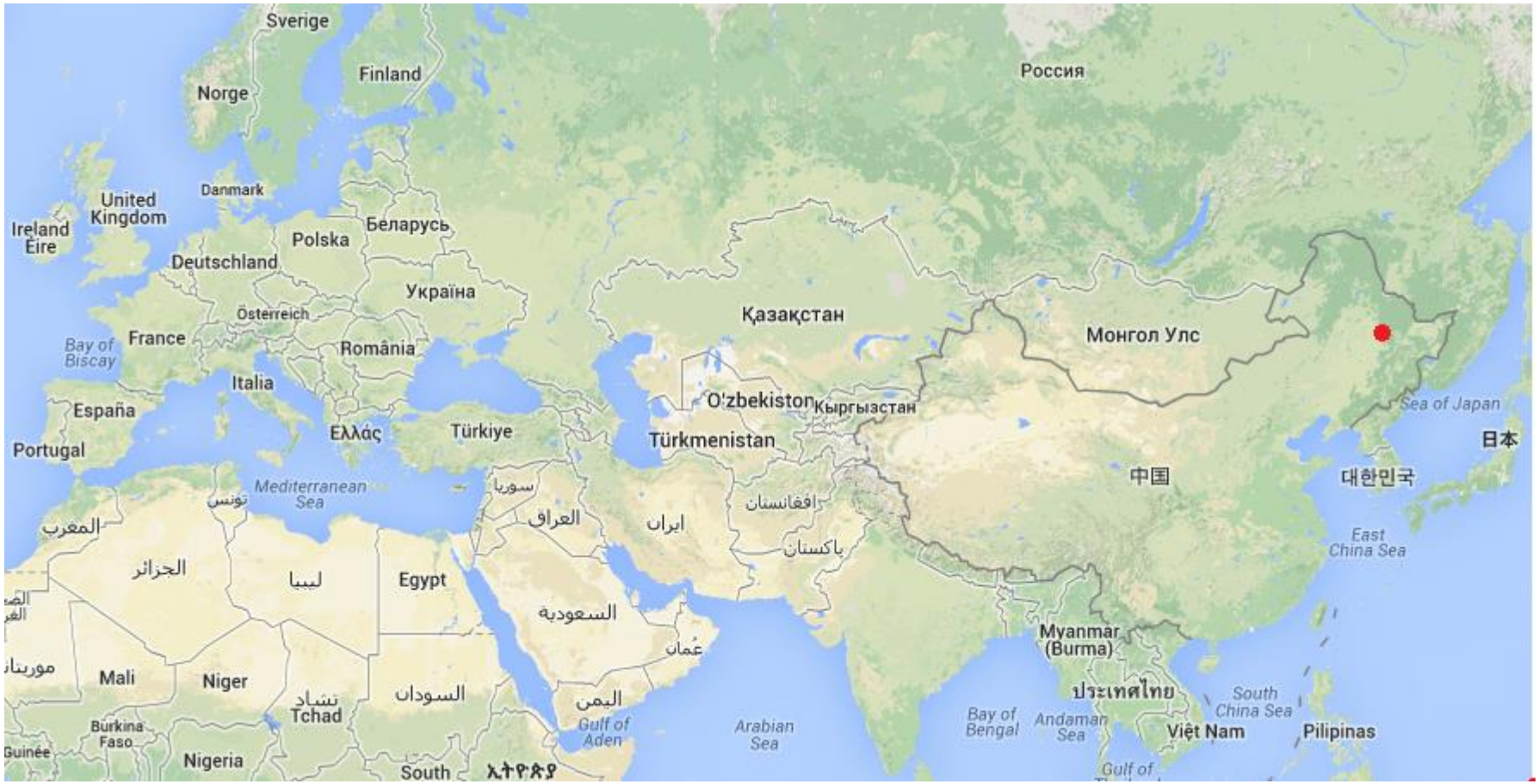
Approaches for Source Retrieval and Text Alignment of Plagiarism Detection

Kong Leilei, Qi Haoliang, Du Cuixia, Wang Mingxing, Han Zhongyuan

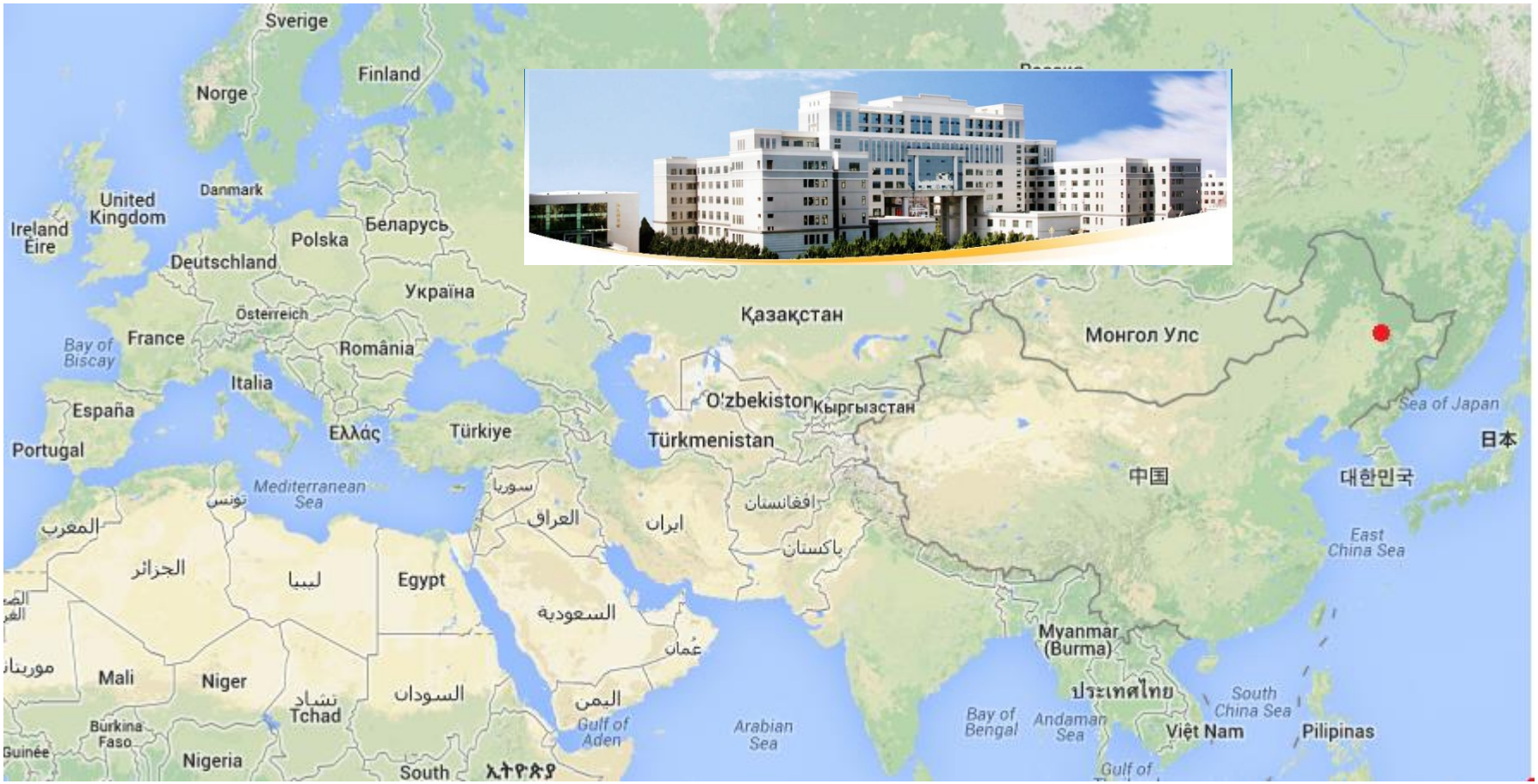
黑龙江工程学院
HEILONGJIANG INSTITUTE OF TECHNOLOGY

Who are we?

Who are we?


