PAN 2015

13th evaluation lab on uncovering plagiarism, authorship, and social software misuse

JOINT TALK ON THREE DATA SUBMISSIONS TO TEXT ALIGNMENT AND ONE SOURCE RETRIEVAL ALGORITHM

Mostafa Dehghani
ICT Research Institute, ACECR, Iran
September, 10, 2015

ACECR

Outline of My Talk

Outline of My Talk

A. **Data Submissions to Text Alignment:**

- Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation
- Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus
- Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus

Outline of My Talk

A. **Data Submissions to Text Alignment:**

- Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation
- Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus
- Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus
- Evaluation of Text Reuse Corpora for Text Alignment Task of plagiarism Detection

Outline of My Talk

A. **Data Submissions to Text Alignment:**
   - Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation
   - Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus
   - Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus
   - Evaluation of Text Reuse Corpora for Text Alignment Task of plagiarism Detection

B. **Source Retrieval Plagiarism Detection based on Noun Phrase and Keyword Phrase Extraction**

A

Data Submissions to Text Alignment

Corpus Construction Steps

-
-
-
-
-

Corpus Construction Steps

- Preprocessing
.
.
.
.

Corpus Construction Steps

- Preprocessing
- Clustering
- .
- .
- .

Corpus Construction Steps

- Preprocessing
- Clustering
- Fragment Extraction

Corpus Construction Steps

- Preprocessing
- Clustering
- Fragment Extraction
- Fragment Obfuscation

Corpus Construction Steps

- Preprocessing
- Clustering
- Fragment Extraction
- Fragment Obfuscation
- Inserting Plagiarism Cases into Documents

Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation

Data resource:

Wikipedia Articles

Mono Lingual Persian Corpus

Mono Lingual Persian Corpus

- Preprocessing
  - Persian is one of the Indo-European languages which have borrowed its script from Arabic, a member of the Semitic language family

Mono Lingual Persian Corpus

- Preprocessing
  - Persian is one of the Indo-European languages which have borrowed its script from Arabic, a member of the Semitic language family
- Clustering
  - In this step, collection of Wikipedia documents clustered into different topically related groups
  - A bipartite graph of documents-categories was created to cluster the documents
  - In the next step, the Infomap community detection algorithm was applied to the graph and all communities were detected
  - Finally, Documents within a community are considered as one cluster

Mono Lingual Persian Corpus

- Fragment Extraction

Mono Lingual Persian Corpus

- Fragment Extraction
  - Divided Documents into Two Categories:
    - 50% Source Documents
    - 50% Suspicious Documents : 25% with Plagiarism – 25% no Plagiarism

Mono Lingual Persian Corpus

- Fragment Extraction
  - Divided Documents into Two Categories:
    - 50% Source Documents
    - 50% Suspicious Documents : 25% with Plagiarism – 25% no Plagiarism
  - The task of the fragment extraction is to extract fragments from source documents.

Mono Lingual Persian Corpus

- Fragment Extraction
  - Divided Documents into Two Categories:
    - 50% Source Documents
    - 50% Suspicious Documents : 25% with Plagiarism – 25% no Plagiarism
  - The task of the fragment extraction is to extract fragments from source documents.

Fragment Length	
Short	30 – 50 words
Medium	150 – 250 words
Long	300 – 500 words

Mono Lingual Persian Corpus

- Fragment Extraction
  - Divided Documents into Two Categories:
    - 50% Source Documents
    - 50% Suspicious Documents : 25% with Plagiarism – 25% no Plagiarism
  - The task of the fragment extraction is to extract fragments from source documents.

Mono Lingual Persian Corpus

- Fragment Extraction
  - Divided Documents into Two Categories:
    - 50% Source Documents
    - 50% Suspicious Documents : 25% with Plagiarism – 25% no Plagiarism
  - The task of the fragment extraction is to extract fragments from source documents.
- Fragment Obfuscation

Mono Lingual Persian Corpus

- Fragment Extraction
  - Divided Documents into Two Categories:
    - 50% Source Documents
    - 50% Suspicious Documents : 25% with Plagiarism – 25% no Plagiarism
  - The task of the fragment extraction is to extract fragments from source documents.
- Fragment Obfuscation
  - Artificial Obfuscation
    - None (No Obfuscation)
    - Random Change of Order
    - POS-preserving Change of Order
    - Synonym Substitution
    - Addition / Deletion

Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents

Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
  - In this step, according to suspicious document's length, one or more plagiarism cases are selected.

Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
  - In this step, according to suspicious document's length, one or more plagiarism cases are selected.

Plagiarism per Document	
Little	5% - 20%
Medium	20% - 50%
Much	50% - 80%
Very Much	80% - 100%

Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
  - In this step, according to suspicious document's length, one or more plagiarism cases are selected.

Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
  - In this step, according to suspicious document's length, one or more plagiarism cases are selected.
  - Each of selected cases inserted at random positions in suspicious document.

Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
  - In this step, according to suspicious document's length, one or more plagiarism cases are selected.
  - Each of selected cases inserted at random positions in suspicious document.
  - Each suspicious document and its corresponding source documents are selected from one cluster.

Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
  - In this step, according to suspicious document's length, one or more plagiarism cases are selected.
  - Each of selected cases inserted at random positions in suspicious document.
  - Each suspicious document and its corresponding source documents are selected from one cluster.

```
<?xml version="1.0" encoding="UTF-8"?>
<document xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" ....
  <feature name="project-gutenberg" etext_number="7243" url="http://www.gutenberg.org/files/....
  <feature name="language" value="en" />
  <feature name="artificial-plagiarism" translation="false" obfuscation="none"
  this_offset="487" this_length="4218" source_reference="source-document03471.txt"
  source_offset="10866" source_length="4226" />

  <feature name="artificial-plagiarism" translation="false" obfuscation="low"
  this_offset="7507" this_length="1872" source_reference="source-document03471.txt"
  source_offset="4846" source_length="1792" />

  <feature name="artificial-plagiarism" translation="false" obfuscation="low"
  this_offset="10626" this_length="805" source_reference="source-document03471.txt"
  source_offset="2399" source_length="800" />
  </document>
```

Mono Lingual Persian Corpus

- Results

Documents	
The number of source documents:	1057
The number of suspicious documents:	
With plagiarism:	529
No plagiarism:	528
Plagiarism Cases	
The number of plagiarism cases:	
No obfuscation cases:	259
With obfuscation cases:	564
Plagiarism per Document	
The number of Little plagiarized documents:	301
The number of Medium plagiarized documents:	80
The number of Much plagiarized documents:	96
The number of Very much plagiarized documents:	52

Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus

Data resources:

- Wikipedia Articles
- SemEval Dataset

Mono Lingual English Corpus

- Clustering

Mono Lingual English Corpus

- Clustering
- Fragment Extraction
  - Method 1: The fragments are extracted from source documents.
  - Method 2: The fragments are generated based on SemEval dataset sentences.

Mono Lingual English Corpus

- Clustering
- Fragment Extraction
  - Method 1: The fragments are extracted from source documents.
  - Method 2: The fragments are generated based on SemEval dataset sentences.

Fragment Length	
Short	3 – 5 sentences
Medium	6 – 8 sentences
Long	9 – 12 sentences

Mono Lingual English Corpus

- Clustering
- Fragment Extraction
  - Method 1: The fragments are extracted from source documents.
  - Method 2: The fragments are generated based on SemEval dataset sentences.

Mono Lingual English Corpus

- Fragment Obfuscation

Mono Lingual English Corpus

- Fragment Obfuscation
  - Artificial Obfuscation

Mono Lingual English Corpus

- Fragment Obfuscation
  - Artificial Obfuscation
  - Simulated Obfuscation
    - The pairs of sentences from the SemEval dataset with their corresponding similarity score are used for constructing the simulated plagiarism cases.
    - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment.

Mono Lingual English Corpus

- Fragment Obfuscation
  - Artificial Obfuscation
  - Simulated Obfuscation
    - The pairs of sentences from the SemEval dataset with their corresponding similarity score are used for constructing the simulated plagiarism cases.
    - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment.

Degree	Similarity Scores of Sentences		
	3	4	5
Low	-	1% -15%	85% - 100%
Medium	25% - 45%		55%- 75%
High	45% - 65%		35% - 55%

Mono Lingual English Corpus

- Fragment Obfuscation
  - Artificial Obfuscation
  - Simulated Obfuscation
    - The pairs of sentences from the SemEval dataset with their corresponding similarity score are used for constructing the simulated plagiarism cases.
    - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment.

