Three Way Search Engine with Queries Multi-feature Document Comparison

Šimon Suchomel and Jan Kasprzak with Michal Brandejs

Faculty of Informatics Masaryk University
Three types of queries

Keywords
- KW extraction based on tf-idf
- KW enriched with collocations
- non-positionable
- unconditionally executed
- 4 KW queries per doc on average
- covering the whole document
  e.g., mobile phone signal prepaid cards

Intrinsic plagiarism
- positionable
- conditionally executed
- covering the suspicious passage
- based on Average Word Frequency Class
- its change indicates suspicious passage
  e.g., in precious ladies shoes placing focus

Headers
- naive headers detection
- positionable
- conditionally executed
- covering the part introduces by the header
  e.g., Differences Between Cell Phones

Suspicious document

<table>
<thead>
<tr>
<th>Query type</th>
<th>Extracted/Optimal Similarities percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>KW</td>
<td>4.16</td>
</tr>
<tr>
<td>Headers</td>
<td>4.45</td>
</tr>
<tr>
<td>Headers</td>
<td>4.04</td>
</tr>
</tbody>
</table>
- KW extraction based on tf-idf
- KW enriched with collocations
- non positionable
- unconditionally executed
- 4 KW queries per doc on average
- covering the whole document

e.g.: mobile phone signal prepaid carriers
Suspicious document

All About Cell Phone

More and more people have been using cell phone nowadays. In fact, some people even consider it as a necessity and we can’t blame them. Cell phone is very useful and helpful for many after all.

What is Cell Phone?
Unless you’re living under a rock, you do know what a cell phone is. But for the uninitiated, here’s what it means. A mobile phone (also known as a ...

Differences Between Cell Phones

The cell phone itself isn’t an ongoing expense. You will need to shell out money for this only once unless you want an upgrade. Your own personal taste, style and need should determine what cell phone you want (not the sales assistant behind the desk).

... The use of “hands-free” was not recommended by the British Consumers’ Association in a statement in November 2000 as they believed that exposure was increased. However, measurements for the (then) UK Department of Trade and Industry and others for the French | Agence française de santé environnementale | showed substantial reductions. In 2005 Professor Lawrie Challis and others said clipping a ferrite bead onto hands-free kits stops the radio waves travelling up the wire and into the head.

Overall Health Risks

Many scientific studies have investigated possible health effects of mobile phone radiations. These studies are occasionally reviewed by some scientific committees to assess overall risks. The most recent assessment was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR). It concludes from the available research that no significant health effect has been demonstrated from mobile phone radiation at normal exposure levels:
• unconditionally executed
• 4 KW queries per doc on average
• covering the whole document

e.g.: mobile phone signal prepaid carriers
Suspicious document

All About Cell Phone

More and more people have been using cell phone nowadays. In fact, some people even consider it as a necessity and we can’t blame them. Cell phone is very useful and helpful for many after all.

What is Cell Phone?
Unless you’re living under a rock, you do know what a cell phone is. But for the uninitiated, here’s what it means. A mobile phone (also known as a ...

**Differences Between Cell Phones**

The cell phone itself isn’t an ongoing expense. You will need to shell out money for this only once unless you want an upgrade. Your own personal taste, style and need should determine what cell phone you want (not the sales assistant behind the desk).

... The use of “hands-free” was not recommended by the British Consumers’ Association in a statement in November 2000 as they believed that exposure was increased. However, measurements for the (then) UK Department of Trade and Industry and others for the French | Agence française de sécurité sanitaire environnemental showed substantial reductions. In 2005 Professor Lawrie Challis and others said clipping a ferrite bead onto hands-free kits stops the radio waves travelling up the wire and into the head.

**Overall Health Risks**

Many scientific studies have investigated possible health effects of mobile phone radiations. These studies are occasionally reviewed by some scientific committees to assess overall risks. The most recent assessment was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR). It concludes from the available research that no significant health effect has been demonstrated from mobile phone radiation at normal exposure levels:
Three types of queries

Keywords
- KW extraction based on TF-IDF
- KW enriched with collocations
- non-positionable
- unconditionally executed
- 4 KW queries per doc on average
- covering the whole document
  e.g. mobile phone signal prepaid cards

Intrinsic plagiarism
- positionable
- conditionally executed
- covering the suspicious passage
- based on Average Word Frequency Class
- it's change indicates suspicious passage
  e.g. in previous debate students stealing finals

Headers
- naive headers detection
- positionable
- conditionally executed
- covering the part introduces by the header
  e.g. Differences Between Cell Phones

Suspicious document

<table>
<thead>
<tr>
<th>Query type</th>
<th>Extracted/Obtained Similarities</th>
<th>similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>KW</td>
<td>416</td>
<td>72.5%</td>
</tr>
<tr>
<td>Headers</td>
<td>1034</td>
<td>24.1%</td>
</tr>
<tr>
<td>NA</td>
<td>405</td>
<td>0.2%</td>
</tr>
</tbody>
</table>
- positionable
- conditionally executed
- covering the suspicious passage
- based on Average Word Frequency Class
- it's change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite
Differences Between Cell Phones

The cell phone itself isn't an ongoing expense. You will need to shell out money for this only once unless you want an upgrade. Your own personal taste, style and need should determine what cell phone you want (not the sales assistant behind the desk).