Sverige
Finland
Norge
Россия
Danmark
United Kingdom
Ireland
Éire
Беларусь
Polska
Deutschland
Україна
Österreich
Қазақстан
Монгол Улс
Bay of Biscay
France
România
Italia
O'zbekiston
Кыргызстан
España
Ελλάς
Türkiye
Türkmenistan
Portugal
Mediterranean Sea
سوريا
العراق
ایران
افغانستان
باكستان
中国
日本
대한민국
Sea of Japan
East China Sea
المغرب
الجزائر
ليبيا
Egypt
السعودية
عُمان
Myanmar (Burma)
Mali
Niger
تشاد
Tchad
السودان
اليمن
Gulf of Aden
Arabian Sea
Bay of Bengal
Andaman Sea
ประเทศไทย
South China Sea
Việt Nam
Pilipinas
Burkina Faso
Guinée
Nigeria
South
ኢትዮጵያ
Gulf of

Who are we?


Sverige
Finland
Norge
Danmark
United Kingdom
Ireland
Éire
Беларусь
Polska
Deutschland
Україна
Österreich
France
Bay of Biscay
România
Italia
España
Portugal
Ελλάς
Türkiye
Mediterranean Sea
تونس
المغرب
الجزائر
ليبيا
Egypt
سوريا
العراق
ایران
Қазақстан
O'zbekiston
Кыргызстан
Türkmenistan
افغانستان
پاکستان
السعودية
عُمان
اليمن
Gulf of Aden
Arabian Sea
Монгол Улс
中国
Sea of Japan
日本
대한민국
East China Sea
Myanmar (Burma)
ประเทศไทย
Bay of Bengal
Andaman Sea
South China Sea
Việt Nam
Pilipinas
Gulf of
Mali
Niger
Tchad
السودان
Burkina Faso
Guinée
Nigeria
South
ኢትዮጵያ

Who are we?


Heilongjiang Institute of Technology
Harbin, Heilongjiang Province, China
Sverige
Finland
Norge
United Kingdom
Ireland
Éire
Danmark
Беларусь
Polska
Deutschland
Україна
Österreich
France
Bay of Biscay
România
Italia
España
Portugal
Ελλάς
Türkiye
Mediterranean Sea
Қазақстан
Монгол Улс
O'zbekiston
Кыргызстан
Türkmenistan
中国
日本
대한민국
Sea of Japan
East China Sea
سوريا
العراق
ایران
افغانستان
پاکستان
المغرب
الجزائر
ليبيا
Egypt
السعودية
عُمان
اليمن
Gulf of Aden
Arabian Sea
Mali
Niger
Tchad
السودان
Burkina Faso
Guinée
Nigeria
South
ኢትዮጵያ
Myanmar (Burma)
ประเทศไทย
Bay of Bengal
Andaman Sea
South China Sea
Việt Nam
Pilipinas
Gulf of

Our University

Our University

Our University

Our University

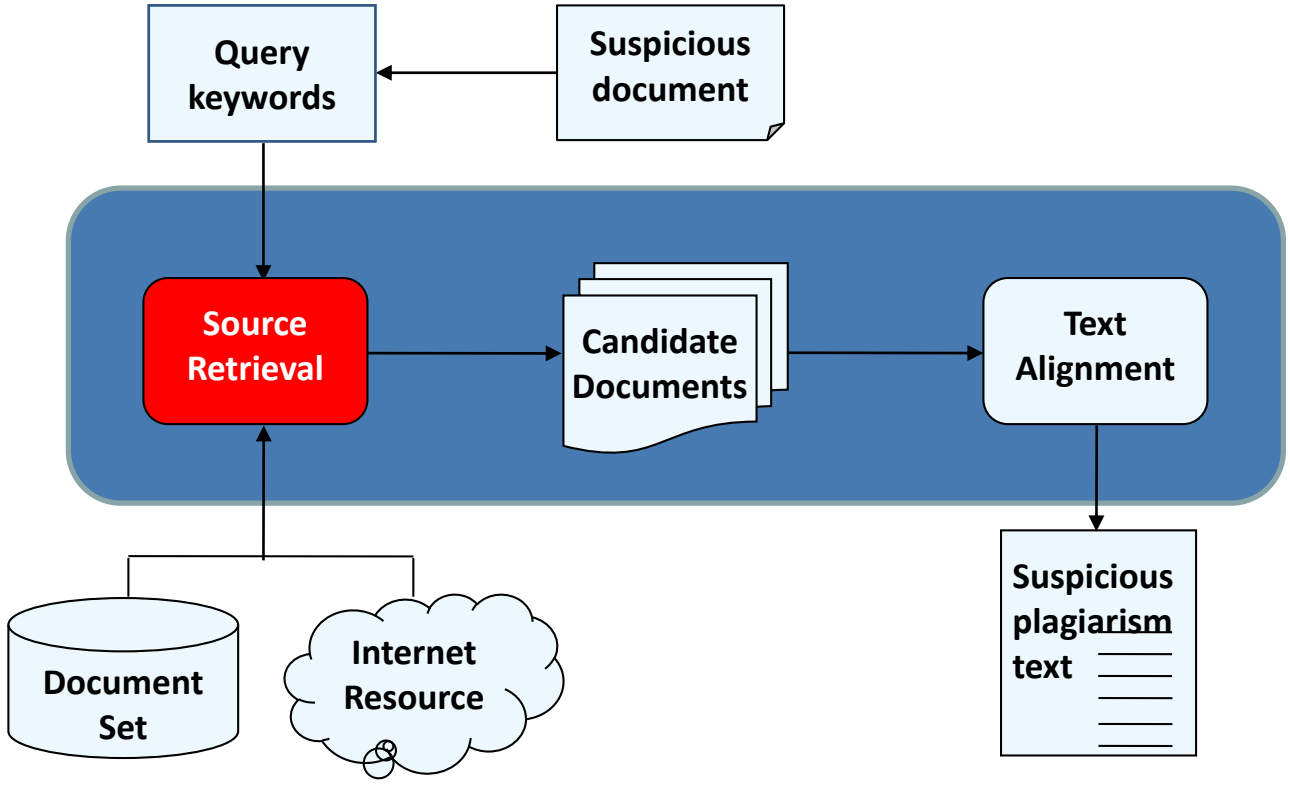
Our University

Our University

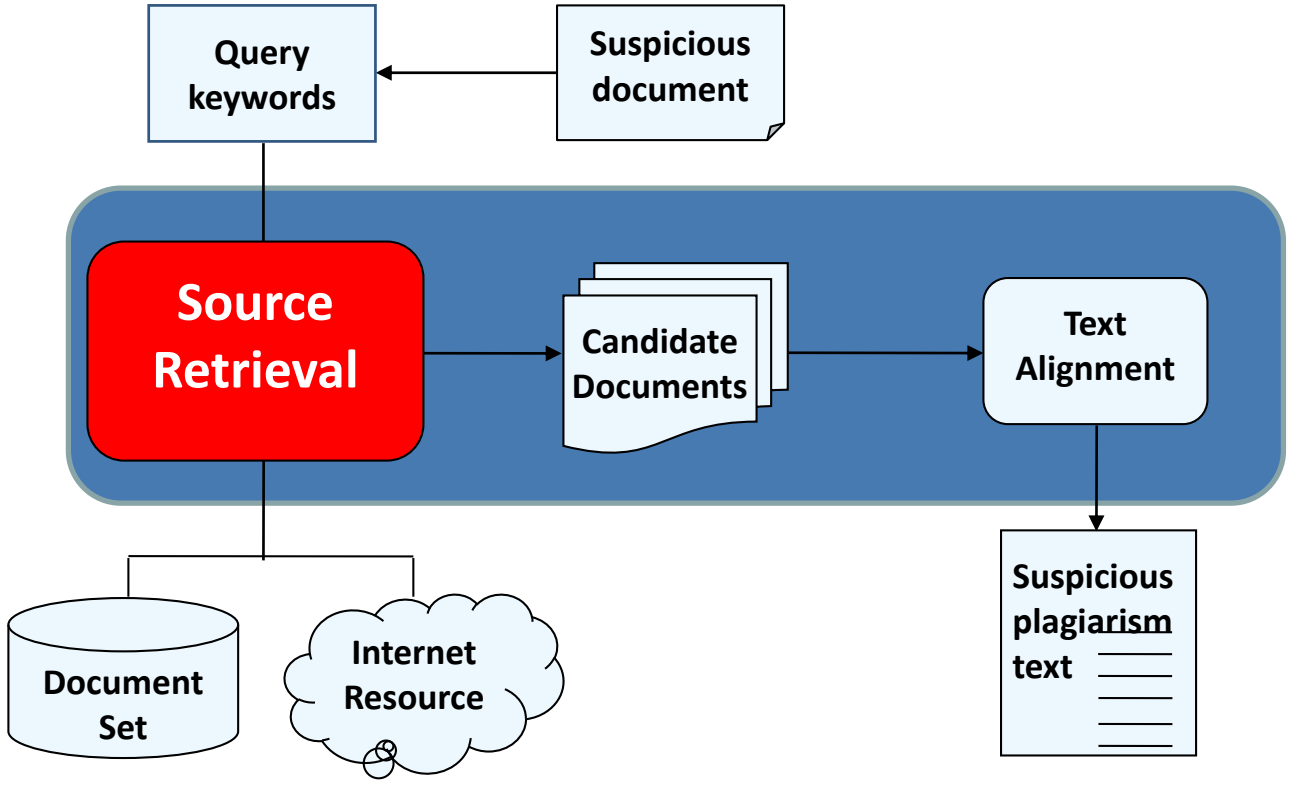
Index

- Approaches for Source Retrieval
- Approaches for Text Alignment
- Further works

