Source Retrieval and Text Alignment Corpus Construction for Plagiarism Detection

Notebook for PAN at CLEF 2015

Kong Leilei[1,2], Lu Zhimao[2], Han Yong[1], Qi Haoliang[1], Han Zhongyuan[1,3],

Wang Qibo[1], Hao Zhenyuan[1], Zhang Jing[1]

1Heilongjiang Institute of Technology, China
2Harbin Engineering University, China
3Harbin Institute of Technology, China

kongleilei1979@hotmail.com

Abstract. For the task of source retrieval, we focus on the process of Download Filtering. For the process from chunking to search control, we aim at high recall, and for the process of download filtering, we devote to improve precision. A vote-based approach and a classification-based approach are incorporated to filter the searching results to get the plagiarism sources. For the task of text alignment corpus construction, we describe the methods we use to construct the Chinese plagiarism cases. At last, we report the statistics of text alignment dataset submissions.

1 Source Retrieval in Plagiarism Detection

Source retrieval is a core task of plagiarism detection. The source retrieval task can be described as: given a suspicious document and a web search engine, the task is to retrieve the source documents from which text has been reused [1]. The research of plagiarism source retrieval algorithm is a valuable work which is more than just for the development of plagiarism software. Finding plagiarism sources from tens of millions of webpages is a challenging job for all of researchers.

PAN organized Source Retrieval Evaluation from 2012. Potthast et al. summarized the general process by analyzing the algorithms committed by contestants [1], shown in Figure 1.

Followed the above process, we focus on download filtering process in this year's evaluation. For the process from chunking to search control, we aim at high recall, and for the process of the download filtering, we devote to improve precision.

Given a fixed suspicious text chunking method and a fixed downloading number of retrieval results, we find there is no outstanding difference on evaluation measure recall if we retain enough retrieval results (for example, 100 retrieval results for a query) without considering precision. So, we decide to achieve a high recall by

submitting as many queries as possible to the search engine and retaining as many retrieval results as possible.

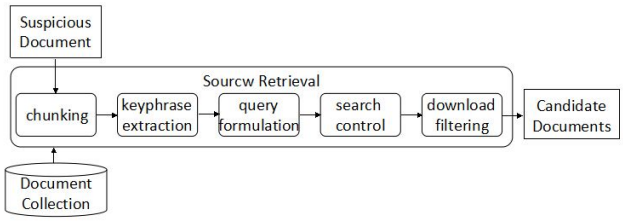

Fig. 1. a general process of plagiarism source retrieval

Chunking. Firstly, the suspicious texts are partitioned into segments that are made up of only one sentence. Especially, it is found that the suspicious documents generally contain some headings. If there are empty lines in front and one behind and the word number of the line is less than 10, the current line are previewed as headings. We try to use only headings as queries to retrieve the plagiarism sources when we did not retrieve any sources on some suspicious documents, but the sources are still not discovered by using these headings. So the headings are merged into the sentence which were adjacent to them.

Keyphrase Extracting. After getting all sentences, each word in each paragraph is tagged using the Stanford POS Tagger[2] and only nouns and verbs are considered as query keyphrase.

Query Formulation. Queries are constructed by extracting each sentence of k keywords, where k = 10. If the number of nouns and verbs in one sentence is more than 10, we retain only top 10 with high term frequencies. And if the number is less than 10, all nouns and verbs are regarded as the query. Then these queries are submitted to ChatNoir search engine[3] to retrieve plagiarism sources.

Search Control. Since each query is generated by only one sentence, it represents the topic which the sentence tries to express, and maybe strayed from the subject which the plagiarism segment which the sentence come from. The result is that many positive plagiarism sources are ranked below. Therefore, for each query, we keep the top 100 results. This tactic make us own a higher recall before download filtering.