Mono Lingual English Corpus

- Fragment Obfuscation
  - Artificial Obfuscation
  - Simulated Obfuscation
    - The pairs of sentences from the SemEval dataset with their corresponding similarity score are used for constructing the simulated plagiarism cases.
    - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment.
- Inserting Plagiarism Cases into Documents

Mono Lingual English Corpus

- Fragment Obfuscation
  - Artificial Obfuscation
  - Simulated Obfuscation
    - The pairs of sentences from the SemEval dataset with their corresponding similarity score are used for constructing the simulated plagiarism cases.
    - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment.

- Inserting Plagiarism Cases into Documents

Plagiarism per Document	
Hardly	5% - 20%
Medium	20% - 40%
Much	40% - 60%

Mono Lingual English Corpus

- Results

Statistics	
Documents	
The number of source documents:	3309
The number of suspicious documents:	952
Plagiarism per Document	
Hardly (5% - 20%)	60%
Medium (20% - 40%)	25%
Much (40% - 60%)	15%
Plagiarism Cases	
The number of plagiarism cases:	
- No obfuscation cases:	10%
- Random obfuscation:	78%
- Simulated obfuscation:	12%
Case Length Statistics	
Short (3 – 5 sentences):	50%
Medium (6 – 8 sentences):	32%
Long (9 – 12 sentences):	18%

Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus

Data resources:

- Wikipedia Articles
- Persian-English Parallel Corpus

Bilingual Persian-English Corpus

- Clustering
  -
  -

Bilingual Persian-English Corpus

- Clustering

  Parallel Sentences Clustering

  1. Persian Wikipedia documents were indexed by the Apache Lucene library.
  2. We built a query from each Persian sentence
  3. The query was searched in the indexed documents and returns the top document.
  4. A bipartite graph of return documents-categories was created. Then, the info- map community detection algorithm was applied to the graph and all communities were detected. Documents within a community are considered as one cluster.
  5. Finally, parallel sentences were assigned to the documents in the same cluster.

Bilingual Persian-English Corpus

Clustering

Parallel Sentences Clustering

1. Persian Wikipedia documents were indexed by the Apache Lucene library.
2. We built a query from each Persian sentence
3. The query was searched in the indexed documents and returns the top document.
4. A bipartite graph of return documents-categories was created. Then, the info- map community detection algorithm was applied to the graph and all communities were detected. Documents within a community are considered as one cluster.
5. Finally, parallel sentences were assigned to the documents in the same cluster.

Documents Clustering

- For each cluster of return documents in the previous stage, the categories of documents have been extracted and considered as label of that cluster.
- The basic documents collected into different topically related clusters based on their categories. The documents are assigned to the cluster with maximum common categories.

Bilingual Persian-English Corpus

- Fragment Extraction

Bilingual Persian-English Corpus

- Fragment Extraction
  - Plagiarism cases are constructed from parallel sentences.
  - Source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.

Bilingual Persian-English Corpus

- Fragment Extraction
  - Plagiarism cases are constructed from parallel sentences.
  - Source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.

Fragment Length	
Short	3 – 5 sentences
Medium	5 – 10 sentences
Long	10 – 15 sentences

Bilingual Persian-English Corpus

- Fragment Extraction
  - Plagiarism cases are constructed from parallel sentences.
  - Source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.

Bilingual Persian-English Corpus

- Fragment Extraction
  - Plagiarism cases are constructed from parallel sentences.
  - Source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.
- Fragment Obfuscation
  - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with different similarity score were chosen.

Bilingual Persian-English Corpus

Fragment Extraction

- Plagiarism cases are constructed from parallel sentences.
- Source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.

Fragment Obfuscation

- To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with different similarity score were chosen.

Degree	Similarity scores of sentences in fragments		
	1- 0.85	0.85 – 0.65	0.65 – 0.85
Low	100%	-	-
Medium	55% - 75%	25% - 45%	-
High	35% - 55%	-	45% - 65%

Bilingual Persian-English Corpus

- Inserting Plagiarism Cases into Documents

Bilingual Persian-English Corpus

- Inserting Plagiarism Cases into Documents
  - In this step, according to suspicious document's length, one or more plagiarism cases are selected.
  - Persian documents considering as suspicious documents and source documents are English documents.

Bilingual Persian-English Corpus

- Inserting Plagiarism Cases into Documents
  - In this step, according to suspicious document's length, one or more plagiarism cases are selected.
  - Persian documents considering as suspicious documents and source documents are English documents.