Source Retrieval



Source Retrieval



2 problmes of source retrieval

- Tow core problem of source retrieval
  - Retrieval source is millions of documents from the Internet
    - This work was done by PAN
  - The query keywords of suspicious document which would be used for retrieval are not specified
    - How to extract query keyword is one of important issues of our work

Query Keywords Extraction

- Query Keywords Extraction Based on TF-IDF
- Query Keywords Extraction Based on Weighted TF-IDF
- Adjacent Query Keywords Extraction by PatTree
- Combination of Queries and Execution of Retrieval

Keywords Based on TF-IDF

- TF - term frequency, denotes the frequency of term i in document j

$$TF_i = tf_{ij}$$

- IDF - inverse document frequency

$$IDF = \log_2 (N / df_j)$$

- TF-IDF of term i is:

$$TFIDF(i) = TF(i) * IDF(i)$$

- *Tips: we found that the top 10 terms with the highest TF-IDF can obtain a good results*

Keywords Based on Weighted TF-IDF

- Weighted TF-IDF

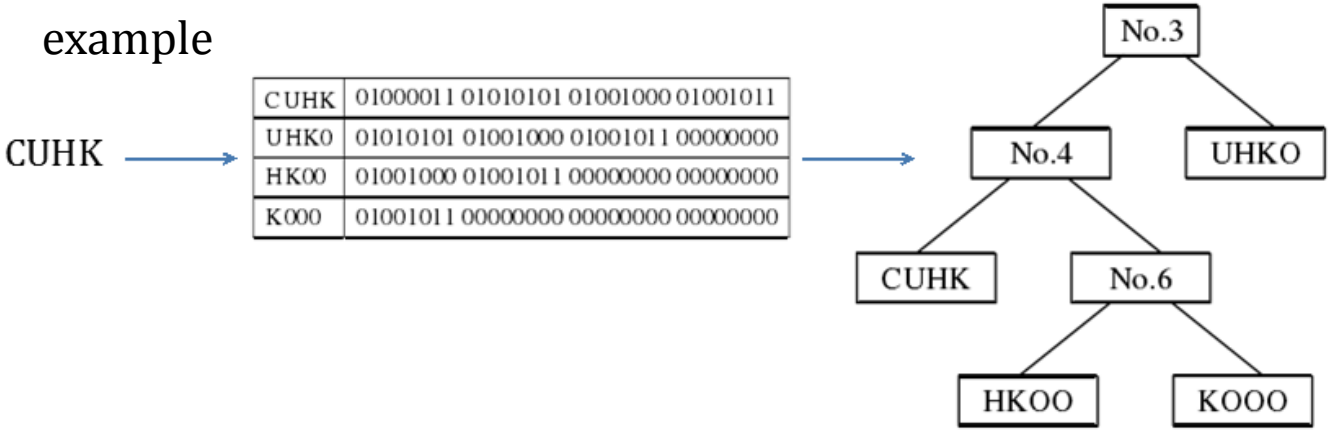
$$WTFIDF_i = weight * TFIDF_i$$

- Where *weight* is a weighted parameter, and we calculate the weight of term i according to its location

- *Tips: the keywords extraction based on the weighted TF-IDF sometimes is useful, sometimes useless.*

Adjacent Query Keywords Extraction by PatTree

- The adjacent string with high frequency is more important than a single word
- We use PatTree - an efficient data structure– to get the adjacent strings and their frequency



Combination of Queries

Table 1: Query Combination and Group

Query	Query Keywords
1	Top 1 to 5 query keywords based on TF-IDF
2	Top 2 to 10 query keywords based on TF-IDF
3	2-Gram query keywords based on PatTree
4	3-Gram query keywords based on PatTree
5	4-Gram query keywords based on PatTree
6	4-Gram query keywords based on PatTree
7	Top 1 to 5 query keywords based on weighted TF-IDF
8	Top 6 to 10 query keywords based on weighted TF-IDF
9	5-Gram query keywords based on PatTree

Results Source Retrieval subtask

Table2： Results Source of PAN@CLEF2013 Retrieval subtask

Workload	Queries	48.50
	Downloads	5691.47
Time to 1st Detection	Queries	2.46
	Downloads	285.66
Retrieved Performance	Precision	0.01
	Recall	0.65
No Detection		3

Results Source Retrieval subtask

Table2： Results Source of PAN@CLEF2013 Retrieval subtask

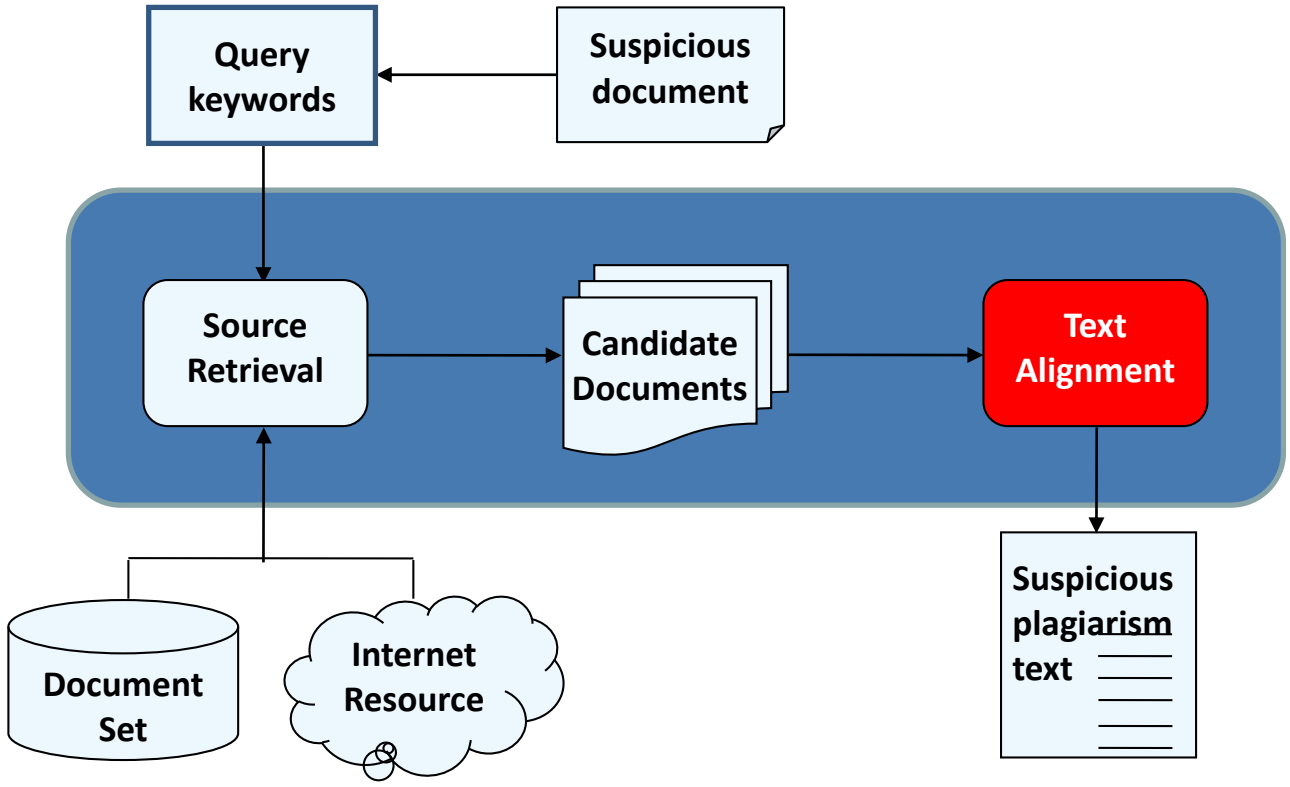
Workload	Queries	48.50
	Downloads	5691.47
Time to 1st Detection	Queries	2.46
	Downloads	285.66
Retrieved Performance	Precision	0.01
	Recall	0.65
No Detection		3