Download Filtering. There can be no argument that the number of retrieval results has a large effect on the performance, and increasing the number will lead to an increase in recall and a decrease in precision. In the steps of keywords extraction, except for the content of suspicious document and its text chunk, we have very little information. Submitting more queries may be the best choice without considering the retrieval cost. But after retrieving, we can get abundant information including various similarity scores between query and document, the length of document, the length of words, sentences and characters of document, the snippet(the length of snippet we requested is 500 characters), and so on. By exploiting the retrieval results and the meta-data returned by ChatNoir API, we design a two-step download filtering algorithm.

As we known, the evaluation algorithm of source retrieval computes recall, precision and fMeasure by using the downloading documents, so before implementing our download filtering algorithm, we decide to filter some retrieval results firstly. We suppose that the queries can retrieve the same plagiarism sources if they come from the same plagiarism segment of suspicious document. Then, for one suspicious document, the same retrieval results will occur many times. The underlying assumption is that more possible plagiarism sources are likely to receive more search results voting from different queries of suspicious document. So, we use a simple vote algorithm to assign a weight to each document of the retrieval results set. If a document is retrieved by a query, the weight of the document will add 1. We have also tried the weighted vote approach by giving the document which ranking at the front more higher weight, but it do not perform better than the simple vote approach.

After implementing vote algorithm, the results of vote are regarded as the candidate plagiarism sources. If the size of result list is less than 20, we choose the top 50 results according to the top voting results as the candidates.

Table 1 shows the performance of source retrieval only using vote approach to filter the retrieval results, which is called Han15 by PAN in [4]. Experiments were performed on the train dataset pan14-source-retrieval-training-corpus-2014-12-01 of source retrieval which contains 98 suspicious documents. The numbers in the column headers means the count of vote, and the row headers are the evaluation measures of source retrieval. We choose vote 8 when we submit our source retrieval software to PAN.

	vote5	vote6	vote 7	vote 8	vote 9	vote 10	vote 12	vote 15
fMeasure	0.2976	0.3081	0.3161	0.3167	0.3177	0.3127	0.3159	0.3129
Recall	0.5109	0.4931	0.4843	0.4795	0.4721	0.4710	0.4608	0.4622
Precision	0.2627	0.2755	0.2820	0.2832	0.2872	0.2861	0.2856	0.2807
Queries	202.27	202.27	202.27	202.27	202.27	202.27	202.27	202.27
Downloads	58.3673	53.5918	50.6429	53.6429	51.9490	61.2449	46.2347	46.2143

Table 1. Results of only using vote approach

The data in above table 1 is evaluated by our own evaluation detector which is designed according to Ref. [1]. But we only implemented the former two-way approach to determine true positive detections because we did not know which algorithm was used to extract plagiarism passages' set which were applied to compute the containment relationship.

In the past year's evaluation, Williams et al.[5] proposed a filtering approach which viewed the filtering process of candidate plagiarism sources as a classification problem. A supervised learning method based on LDA(Linear Discriminant Analysis) was used to learn a classification model to decide which candidate plagiarism source was the positive detections before downloading them. This year, we followed their idea and added four new features. They are Document-snippet word 2-gram, 3-gram, 4-gram and 8gram intersection. The set of word 2, 3, 4 and 8 grams from the suspicious document and snippet are extracted separately, and the common n-grams are computed. We chose SVM as our classification model. The open tools SVM

light(http://www.cs.cornell.edu/People/tj/svm_light/) is used as our classifier. We only trained the parameter c in training set which was constructed according to Ref. [6]. After voting, all the results which are positive case judged by classifier are downloaded. The vote strategy follows Han15. This approach based on vote and classification is called Kong15 by PAN in [4].