Plagiarism per Document	
Low	5% - 20%
Medium	20% - 40%
High	40% - 60%

Bilingual Persian-English Corpus

- Inserting Plagiarism Cases into Documents
  - In this step, according to suspicious document's length, one or more plagiarism cases are selected.
  - Persian documents considering as suspicious documents and source documents are English documents.
  - English fragment inserted at random positions in source documents and its corresponding Persian fragments has been inserted into suspicious documents.
  - Each suspicious document and its corresponding source documents are selected from one cluster.

Plagiarism per Document	
Low	5% - 20%
Medium	20% - 40%
High	40% - 60%

Bilingual Persian-English Corpus

- Results

Documents	
The number of source documents (English):	19973
The number of suspicious documents (Persian):	
• With plagiarism:	3571
No plagiarism:	3571
Plagiarism cases	
The number of plagiarism cases:	11200
Plagiarism per Document	
The number of Little plagiarized documents	2035
The number of Medium plagiarized documents	536
The number of Much plagiarized documents	642
The number of Very much plagiarized documents	58

Evaluation of Text Reuse Corpora for Text Alignment Task of plagiarism Detection

Evaluation of Corpus Submissions to PAN 2015

Corpora Statistical Information

Corpora Statistical Information

	cheema15	hanif15	Kong15	Alvi15	Palkovskii15
Type of Corpus	Mono-Lingual	Bi-Lingual	Mono-Lingual	Mono-Lingual	Mono-Lingual
Source–Suspicious Language	English-English	Urdu-English	Chinese-Chinese	English-English	English- English
Resource Documents	Gutenberg books and Wikipedia	Wikipedia pages	Chinese thesis and http://wenku.baidu.com/ website	"The Complete Grimm's Fairy Tales" book	Internet web pages crawling

Corpora Statistical Information

Corpora Statistical Information

	Cheema15	Hanif15	Kong15	Alvi15	Palkovskii15
Number of Docs					
• **Suspicious Docs**	248	250	4	90	1175
• **Source Docs**	248	250	78	70	1950
Length of Docs (in chars)					
• **Min Length**	2263	361	394	514	519
• **Max Length**	22471	74083	121829	45222	517925
• **Average Length**	7239	4382	42839	7718	6512
Length of Plagiarisms Cases (in chars)					
• **Min Length**	134	78	62	259	157
• **Max Length**	2439	849	2748	1160	14336
• **Average Length**	503	361	423	464	782

Corpora Statistical Information

Corpora Statistical Information

Obfuscation Strategies	Cheema15	Hanif15	Kong15	Alvi15	Palkovskii15
Simulated	123	135	-	-	-
Real	-	-	109	-	-
Automatic	-	-	-	25	-
Retelling-Human	-	-	-	25	-
Character-Substitution	-	-	-	25	-
Translation	-	-	-	-	618
Summary	-	-	-	-	1292
Random	-	-	-	-	626
None	-	-	-	-	624
Sum	123	135	109	75	3160

Manual Evaluation of Corpora

Manual Evaluation of Corpora

- Manually investigate twenty pairs of corresponding source and suspicious fragments in each corpus

Manual Evaluation of Corpora

- Manually investigate twenty pairs of corresponding source and suspicious fragments in each corpus
  - Changes in syntactic structure between source and plagiarized passage

.

.

Manual Evaluation of Corpora

- Manually investigate twenty pairs of corresponding source and suspicious fragments in each corpus
  - Changes in syntactic structure between source and plagiarized passage
  - Concept preserving from source passage to plagiarized passage

Manual Evaluation of Corpora

- Manually investigate twenty pairs of corresponding source and suspicious fragments in each corpus
  - Changes in syntactic structure between source and plagiarized passage
  - Concept preserving from source passage to plagiarized passage
  - Distribution of obfuscation types in suspicious documents

Automatic Evaluation of Corpora

Automatic Evaluation of Corpora

- Evaluating two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus

Automatic Evaluation of Corpora

- Evaluating two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus
- For Kong15 corpus

Automatic Evaluation of Corpora

- Evaluating two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus
- For Kong15 corpus
  - All source and correspond suspicious fragments are extracted, and the total number of similar “characters n-grams” between source and suspicious plagiarized passages are calculated for n in range of one to four .

Automatic Evaluation of Corpora

- Evaluating two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus
- For Kong15 corpus
  - All source and correspond suspicious fragments are extracted, and the total number of similar “characters n-grams” between source and suspicious plagiarized passages are calculated for n in range of one to four .
- For evaluation of summary obfuscation

.