Results Source Retrieval subtask

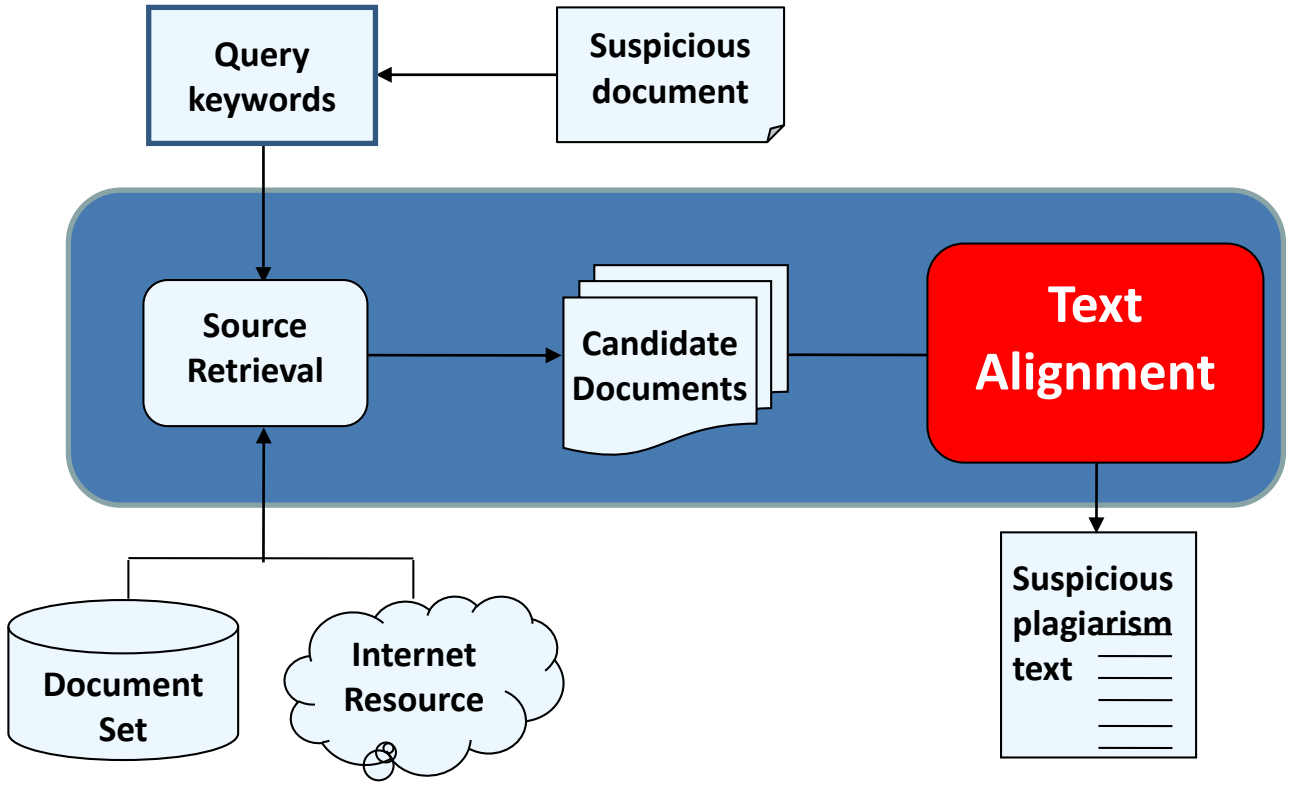
Table2： Results Source of PAN@CLEF2013 Retrieval subtask

Workload	Queries	48.50
	Downloads	5691.47
Time to 1st Detection	Queries	2.46
	Downloads	285.66
Retrieved Performance	Precision	0.01
	Recall	0.65
No Detection		3