Using the Source Oracle, we filtered our results. The final log file reported the filtered results of source retrieval. Table 2 shows the results by using the classification tactics.

	vote5	vote6	vote 7	vote 8	vote 9	vote 10	vote 12	vote 15
F1	0.4528	0.4536	0.4554	0.4541	0.4531	0.4522	0.4528	0.4536
Recall	0.5022	0.4826	0.4744	0.4703	0.4629	0.4618	0.5022	0.4826
Precision	0.5318	0.5363	0.5436	0.5451	0.5459	0.5453	0.5318	0.5363
Queries	202.27	202.27	202.27	202.27	202.27	202.27	202.27	202.27
downloads	61.2449	46.2347	46.2143	58.3673	53.5918	50.6429	61.2449	46.2347

Table 2. Results of combining vote and classification approach

Our two evaluation results reported by PAN are shown in Table 3.

	Kong15	Han15
fMeasure	0.38487	0.36192
Recall	0.42337	0.31769
Precision	0.45499	0.54954
Downloads	38.3	11.8
DownloadUntilFirstDetection	3.5	1.7
queries	195.1	194.5
QueriesUnitilFirstDetection	197.5	202.0

Table 3. Results of PAN@CLEF2015 Source Retrieval subtask

2 Text Alignment Corpus Construction

For the task of text alignment corpus construction, we submit a corpus which contains 7 plagiarism cases. The plagiarism cases are constructed by using real plagiarism.

Firstly, we recruited 10 volunteers to write a paper according to a topic we proposed. We choose 7 of 10 to submit our corpus. Table 4 lists the 7 topic.

For each essay, we request ten thousand Chinese characters at least. The volunteers retrieved the related contents on the subject by using the specified search engine and wrote the paper. Especially, the Baidu is used to search engine. The number of sources has not been not limited.

Then papers were submitted to a famous Chinese plagiarism detection software which are used in many Chinese colleges and universities. This plagiarism detection software uses the fingerprint technology to detect the plagiarism. Next, the volunteers modified the contents which were detected by this software. The modification tactics include: adjusting the words' order, replacing the words and paraphrasing

modification. But no matter what kinds of modifying tactics they adopted, they must ensure that the paper after revising is readable and consistent with the original paper's meaning. Lastly, the modified papers were submitted to the plagiarism detection software until the software could no longer detect any plagiarism. The modified papers were submitted to PAN as the text alignment corpus.

Suspicious Document	Topic
suspicious-document00000	Campus Second-hand Book Trade
suspicious-document00001	Online Examination
suspicious-document00002	Online Examination
suspicious-document00003	Second-hand Car Trade
suspicious-document00004	Automobile 4S Shop
suspicious-document00005	Multimedia Material Management Library
suspicious-document00006	Driving license exam
suspicious-document00007	Supermarket Management System

Table 4. Topics of text alignment corpus construction

The statistics of the corpus is shown in table 5.

Corpus characteristic	Total							
	00000	00001	00002	00003	00004	00005	00006	00007
Average lengths of suspicious documents	33688	27211	28881	46167	35733	21858	23251	52531
Average lengths of plagiarism cases	188	330	543	577	1288	1066	827	687
Number of plagiarism cases per document	4	1	12	3	9	4	5	13
Jaccard coefficient	0.4665	0.4215	0.6856	0.5439	0.7044	0.3252	0.6913	0.4705

Table 5. Statistics of corpus characteristic by Chinese characters

We peer-review pan15 text alignment dataset submissions[7] and the statistics of corpus are shown in table 6.

Corpus characteristic	Total(alvi15-English)				Total(khoshnavataher15-persian)		
	01	02	03	04	01	02	03
Number of suspicious document	15	25	25	25	400	117	232
Number of source document	19	25	25	25	489	118	243
Average length of suspicious documents	13577	7134	7388	666	4366	9879	8968
Average length of source documents	13619	9402	6981	8730	4952	4990	5906
Average lengths of plagiarism cases	-	523	393	447	-	901	925
Number of plagiarism cases	-	25	25	25	-	129	282
Jaccard coefficient	-	0.2431	0.5193	0.2057	-	09453	0.7101

Table 6.1. Statistics of text alignment dataset submissions (alvi15 and khoshnavataher15)