Automatic Evaluation of Corpora

- Evaluating two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus
- For Kong15 corpus
  - All source and correspond suspicious fragments are extracted, and the total number of similar “characters n-grams” between source and suspicious plagiarized passages are calculated for n in range of one to four .
- For evaluation of summary obfuscation
  - From the point of “concept preserving” measure, we have extracted 10% of top words from source fragments based on tf.idf weight.

B

Source Retrieval based on Noun and Keyword Phrase Extraction

Data resources:

External PD Corpus of PAN 2011

Approach in Use: Five Steps

.

.

.

.

.

Approach in Use: Five Steps

- Suspicious Document Chunking

Approach in Use: Five Steps

- Suspicious Document Chunking
- Noun Phrase and Keyword Phrase Extraction

Approach in Use: Five Steps

- Suspicious Document Chunking
- Noun Phrase and Keyword Phrase Extraction
- Query Formulation

Approach in Use: Five Steps

- Suspicious Document Chunking
- Noun Phrase and Keyword Phrase Extraction
- Query Formulation
- Search Control

Approach in Use: Five Steps

- Suspicious Document Chunking
- Noun Phrase and Keyword Phrase Extraction
- Query Formulation
- Search Control
- Document Filtering and Downloading

Suspicious Document Chunking

Suspicious Document Chunking

- Segmentation of suspicious documents into parts called chunks

Suspicious Document Chunking

- Segmentation of suspicious documents into parts called chunks
- No fixed pattern to put one plagiarism fragment per chunk

Suspicious Document Chunking

- Segmentation of suspicious documents into parts called chunks
- No fixed pattern to put one plagiarism fragment per chunk
- Sufficient length of chunks, In order to comprise:
  1. At least one plagiarism fragment per chunk,
  2. And Maximum numbers of extracted queries from the chunks.

Suspicious Document Chunking

- Segmentation of suspicious documents into parts called chunks
- No fixed pattern to put one plagiarism fragment per chunk
- Sufficient length of chunks, In order to comprise:
  1. At least one plagiarism fragment per chunk,
  2. And Maximum numbers of extracted queries from the chunks.
- Individual sentences sets of 500 words Chunks as results.

Noun phrase and keyword phrase Extraction

Noun phrase and keyword phrase Extraction

Operation number	Operation Description
1	Selection of top 80% long sentences (based on length in chars)
2	Selection of top 80% sentences (based on number of nouns)
3	Selection of top three sentences (based on average tf.idf1 values)
4	Selection of top three sentences (based on number of words with highest values)

Noun phrase and keyword phrase Extraction

Operation number	Operation Description
1	Selection of top 80% long sentences (based on length in chars)
2	Selection of top 80% sentences (based on number of nouns)
3	Selection of top three sentences (based on average tf.idf1 values)
4	Selection of top three sentences (based on number of words with highest values)

- Scenario1: Operation 1→ Operation 2 → Operation 3 for noun phrase extraction

Noun phrase and keyword phrase Extraction

Operation number	Operation Description
1	Selection of top 80% long sentences (based on length in chars)
2	Selection of top 80% sentences (based on number of nouns)
3	Selection of top three sentences (based on average tf.idf1 values)
4	Selection of top three sentences (based on number of words with highest values)

- Scenario1: Operation 1→ Operation 2 → Operation 3 for noun phrase extraction
- Scenario2: Operation 1 → Operation 2 → Operation 4 for keyword phrase extraction

Noun phrase and keyword phrase Extraction

Operation number	Operation Description
1	Selection of top 80% long sentences (based on length in chars)
2	Selection of top 80% sentences (based on number of nouns)
3	Selection of top three sentences (based on average tf.idf1 values)
4	Selection of top three sentences (based on number of words with highest values)

- Scenario1: Operation 1→ Operation 2 → Operation 3 for noun phrase extraction
- Scenario2: Operation 1 → Operation 2 → Operation 4 for keyword phrase extraction
- Three sentences from each scenario1 and scenario2 selected to query formulation

Query Formulation

Query Formulation

- From each selected sentence, one query is extracted.

Query Formulation

- From each selected sentence, one query is extracted.
- The threshold for the number of words in each query is limited to ten.

Query Formulation

- From each selected sentence, one query is extracted.
- The threshold for the number of words in each query is limited to ten.
- Selection of high weighted terms to reach the ChatNoir limitation.

Query Formulation

- From each selected sentence, one query is extracted.
- The threshold for the number of words in each query is limited to ten.
- Selection of high weighted terms to reach the ChatNoir limitation.
- The terms are placed next to each other based on the order in sentence.