Text Alignment



Text Alignment



Text Alignment

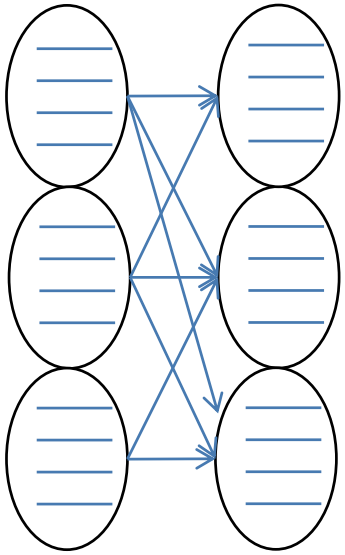
Text Alignment

Text Alignment

Seeding

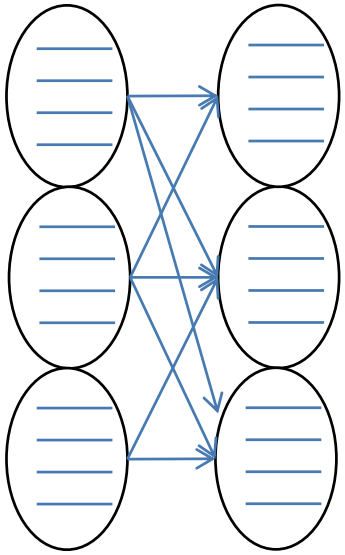
Text Alignment

Seeding

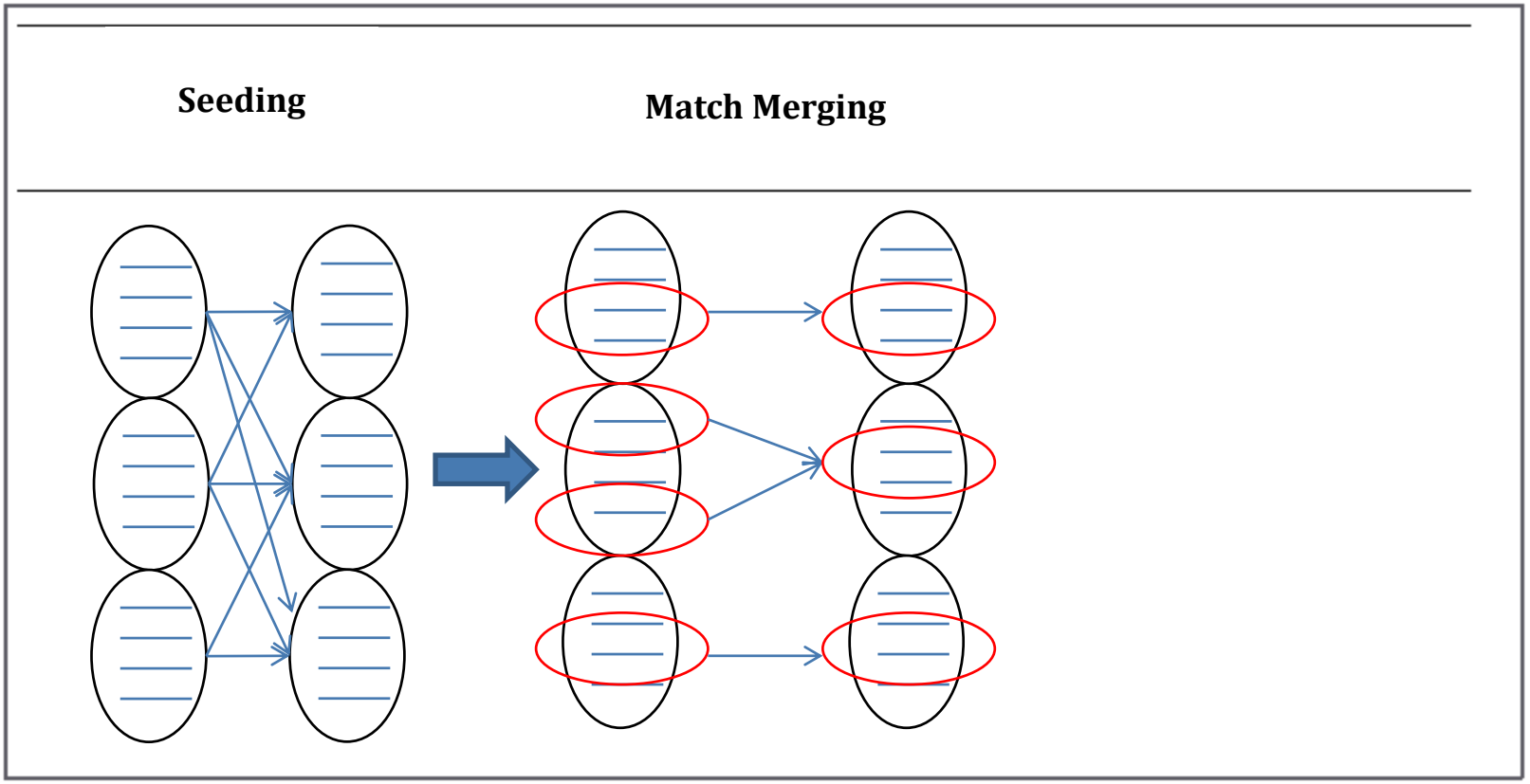


Text Alignment

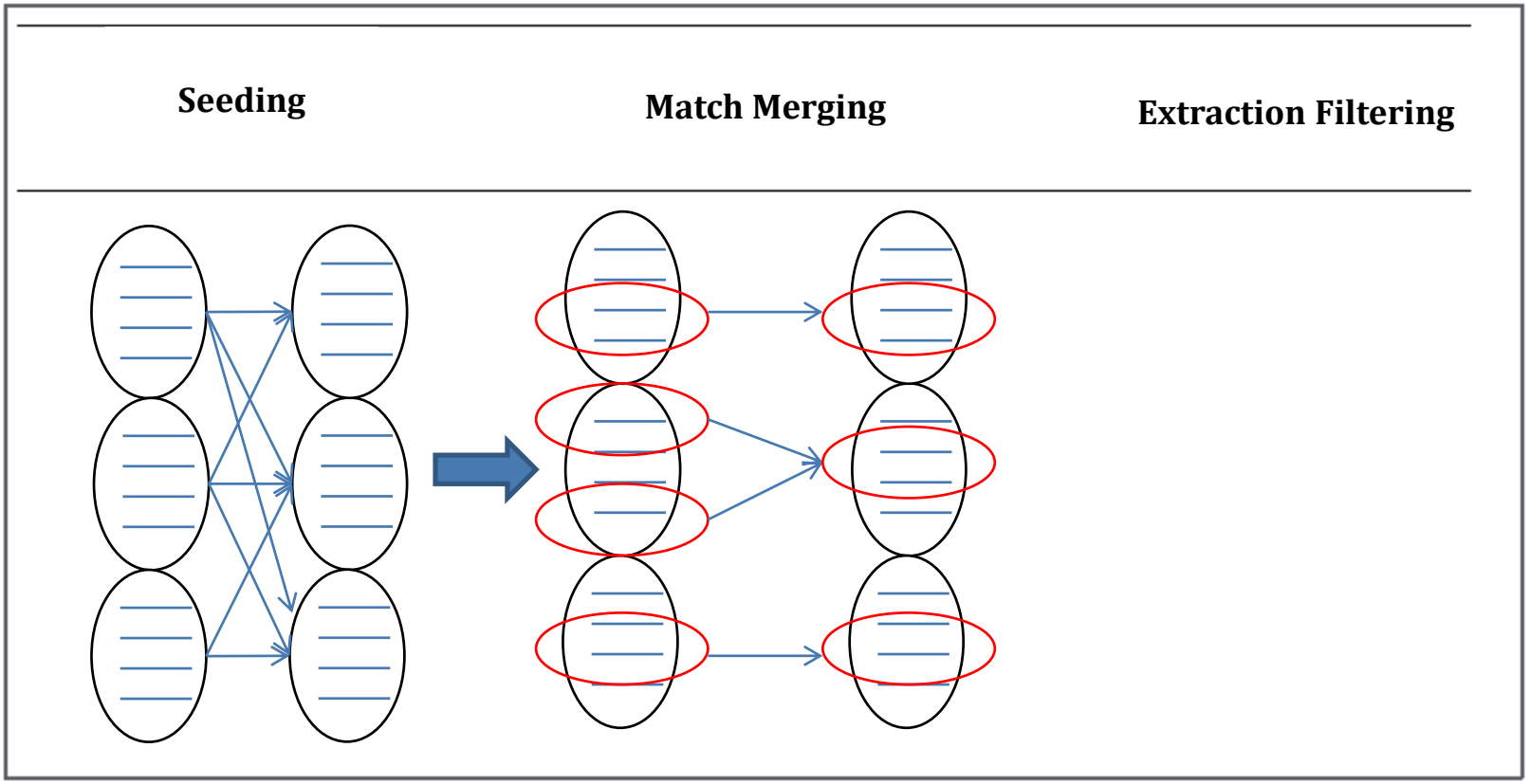
Seeding **Match Merging**



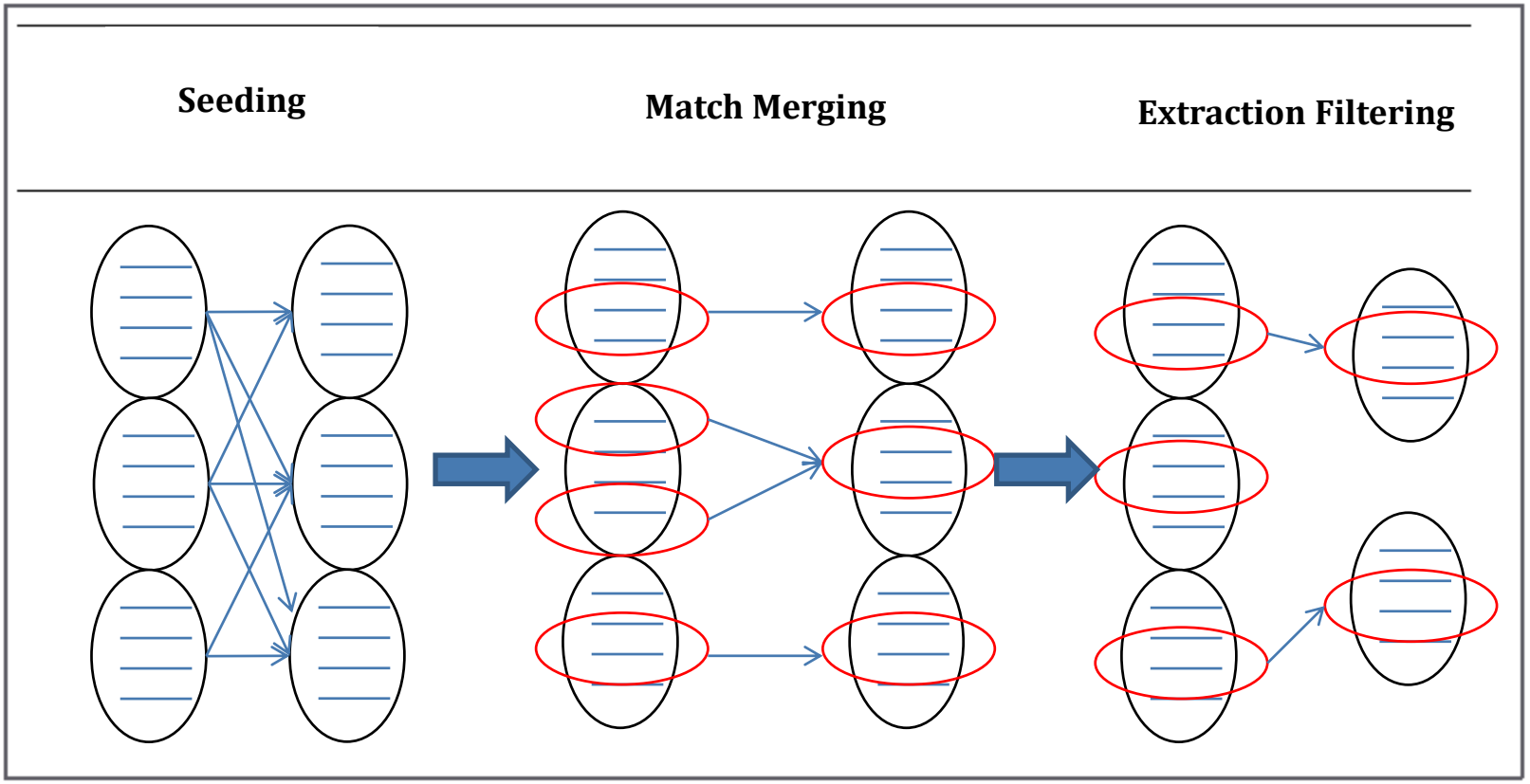