The use of "hands-free" was not recommended by the British Consumers' Association in a statement in November 2000 as they believed that exposure was increased. However, measurements for the (then) UK Department of Trade and Industry and others for the French Agence nationale de sécurité sanitaire environmental showed substantial reductions. In 2005 Professor Lawrie Challis and others said clipping a ferrite bead onto hands-free kits stops the radio waves travelling up the wire and into the head.

Overall Health Risks
Many scientific studies have investigated possible health effects of mobile phone radiations. These studies are occasionally reviewed by some scientific committees to assess overall risks. The most recent assessment was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR). It concludes from the available data that no significant health effects have been substantiated.
Based on Average Word Frequency Class

- it's change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite
Three types of queries

**Keywords**
- KW extraction based on tf-idf
- KW enrichment with collocations
- non postionable
- unconditionally executed
- 4 KW queries per doc on average
- covering the whole document
- e.g., mobile phone signal, prepaid cards

**Intrinsic plagiarism**
- positional
- conditionally executed
- covering the suspicious passage
- based on Average Word Frequency Class
- it's change indicates suspicious passage
- e.g., in protractor labor: should doping term

**Headers**
- naive headers detection
- postionable
- conditionally executed
- covering the part introduces by the header
- e.g., Differences Between Cell Phones

---

**Table:**

<table>
<thead>
<tr>
<th>Query type</th>
<th>Extracted</th>
<th>Desired</th>
<th>Similarities percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>KW</td>
<td>416</td>
<td>N/A</td>
<td>72.5%</td>
</tr>
<tr>
<td>Headers</td>
<td>1034</td>
<td>405</td>
<td>82.6%</td>
</tr>
</tbody>
</table>
- naive headers detection
- positionable
- conditionally executed
- covering the part introduces by the header

e.g.: Differences Between Cell Phones
people even consider it as a necessity and we can't see it in our lives. It is very useful and helpful for many after all.

What is Cell Phone?
Unless you're living under a rock, you do know what a cell phone is for the uninitiated, here's what it means. A mobile...

Differences Between Cell Phones

The cell phone itself isn't an ongoing expense. You pay money for this only once unless you want an upgrade (taste, style and need should determine what cell phone you sales assistant behind the desk).

The use of "hands-free" was not recommended by the...
covering the part introduced by the header

e.g.: Differences Between Cell Phones
<table>
<thead>
<tr>
<th>Query type</th>
<th>Extracted</th>
<th>Omitted</th>
<th>Similarities portion</th>
</tr>
</thead>
<tbody>
<tr>
<td>KW</td>
<td>4.16</td>
<td>N/A</td>
<td>72.5%</td>
</tr>
<tr>
<td>Intrinsic</td>
<td>5.16</td>
<td>2.35</td>
<td>24.3%</td>
</tr>
<tr>
<td>Headers</td>
<td>10.34</td>
<td>4.75</td>
<td>3.2%</td>
</tr>
</tbody>
</table>
Three types of queries

Keywords
- KW extraction based on tf-idf
- KW enriched with collocations
- non positionable
- unconditionally executed
- 4 KW queries per doc on average
- covering the whole document
- e.g., mobile phone signal prepaid cards

Intrinsic plagiarism
- positionable
- conditionally executed
- covering the suspicious passage
- based on Average Word Frequency Class
- it's change indicates suspicious passage
- e.g., in protector fights shifting lemma

Headers
- naive headers detection
- positionable
- conditionally executed
- covering the part that introduces by the header
- e.g., Differences Between Cell Phones

Suspicious document

<table>
<thead>
<tr>
<th>Query type</th>
<th>Extracted/Expected Similarities percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>KW</td>
<td>4.16</td>
</tr>
<tr>
<td>Pattern</td>
<td>3.14</td>
</tr>
<tr>
<td>Headers</td>
<td>4.35</td>
</tr>
<tr>
<td>Index</td>
<td>1.04</td>
</tr>
<tr>
<td>Overall</td>
<td>4.05</td>
</tr>
</tbody>
</table>

Prezi
Three Way Search Engine with Queries
Multi-feature Document Comparison

Šimon Suchomel and Jan Kasprzak
with Michal Brandejs

Faculty of Informatics
Masaryk University
Detailed Comparison

Algorithm

Tokenization
- Sequences of Unicode letters
- Words

Common Features

Valid intervals of common features

Postprocessing

Performance
- 24-core server
- 4x AMD 8139

Implementation
- Pure Perl
- 669 lines of code

41 minutes
Q(# of documents)

14 seconds!
Q(# of document pairs)
Algorithm

Tokenization
Sequences of Unicode letters

Words

Common Features

Postprocessing

Valid intervals of common features

Overlapping detections

Word 5-grams
3+ characters, sorted

Stopword 8-grams
50 most-frequent English words

[Stamatatos, 2011]

Neighbouring detections
Heuristics for merging
Up to 30,000 characters gap
Tokenization
Sequences of Unicode letters
Words
Features

Word 5-grams
What do you get if you multiply six by nine?

