Efficient Paragraph based Chunking and Download Filtering for Plagiarism Source Retrieval

Notebook for PAN at CLEF 2015

Riya Ravi N, Deepa Gupta
Amrita Vishwa Vidyapeetham
Amrita School of Engineering, Bangalore Campus, India
riya.sanjesh@gmail.com, g_deepa@blr.amrita.edu

Abstract. This paper describes the approach of the system that we built as part of the participation in 'PAN 2015 Source Retrieval' task. Chunking of documents based on paragraphs and efficient download filtering improved the overall performance of the system. Source Retrieval is an important task of a Plagiarism Detection system

Keywords: API Search Engine · ClueWeb09 corpus · External Plagiarism Detection System · PAN · POS Tagging · Source Retrieval · TF-IDF.

1 Introduction

"Plagiarism is an act or instance of using or closely imitating the language and thoughts of another author without authorization and there presentation of that author's work as one's own, as by not crediting the original author" [1]. Over past several years, a number of researchers have been working on Plagiarism Detection systems to make them more efficient and fast. Evaluation Labs such as the one conducted by PAN [2] for uncovering plagiarism, every year, encourages the researchers to aim higher and come up with better systems.

External Plagiarism Detection task in the PAN evaluation lab is divided into two subtasks – Source Retrieval and Text Alignment. Source Retrieval task involves in retrieving the source documents from which the given suspicious document is plagiarized. Text Alignment task on the other hand, involves in identifying the actual plagiarized portions of the given suspicious document along with the source of the plagiarism. In this paper we concentrate on the Source Retrieval task. Keywords are first extracted from the suspicious documents, provided as input to the system. Queries are then formulated from these keywords for submission to the Search Engine which searches the ClueWeb09 corpus [3] for candidate plagiarism source documents [4]. In this source retrieval subtask we have used the ChatNoir search engine [5]. The other search engine available for the search is the Lemur Indri search engine. The resultant URLs are first filtered by multiple means and then passed on to the Download API to download the source document from the ClueWeb09 corpus.

2 Source Retrieval Sub Task

As part of the source retrieval sub task, suspicious documents are made available and it is expected to build a source retrieval system which identifies and retrieves all the source documents from which the text of the suspicious documents has been reused while minimizing the retrieval costs. Each suspicious document provided is based on a topic and are plagiarized from the documents available in the ClueWeb09 corpus. Two search engines are made available to search the documents in this corpus – ChatNoir and Lemur Indri. To make it easier for the participants of the task, a common search API is provided by PAN to access these search engines [4]. ChatNoir search engine is fast and is good in searching keywords in the documents. Lemur Indri search engine provides the facility to search for phrases but the search is slower than the ChatNoir search engine. These search engines along with the URLs of the source document also provides many facets of the search result, such as readability, word count, page rank, BM25 values etc. Participants may use any of these facets to filter out the URLs before downloading the source documents. This search API expects the keywords/keyphrases in the form of queries which are passed on to the corresponding search engine. Participants are expected to form these queries from the keywords or keyphrases extracted from the suspicious document.

To facilitate the downloading of the documents PAN provides a download API which retrieves a source document from the ClueWeb09 corpus given the URL of that document. Along with this the download API also provides an 'oracle' feature which identifies if the requested document is the plagiarism source of the given suspicious document [4, 6].

3 Text Alignment Sub Task

Text Alignment is the second sub task of the Plagiarism Detection task in PAN evaluation lab and it follows the Source Retrieval sub task. The expectation here is to identify plagiarized passages and the corresponding source passages. A set of suspicious documents along with their identified source documents is provided. The key is the ability to identify obfuscated passages. Not all reuse is direct copy and paste but obfuscated, which means there would be very little lexical similarity between the source and plagiarized passages in certain cases.

The performance of the plagiarism detection system is measured based on plagdet score, precision, recall and granularity, along with the cross year evaluation comparison [4, 6].

4 Proposed Algorithm

The algorithm used for developing the proposed system is listed below.

```
Algorithm 1.
Input: Set of Suspicious documents
Output: Set of source documents and their URLs
Begin
  Repeat below steps for each suspicious document $D_{susp}$
    Chunking:
    Divide $D_{susp}$ into paragraph chunks
    Keyword Extraction:
    PoS tag each paragraph and extract Nouns, Adjectives and Verbs
    Find the TF-IDF values of these keywords
    Sort these keywords according to their TF-IDF values
    Select the top n keywords from this sorted list
    Query Formulation:
    Form two queries with max n/2 keywords in each
    Filter out duplicate queries
    Download
    Repeat below steps for each Query
      Call ChatNoir Search API passing the query
      If the resultant URL is already processed skip it
      Call ChatNoir Search API for the snippet by passing the URL
      Match the snippet and the original paragraph
      If the cosine similarity is less than threshold($\Theta$) then skip the URL
      Download the source document by calling the Download API
End
```

5 Proposed System

5.1 Document Chunking

It is difficult to process a big document as a whole and that's why chunking of the document is important. The chunks are considered as one unit and forms the starting point of the search and download of the plagiarism sources. In our approach we chunk the document into paragraphs. The paragraphs are identified by one or more blank lines between them. The thinking behind keeping the paragraphs as chunks instead of fixed length words or sentences is that the paragraphs form a logical separation of the flow of the document and it is more common to copy a paragraph rather than copy few words or sentences.