Download Filtering and Search Control

Download Filtering and Search Control

- **Download Filtering**

Download Filtering and Search Control

- **Download Filtering**
  - 14 top results are selected for each query

Download Filtering and Search Control

- **Download Filtering**
  - 14 top results are selected for each query
  - The query is divided into two sub-queries:
    - Snippet with the length of 500 characters are extracted as a sub-query.
    - Snippets are combined with each other and make a passage.

Download Filtering and Search Control

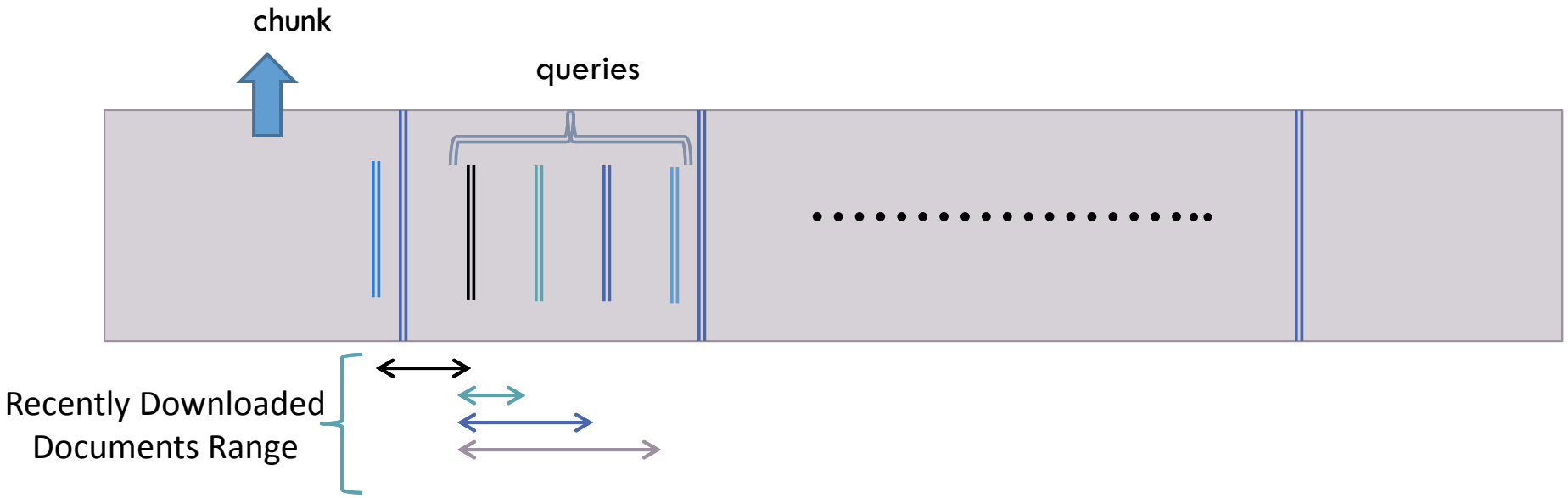
- **Download Filtering**
  - 14 top results are selected for each query
  - The query is divided into two sub-queries:
    - Snippet with the length of 500 characters are extracted as a sub-query.
    - Snippets are combined with each other and make a passage.
  - If the resulted passage contains at least 50% words of the query
    - The related document is downloaded
    - The document is maintained for search control operation

Download Filtering and Search Control

- **Download Filtering**
  - 14 top results are selected for each query
  - The query is divided into two sub-queries:
    - Snippet with the length of 500 characters are extracted as a sub-query.
    - Snippets are combined with each other and make a passage.
  - If the resulted passage contains at least 50% words of the query
    - The related document is downloaded
    - The document is maintained for search control operation
- **Search Control**
  - Drop a query when at least 60% of its terms are contained in recently downloaded documents set

Search Control

- Drop a query when at least 60% of its terms are contained in recently downloaded documents set



Evaluation

Evaluation

Downloads	F1	No Detection	Precision	Queries	Recall	Runtime
183.3	0.115	1	0.07539	43.5	0.41381	8:32:37

.

.

Evaluation

Downloads	F1	No Detection	Precision	Queries	Recall	Runtime
183.3	0.115	1	0.07539	43.5	0.41381	8:32:37

- Highest rank in “No Detection” measure.
- Highest rank in “Runtime” measure.

Thank you for Your Attention