Corpus characteristic	Total(khoshnavataher15-English)				Total(kong15-Chinese)
	01	02	03	04	01
Number of suspicious document	199	54	391	39	4
Number of source document	448	132	1117	39	5
Average length of suspicious documents	19019	21788	23290	25136	33986
Average length of source documents	16171	19029	18743	27477	21319
Average lengths of plagiarism cases	-	406	436	486	569
Number of plagiarism cases	-	143	1207	39	20
Jaccard coefficient of plagiarism cases	-	0.6815	0.3416	0.3080	0.60738

Table 6.2. Statistics of text alignment dataset submissions (khoshnavataher15 and kong15)

Corpus characteristic	Total(najib15-English)					Total(khoshnavataher15-English-persian)	
	01	02	03	04	05	01	02
Number of suspicious document	125	21	76	7	19	2742	2728
Number of source document	125	21	76	7	19	3839	4571
Average length of suspicious documents	6344	8579	6689	6375	5871	4308	6052
Average length of source documents	8178	8217	7353	7386	7794	18494	18744
Average lengths of plagiarism cases	-	699	463	834	342	-	299
Number of plagiarism cases	-	21	76	7	19	-	5606
Jaccard coefficient of plagiarism cases	-	0.4698	0.3221	0.3341	0.3611	-	0.0033

Table 6.3. Statistics of text alignment dataset submissions (najib15 and khoshnavataher15)

Corpus characteristic	Total(palkovskii15-English)					Total(cheema15-English)	
	01	02	03	04	05	01	02
Number of suspicious document	138	153	146	146	592	115	135
Number of source document	500	478	482	480	223	115	135
Average length of suspicious documents	5399	16438	14074	17299	6546	6448	6581
Average length of source documents	3926	4187	4274	4823	5138	2054	2371
Average lengths of plagiarism cases	-	564	434	511	627	-	344
Number of plagiarism cases	-	624	626	618	108	-	135
Jaccard coefficient of plagiarism cases	-	0.0298	0.0166	0.0144	0.0073	-	0.00694

Table 6.4. Statistics of text alignment dataset submissions (palkovskii15 and cheema15)

Table 6. Statistics of text alignment dataset submissions

Acknowledgments This work is supported by Youth National Social Science Fund of

China (No. 14CTQ032), National Natural Science Foundation of China(No. 61272384), and Heilongjiang Province Educational Committee Science Foundation(No. 12541649, No. 12541677).
Remark This work was done in Heilongjiang Institute of Technology.

Reference

1. Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, Benno Stein: Overview of the 6th International Competition on Plagiarism Detection. CLEF (Working Notes) 2014: 845-876.
2. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03. vol. 1, pp. 173–180 (May 2003)
3. Martin Potthast, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. ChatNoir: A Search Engine for the ClueWeb09 Corpus. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12), pages 1004, August 2012. ACM. ISBN 978-1-4503-1472-5.
4. Matthias Hagen, Martin Potthast, and Benno Stein. Source Retrieval for Plagiarism Detection from Large Web Corpora: Recent Approaches. In Working Notes Papers of the CLEF 2015 Evaluation Labs, CEUR Workshop Proceedings, September 2015. CLEF and CEUR-WS.org. ISSN 1613-0073.
5. Williams, K., Chen, H.H., Giles, C.: Supervised Ranking for Plagiarism Source Retrieval—Notebook for PAN at CLEF 2014. 15-18 September, Sheffield, UK. CEUR Workshop Proceedings, CEUR-WS.org (2014), http://www.clef-initiative.eu/publication/working-notes.
6. Williams K, Chen H H, Giles C L. Classifying and ranking search engine results as potential sources of plagiarism[C]//Proceedings of the 2014 ACM symposium on Document engineering. ACM, 2014: 97-106.
7. Martin Potthast, Matthias Hagen, Steve Göring, Paolo Rosso, and Benno Stein. Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment. In Working Notes Papers of the CLEF 2015 Evaluation Labs, CEUR Workshop Proceedings, September 2015. CLEF and CEUR-WS.org. ISSN 1613-0073.