3+ characters, sorted

Stopword 8-grams
50 most-frequent English words

[Stamatatos, 2011]
Word 5-grams

What do you get if you multiply six by nine?

3+ characters, sorted

get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

42
Features

Word 5-grams
What do you get if you multiply six by nine?

3+ characters, sorted

Stopword 8-grams
50 most-frequent English words

[Stamatatos, 2011]
Stopword 8-grams

50 most-frequent English words

[Stamatatos, 2011]
Common Features

Word 5-grams
3+ characters, sorted

Stopword 8-grams
50 most-frequent

What do you get by it being

unsorted
Valid intervals of common features

Ordered features

Standard approach

Kasprzak, 2009
Stamatatos, 2011

Suspicious

source

Word 5-gram

Stopword 8-gram

No strict ordering

Identify by (offset, length)
Not by an ordinal number
Standard approach

Ordered features

[suspicious source]

[Kasprzak, 2009]
[Stamatatatos, 2011]
suspicious

word 5-gram

source

stopword 8-gram

No strict ordering
No strict ordering

Identify by (offset, length)!

Not by an ordinal number
Valid intervals of common features

Ordered features

Standard approach

[s Kasprzak, 2009]
[Stamatatos, 2011]

suspicious

source

suspicious

source

word 5-gram

stopword 8-gram

No strict ordering

Identify by (offset, length)!

Not by an ordinal number
Algorithm

Tokenization
Sequences of Unicode letters

Common Features
Words

Postprocessing

Valid intervals of common features

Word 5-grams
3+ characters, sorted

Stopword 8-grams
50 most-frequent English words, unsorted

Overlapping detections
> 300 characters?
NO

Neighbouring detections
Heuristics for merging
Up to 30,000 characters gap

Ordered features

Suspicious word 5-gram
Stopword 8-gram

No strict ordering
Identity by offset, length only!
Postprocessing

**Overlapping detections**

> 300 characters?

- **NO**
  - Drop both

- **YES**
  - Keep the longer one

**Neighbouring detections**

- Heuristics for merging
  - Up to 30,000 characters gap
  - Biased for recall
Overlapping detections

> 300 characters?

NO
Drop both

YES
Keep the longer one
Postprocessing

Overlapping detections

> 300 characters?

- NO: Drop both
- YES: Keep the longer one

Neighbouring detections

- Heuristics for merging
- Up to 30,000 characters gap
Neighbouring detections

Heuristics for merging

Up to 30,000 characters gap
30,000 biased for recall

that's several pages!
Algorithm

Tokenization
Sequences of Unicode letters

Common Features

Words

Valid intervals of common features

Postprocessing

Word 5-grams
3+ characters, sorted

Stopword 8-grams
50 most-frequent English words

0
Word 5-grams
3+ characters, sorted

Stopword 8-grams
50 most-frequent English words

Overlapping detections
> 300 characters?
NO
YES

Neighbouring detections
Heuristics for merging
Up to 30,000 characters gap
Performance

24-core server

4x AMD 8139
41 minutes

0 (# of documents)

Tokenization
Sequences of Unicode letters

Words

Common Features

Word
5-grams
3+ characters, sorted
What do you get if you multiply six by nine?

Stopword
8-grams
50 most-frequent English words

[Stamatatos, 2011]
Valid intervals of common features

Postprocessing

Overlapping detections

> 300 characters?

No / Yes

Neighbouring detections

Heuristics for merging

14 seconds!

O(# of document pairs)
Implementation

Pure Perl
669 lines of code
Detailed Comparison

Algorithm

Tokenization
Sequences of Unicode letters

Common Features

Valid intervals
of common features

Postprocessing

Performance
24-core server
4x AMD 8139

Implementation
Pure Perl
669 lines of code

41 minutes
Q(# of documents)

14 seconds!
Q(# of document pairs)
Conclusions

Three types of queries

Three Way Search Engine with Queries

Multi-feature Document Comparison

Šimon Suchomel and Jan Kasprzak with Michal Brandejs

Faculty of Informatics Masaryk University

Narrowed search

KW, Intrinsic, Headers with befitting aiming and combination

Multiple types of features

Purely based on (offset, length)
Narrowed search

KW, Intrinsic, Headers with befitting aiming and combination
Multiple types of features

Purely based on (offset, length)
Conclusions

Three types of queries
- Keywords
- Intrinsic plagiarism
- Headers

Three Way Search Engine with Queries
Šimon Suchomel and Jan Kasprzak with Michal Brandejs
Faculty of Informatics, Masaryk University

Multi-feature Document Comparison

Detailed Comparison

Narrowed search
KW, Intrinsic, Headers with befitting aiming and combination

Multiple types of features
Purely based on (offset, length)