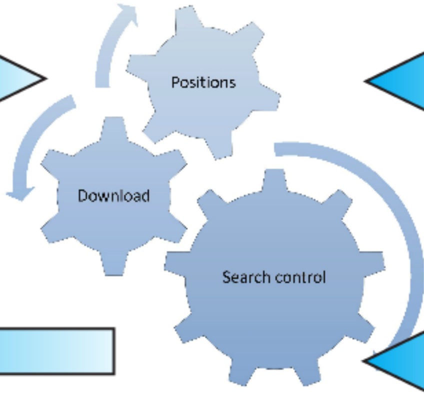
Each volume begins with pages of definitions and postulates, followed by his theorems. Euclid then proves each one of his theorems using the definitions and postulates, mathematically proving even the most obvious. His work was translated into Latin and Arabic, and was first printed in mass quantity in 1482, ten years before Columbus, but 1800 years AFTER it was written! From that point until the early 1900's, The Elements was considered by far the best geometry textbook in the world.

- 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.

- 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

Snippet

More than 1 per result.

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

Downloading and post-processing

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.

Tuned for text alignment task, which is surprisingly not ideal for source retrieval task.

In the post-processing phase a similarity between the suspicious and the source document was calculated. If any similarity was detected, the

Conclusion

- There is no optimal chunking method – without computation of text characteristics.
- The keywords-based queries are possibly the most profitable.

Indri

- pilot query
- phrasal queries

Snippet

More than 1 per result.

- 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.

Tuned for text alignment task, which is surprisingly not ideal for source retrieval task.

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

- 20% concordance for download

Conclusion

- There is no optimal chunking method – without computation of text characteristics.
- The keywords-based queries are possibly the most profitable.

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

Heterogeneous Queries for Synoptic and Phrasal Search

Šimon Suchomel, and Michal Brandejs

Faculty of Informatics, Masaryk University, Brno, Czech Republic

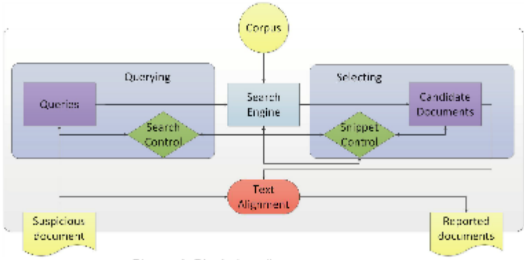
Introduction

Main presumptions

- Queries are the most expensive.
- Downloads are cheap.
- Single-themed documents.
- What is the minimum copied unit?
  - words, sentences, paragraphs

Methodology

- Retrieve theme-similar documents.
- Retrieve documents with the same text according to chosen chunking method.
- Proceed iteratively over queries.



Picture 1: Plagiarism discovery process.

Building of the Queries

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

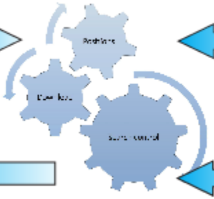
Paragraphs

- 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.

- 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

Snippet

More than 1 per result.

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

Downloading and post-processing

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.

Tuned for text alignment task, which is surprisingly not ideal for source retrieval task.

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

Conclusion

- There is no optimal chunking method – without computation of text characteristics.
- The keywords-based queries are possibly the most profitable.

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