Text Alignment



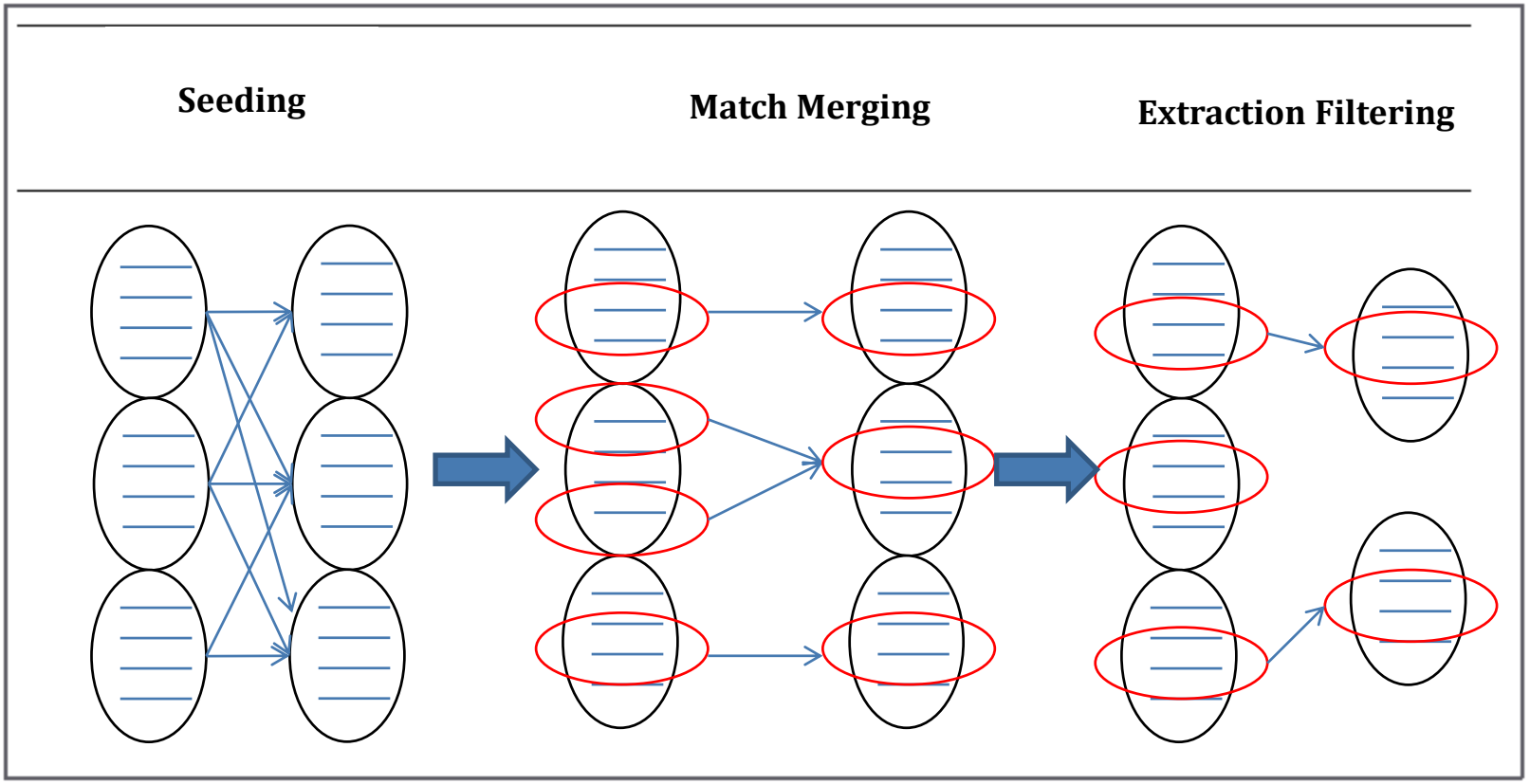
Text Alignment



Text Alignment

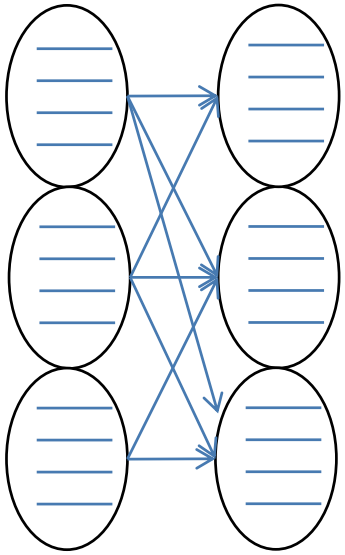


Text Alignment



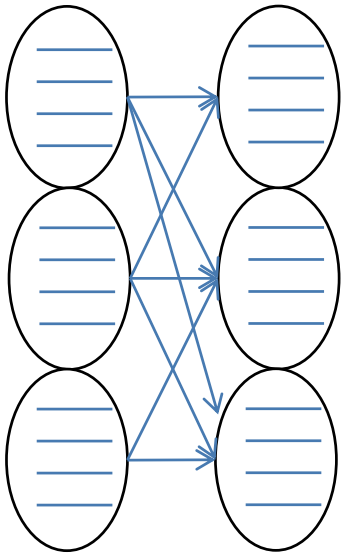
Text Alignment

Seeding	Match Merging	Extraction Filtering



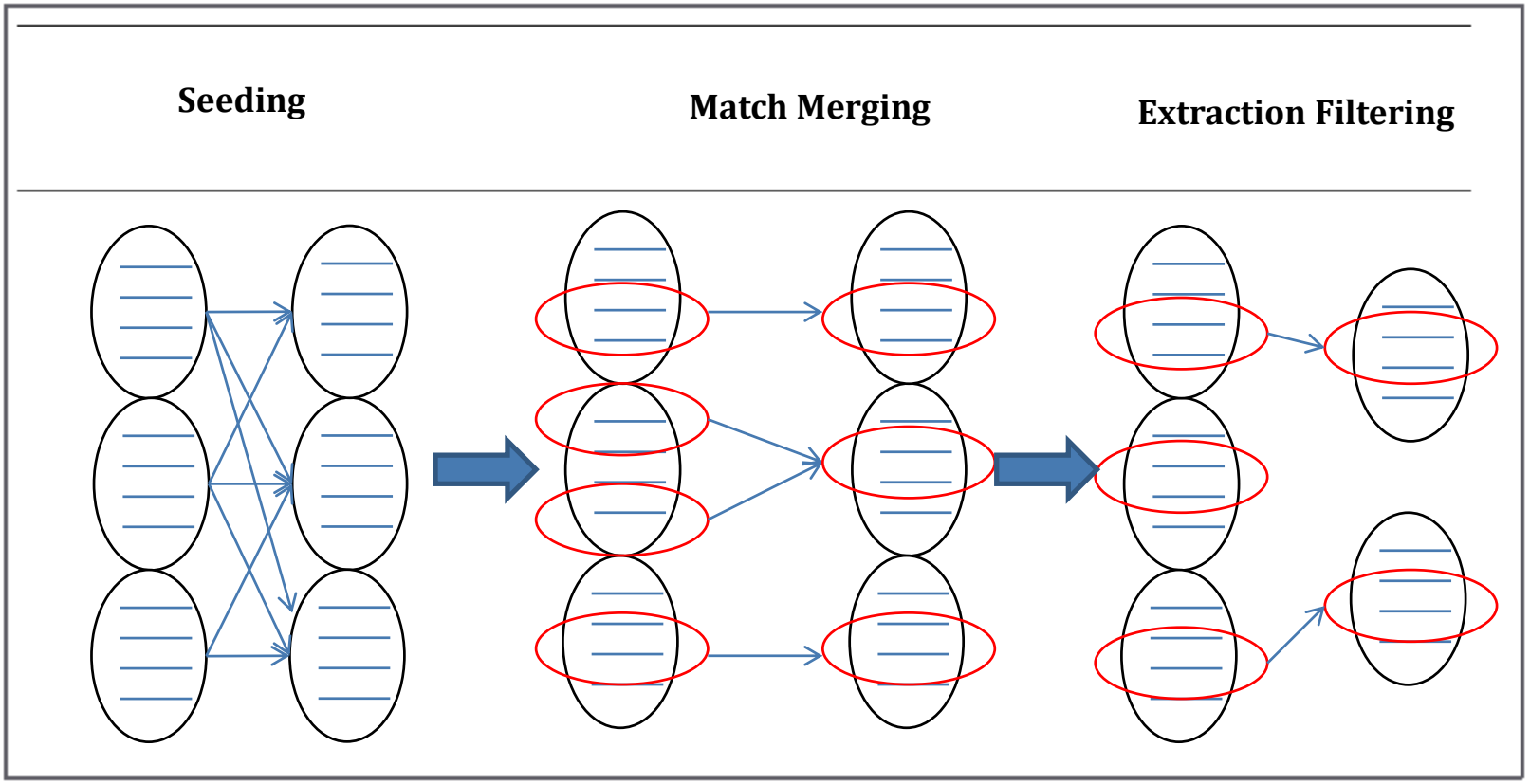
Text Alignment

Seeding **Match Merging** **Extraction Filtering**



