Heterogeneous Queries for Synoptic and Phrasal Search

Šimon Suchomel, and Michal Brandejs
Faculty of Informatics, Mendel 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.

Keywords
- Case studies.
- Enriched query.
- Synoptic queries.
- Phrasal queries.
- Query classification.

Building of the Queries

Query
- Postulates word-grammar elements based on a skeleton preselected grammar.
- Provided with the single proof of a criterion.
- Provides a list of potential sentences.

Source
- Time-dependent.
- Distance.
- Positional.

Post-processing

The system uses the same basic principles as in PAN 2010.
- Common features between source and suspicious documents;
- word bigrams;
- adjacent bigrams;
- "alternative features";
- plain word-grams;
- Overlapping word-grams.

In the postprocessing phase a similarity between the suspicious and the source document was calculated. If any similarity was detected, the suspicious document was 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@fjfi.stuba.sk
http://www.fjfi.stuba.sk/mucbi/
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.
Picture 1: Plagiarism discovery process.
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.
Building of the Queries

Keywords

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

Paragaphs

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

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.

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
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 geometry called The Elements. It is said that his treatise may make Euclid the leading mathematician 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 theorems, collecting many of the results of his predecessors, giving demonstrations of the truth where they were possible, filling in the gaps, and fashioning the whole into a deductive system, a Euclidean element of geometrical knowledge." He also described several geometric results which could be demonstrable, and provided incontestable demonstrations for things which were only somewhat loosely proved by his predecessors. This man lived in the time of the first Ptolemy, three 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 contemporaries with one another, as Eratosthenes somewhere says.4

The "First Elements" is Ptolemy's, Alexander theCleatic's general and ruler of Egypt. From the clues in this passage it can be assumed that Euclid taught around 300 B.C. It is most probable that Euclid received his mathematical training in Athens from the pupils of Plato, mathematicians 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 teaching resource for over 2000 years, and its lessons could still be used today, although it contains 13 volumes, much of the work may not be Euclid. Some of the chapters seem to be written with different styles, and other 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 1300 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.

Keywords

- Euclid
- Geometry
- Elements
- Mathematics

Query

- Postulate
- Euclid
- Geometry
- Elements
- Axioms
- Algebra

- Elements
- Considered
- Geometry
- Textbook
- World

- Parallel
- Postulate
- Line
- Angle
- Proof

- Euclidean
- Geometry
- Textbook
- World
- Mathematician
- Work
- Obtuse

- Equal
- Point
- Volumes
- Assumption
- Publish
- Haytham

- Girolamo Eratosthenes
- Year
- Theory
- Factorization

- Equidistant
- Proclus
- Consider
- Praise
- Number
- Father

- Proceed
- Volume
- Obvious
- Lambert
- Uniqueness

- Definite
- If
- Only
- Somewhat
- Loosely
- Proven
- By

- Pupils
- Platonists
- Mathematicians
- Whose

- Works
- One
- To
- Believe
- That
- He
- Inserted

- Mass
- Quantity
- Ten
- Years

- Lemmatization
- Stop-words
- Removal

- TF-IDF
- Scoring
- Keywords

- From
- Top
- KW,
- There
- Collocations
- Extracted

- Longest
- Collocations
- Form
- Phrasal
- Queries

- Tokens
- Long
- Queries
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 mathematician of all time. Much of what we know about Euclid comes from a summary by the Greek philosopher Proclus in 460 AD, which states that Euclid, "and together with the Elements, collected many of Ptolemy's theorems, perfected many of Thales', and provided indubitable 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 Pythagoras and 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 clue in this passage it can be surmised that Euclid resided around 300 BC. It is most probable that Euclid received his mathematical training in Athens from the pupils of Plato, mathematicians 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 work. 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 1600 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;
- positional queries.
postulate euclid geometry elements axiom al
elements considered geometry
gometry textbook world
parallel postulate line angle proof euclidean
gometry textbook world mathematician work obtuse
equal point volumes assumption publish haytham
book girolamo eratosthenes year theory factorization
equidistant proclus consider praise number father
proceed volume obvious lambert uniqueness definition
were only somewhat loosely proved by
pupils of platomathematicians on whose works
one to believe that he inserted
mass quantity in ten years before
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.
- Overlapping 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/
Heterogeneous Queries for Synoptic and Phrasal Search

Simon 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 coded 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

- Lorraineization: stop-words removal.
- TF-IDF scoring for keywords.
- From top 3,000, there were collections extracted.
- Longest collations form physical queries.
- It takes long queries.

Chat Noir
- pithy query
- non-physical query

Downloading and post-processing

Paragaphs

- Longest sentence from each paragraph.
- No stop-words removal.
- 6 tokens long phrasal queries.
- Positional queries.

Snippets
- More than 1 per result.
- For each document query.
- 2-tuples measurement.
- 27% concordance for download.

Post-processing

The system uses the same basic principles as in PAM 2013.
- Common features between source and suspicious documents.
- Word 5-grams.
- Stop-word 8-grams.
- Alternative features: contextual 4-grams: plain word 4-grams.
- Overlapping deletion removal.

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