5.2 Keyword Extraction

Keywords, as the name suggests are the most important words of a text and they should be able to uniquely identify that text. In our approach, we first process the suspicious document and assign TF-IDF values to the words of the text and then PoS (Parts-of-Speech) tag using the Stanford PoS tagger [7] to extract the nouns, verbs and adjectives. After extracting these words we sort them according to their TF-IDF values and pick the top n keywords. All the chunks of a document are considered while calculating the IDF values.

5.3 Query Formulation and Search Control

Set of extracted keywords are divided into two parts to form two queries. Duplicate queries are filtered out. Remaining queries are passed onto the search API for document search. The top ranked result is considered for further processing.

5.4 Download Filtering

The URLs returned by the search result are passed onto the snippet matching process. For this we call the ChatNoir search API and ask for a 500 character snippet. The returned snippet is matched with the corresponding paragraph text. We calculate the cosine similarity of the snippet and the paragraph text and if the similarity score is more than threshold (Θ), then the URL is passed onto the download API for downloading the document from ClueWeb09 corpus.

6 Dataset and Result Analysis

6.1 Data Set

Data set [8] involved is a set of 99 suspicious documents plagiarized from the various documents available in the ClueWeb09 corpus. ClueWeb09 corpus [3] consists of 1 billion documents in ten languages and is a good representation of the web. The total size of the corpus is around 25 TB in its uncompressed form. It is one of the widely used corpus by researchers.

6.2 Results

Before submitting the system for the PAN source retrieval sub task, it was tuned by running the system on the training data sets provided by PAN. After tuning the system, the value for Θ was fixed to 0.4 and 'n' to 20. This meant the system generated a maximum of two queries per chunk as the underlying search engine – ChatNoir, supports queries with maximum length of 10. The system was later submitted for the PAN 2015 Source Retrieval sub task [9] and the results on the Data Set is shown in Table 1. The results of the proposed system is shown under the user name of 'sanjesh15'. These results show that the system performed well in reducing the load of the system by minimizing the number of downloads while maximizing the value of the F Measure. The system also outperformed other systems in the downloads before the first detection. Along with this the system's runtime was in the lower side.

Table 1. Results of the run of the proposed system on the PAN 2015 Test data set

User	Downloads	Downloads before 1st detection	F Measure	No Detection	Precision	Queries	Queries before 1st Detection	Recall	Runtime
suchomel15	331.3	40	0.09767	5	0.06498	43.8	**4.1**	0.4135	175:13:52
sanjesh15	**8.5**	**1.6**	**0.42726**	8	**0.61303**	90.3	17.5	0.38739	9:17:20
rafiei15	183.3	24.9	0.1154	**1**	0.07539	**43.5**	5.6	0.41381	**8:32:37**
han15	11.8	1.7	0.36192	12	0.54954	194.5	202	0.31769	20:43:02
kong15	38.3	3.5	0.38487	3	0.45499	195.1	197.5	**0.42337**	17:56:55

7 Conclusion

The proposed system is successful in reducing the workload of the system by dropping the number of downloads required to detect the plagiarism. This the system achieved even while maximizing the F Measure value. The system shows not so good value for the recall. It is primarily due to the fact that the system processes only the top ranked query result and ignores the others. As a future scope of work, other query results could also be used and fed to the snippet matching mechanism so that some of the sources are not missed.

8 References

1. http://dictionary.reference.com
2. http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/index.html
3. http://lemurproject.org/clueweb09/
4. Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., and Stein, B.: Overview of the 4th International Competition on Plagiarism Detection. In: Forner, P., Karlgren, J. and Womser-Hacker, C., editors, Working Notes Papers of the CLEF 2012 Evaluation Labs, September 2012. ISBN 978-88-904810-3-1.
5. Potthast, M., Hagen, M., Stein, B., Graßegger, J., 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.
6. Martin Potthast, Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 5th International Competition on Plagiarism Detection. In: Forner, P., Navigli, R., Tufis, D. (eds.) Working Notes Papers of the CLEF 2013 Evaluation Labs (Sep 2013).
7. http://nlp.stanford.edu/software/tagger.shtml
8. Martin Potthast, Matthias Hagen, Michael Völske, and Benno Stein. Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In

Pascale Fung and Massimo Poesio, editors, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 13), pages 1212–1221. ACL, August 2013

9. Martin Potthast, Matthias Hagen, 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.