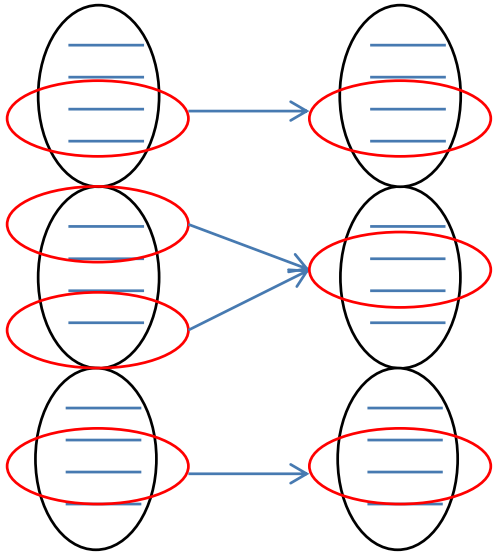
$$Sim(S,R) = \cos\theta = \frac{\sum_{k=1}^{n} W_{Sk} * W_{Rk}}{\sqrt{(\sum_{k=1}^{n} W_{Sk}^{2})(\sum_{k=1}^{n} W_{Rk}^{2})}} > t1$$

Text Alignment

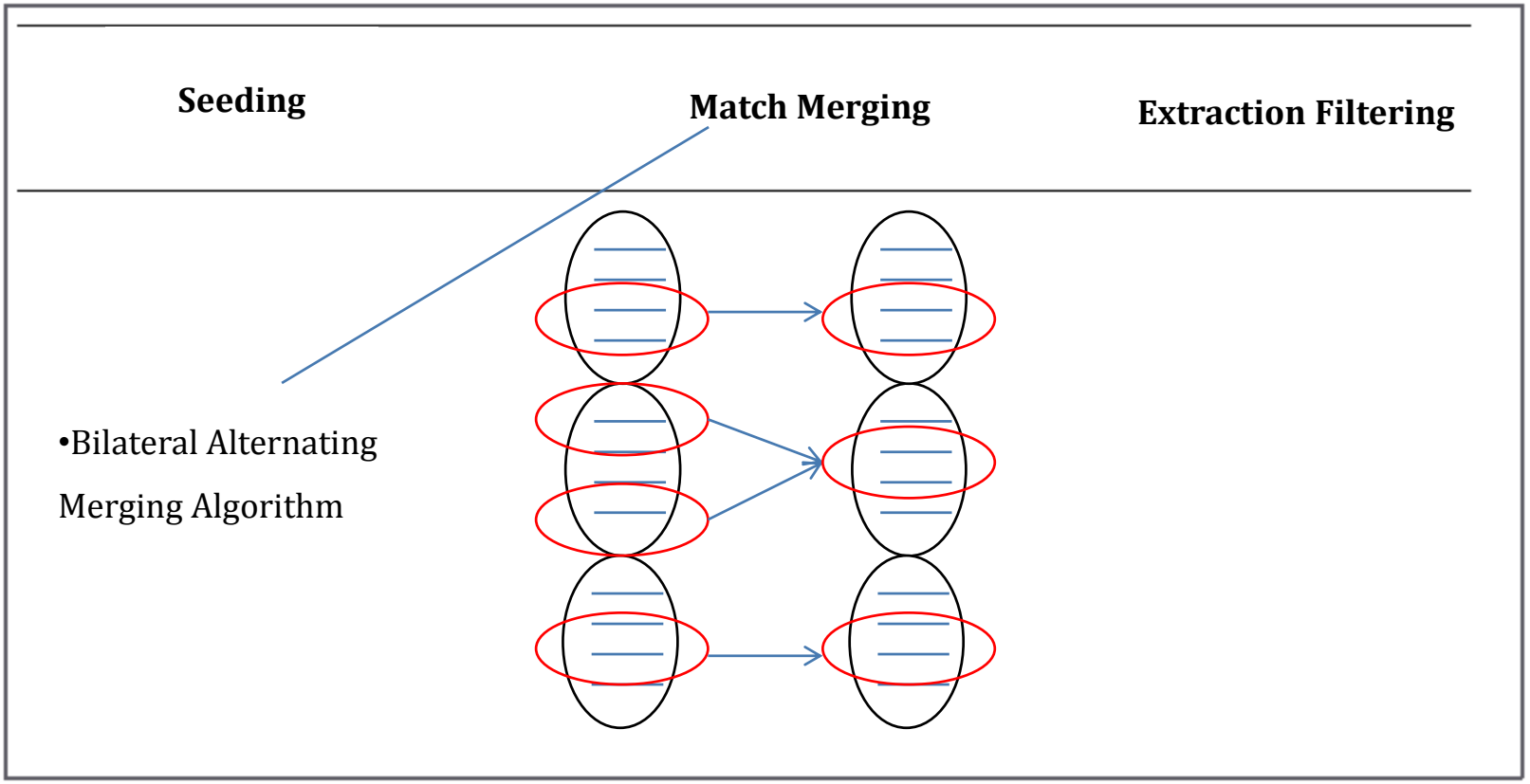


Text Alignment

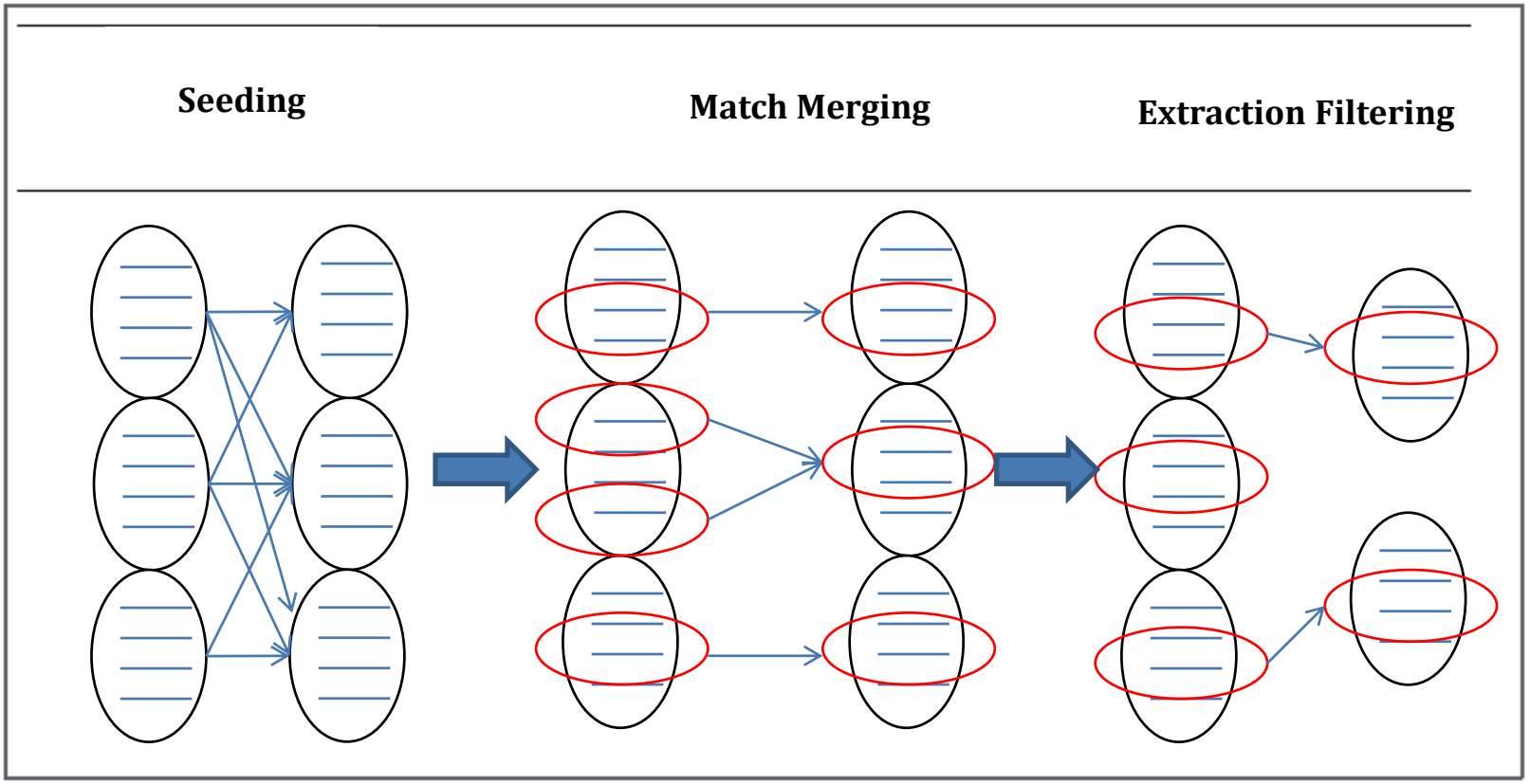
Seeding	Match Merging	Extraction Filtering



Text Alignment

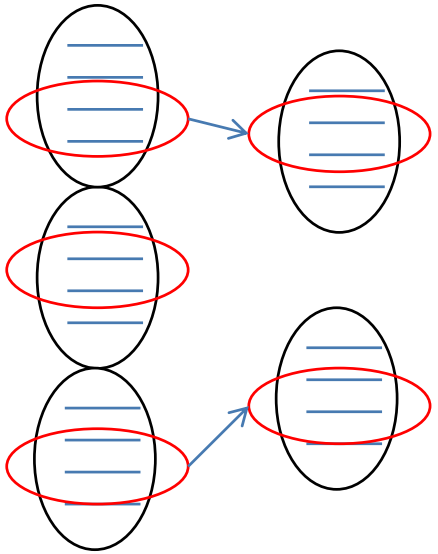


Text Alignment



Text Alignment

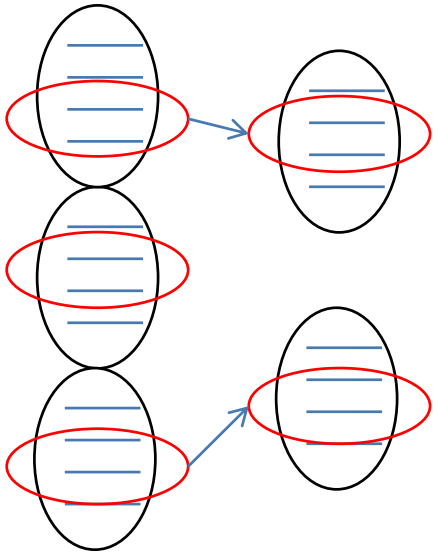
Seeding	Match Merging	Extraction Filtering



Text Alignment

Seeding **Match Merging** **Extraction Filtering**

$$J(I_S, I_R) = \frac{2 * \sum_{t \in I_S \cap I_R} Min(N_{I_S}(t), N_{I_R}(t))}{|I_S| + |I_R|}$$



Performance on the PAN2012 test corpus

Table 3: Overall evaluation results for the final test corpus

	Submission	PlagDet	Recall	Precision	Granularity	Runtime
1	torrejon13	0.82220	0.76190	0.89484	1.00141	72508 ms
2	kong13	0.81896	0.81344	0.82859	1.00336	364064 ms
3	suchomel13	0.74482	0.76593	0.72514	1.00028	1681835 ms
4	saremi13	0.69913	0.77123	0.86509	1.24450	26762432 ms
5	shrestha13	0.69551	0.73814	0.87461	1.22084	41070802 ms
6	palkovskii13	0.61523	0.53561	0.81699	1.07295	390020 ms
7	nourian13	0.57716	0.43381	0.94707	1.04343	2406514 ms
-	baseline	0.42191	0.34223	0.92939	1.27473	1831812 ms
8	gillam13	0.40059	0.25890	0.88487	1.00000	1278518 ms
9	jayapal13	0.27081	0.38187	0.87901	2.90698	289229 ms

Performance on the PAN2012 test corpus

Table 4: Results for the 02-no-obfuscation sub-corpus

	Submission	Dataset	PlagDet	Recall	Precision	Granularity
-	baseline	02	0.93404	0.99960	0.88741	1.00912
1	torrejon13	02	0.92586	0.95256	0.90060	1.00000
2	nourian13	02	0.90136	0.87626	0.92921	1.00092
3	shrestha13	02	0.89369	0.99902	0.80933	1.00083
4	gillam13	02	0.85884	0.83788	0.88088	1.00000
5	saremi13	02	0.84963	0.95416	0.82676	1.06007
6	kong13	02	0.82740	0.90682	0.76077	1.00000
7	palkovskii13	02	0.82431	0.85048	0.79971	1.00000
8	suchomel13	02	0.81761	0.99637	0.69323	1.00000
9	jayapal13	02	0.38780	0.86040	0.91989	3.90017

Performance on the PAN2012 test corpus

Table 5: Results for the 03-random-obfuscation

	Submission	Dataset	PlagDet	Recall	Precision	Granularity
1	kong13	03	0.82281	0.78682	0.86224	1.00000
2	suchomel13	03	0.75276	0.68886	0.82973	1.00000
3	torrejon13	03	0.74711	0.63370	0.90996	1.00000
4	shrestha13	03	0.66714	0.71461	0.92335	1.30962
5	saremi13	03	0.65668	0.68877	0.91810	1.29511
6	palkovskii13	03	0.49959	0.36420	0.93137	1.06785
7	nourian13	03	0.35076	0.23609	0.96274	1.11558
8	jayapal13	03	0.18148	0.18182	0.92314	2.19096
-	baseline	03	0.07123	0.04181	0.98101	1.18239
9	gillam13	03	0.04191	0.02142	0.95968	1.00000

Performance on the PAN2012 test corpus

Table 6: Results for the 04-translation-obfuscation

	Submission	Dataset	PlagDet	Recall	Precision	Granularity
1	kong13	04	0.85181	0.84626	0.85744	1.00000
2	torrejon13	04	0.85113	0.81124	0.89514	1.00000
3	saremi13	04	0.70903	0.80473	0.84819	1.24204
4	suchomel13	04	0.67544	0.66621	0.68494	1.00000
5	shrestha13	04	0.62719	0.63618	0.88008	1.26184
6	palkovskii13	04	0.60694	0.49667	0.82207	1.02825
7	nourian13	04	0.43864	0.28568	0.95856	1.00485
8	jayapal13	04	0.18181	0.19411	0.85653	2.34218
-	baseline	04	0.10630	0.08804	0.97825	1.86726
9	gillam13	04	0.01224	0.00616	0.97273	1.00000

Performance on the PAN2012 test corpus

Table 7: Evaluation results for the 05-summary-obfuscation

	Submission	Dataset	PlagDet	Recall	Precision	Granularity
1	suchomel13	05	0.61011	0.56296	0.67088	1.00476
2	kong13	05	0.43399	0.30017	0.96384	1.07742
3	torrejon13	05	0.34131	0.21593	0.90750	1.03086
4	shrestha13	05	0.11860	0.09897	0.90455	1.83696
5	nourian13	05	0.11535	0.07622	0.99972	1.34234
6	saremi13	05	0.11116	0.10209	0.94600	2.15556
7	palkovskii13	05	0.09943	0.08082	0.67604	1.73596
8	jayapal13	05	0.05940	0.07236	0.68832	3.60987
-	baseline	05	0.04462	0.03649	0.91147	1.97436
9	gillam13	05	0.00218	0.00109	0.99591	1.00000

Further work

- Use different methods to deal with different plagiarism problems to obtain a better performance
- Query keywords extraction and ranking

Thank you for your attention!

黑龙江工程学院
HEILONGJIANG INSTITUTE OF TECHNOLOGY