Source Retrieval Plagiarism Detection based on Noun Phrase and Keyword Phrase Extraction

Notebook for PAN at CLEF 2015

Javad Rafiei, Salar Mohtaj, Vahid Zarrabi, Habibollah Asghari

ICT Research Institute,
Academic Center for Education, Culture and Reseach (ACECR), Iran

{javad.rafiei, salar.mohtaj, vahid.zarrabi, habib.asghari}
@ ictrc.ir

Abstract. This paper describes an approach for source retrieval task of PAN 2015 competition. We apply two methods to extract important terms, namely weighted noun phrases and keyword phrases which are extracted from long sentences in terms of word count. Queries are constructed from top marked sentences. The prepared system tries to gather a complete dataset of downloaded sources and employ it in query filtering operations. The ChatNoir search API is used for submitted queries. Each query is split into two sub-queries and the system extract one snippet for each of sub-queries and exploits them in downloading operation. The evaluation results show high scores for three measures: recall, total queries number and no detection.

Keywords: Plagiarism Detection, Source retrieval, Keyword phrase extraction, Noun phrase extraction

1 Introduction

The advent and rapid development of the World Wide Web facilitate public access to digital information, so that everybody can easily read and alter the content of web pages for personal use. There can be considered two aspect of information generation. First, using the existing information with their references as input and extending them with new innovation as output, and second, taking information and altering the content while maintaining the concept without any reference to source which is called plagiarism. In simple terms, plagiarism is the unreferenced use of other's ideas or text. Plagiarism detection in PAN [1] is divided into source retrieval and text alignment subtasks. The former task uses two search engines [2, 3] to retrieve all plagiarized sources for input suspicious documents. In this paper we focus on source retrieval task at PAN 2014 and present a simple solution for keyword extraction and query

building with regard to high quality of query's terms and minimum use of web search API. We have used ChatNoir API to retrieve candidate source documents.

The rest of this paper is structured as follows: Section 2 describes our approach that includes document segmentation, keyword extraction methods, query formulation and query and document filtering. Section 3 presents the results of our method and a discussion on the results. Conclusion and steps for future works will be explained in section 4.

2 Our Methodology

Our approach has been divided into five steps as follows:

- Suspicious Document Chunking,
- Noun phrase and keyword phrase Extraction,
- Query Formulation,
- Search Control,
- Document Filtering and Downloading

These steps are equal to those described in [1]. Before these main steps, raw suspicious documents are passed through a preprocessing block that includes stop words removal and punctuation deletion.

2.1 Suspicious Document Chunking

After the preprocessing step, the documents are prepared for suspicious document chunking. Each document is segmented into some parts called chunks. These chunks are separately used for keyword phrase and noun phrase extraction and also query construction. Therefore, their length should be long enough to extract meaningful queries. On the other hand, these chunks may contain unknown numbers of plagiarism cases from source documents. Suspicious documents are divided into chunks of 500 words length and then each chunk is tokenized into individual sentences. As a result, we have some sentences that are used to extract appropriate keywords.

2.2 Noun phrase and keyword phrase Extraction

This step has the main role in source retrieval task. Extracting appropriate keywords help us to efficiently perform the next steps. There are many previous studies that have tried to extract the keywords by investigating the contents [4, 5]. We have used two types of keyword extraction in our approach: 1) Keyword phrase extraction. 2) Noun phrase extraction.

The input to this step is a series of sentences that are extracted from a specific chunk at the chunking step. Before starting extraction process, sentences with low information content are discarded: We rank the input sentences based on their length and the number of nouns, and then discard the lower 20% of the sentences in the rank-

ing. The resulting sentences are long enough and have rich content for keyword extraction.

In this stage, two different types of tf.idf weighting scheme are used for rating important words in the sentences:

- Type #1 (tf.idf1): In this type, Tf = term frequency in the chunk, and Idf = inverse term frequency in the whole suspicious document
- Type #2 (tf.idf2): in this type, Tf = term frequency in the chunk, and Idf = inverse document frequency in PAN 2011 corpus

Keyword phrase Extraction. Keyword phrases are considered as a collection of keywords with high tf.idf weights in a sentence. The number of keywords that can be used as input to ChatNoir search engine is limited, so the number of query words should be restricted in such a way to maximize the information content. The process of keyword phrase extraction is done in four steps as follows: In the first step, for each type of tf.idf weighting scheme that mentioned above, ten words with highest tf.idf value are selected from the whole chunk. In the second step, the sentences that contain words with high values of tf.idf1 AND tfidf2 are selected. In the next step, among the selected sentences, those that contain words with highest tf.idf1 and tfidf2 are selected for keyword phrase extraction. As a result, three sentences are selected in this step. Finally, in the fourth step, the keywords are extracted from the resulting sentences as follows:

- Nouns with high tf.idf values
- Remaining nouns in the sentence
- Adjective and verbs with high tf.idf1 values

It should be noted that the selection process is done based on the above mentioned priority. When we reach the maximum number of keywords in each stage, then the process would be stopped.

The four steps mentioned above are repeated for each chunk and the extracted keyword phrases are passed to the next step for query building.

Noun phrase Extraction. Noun phrase extraction is accomplished by processing the remaining sentences. The formulation has been deployed based on the formal English noun phrase structure [6]. For each noun phrase, a score is calculated based on tf.idf1 values. This score is the average of tf.idf1 values. Next, the noun phrases are ranked based on their scores. From the three top ranked noun phrases, the top tf.idf2 weighting words are passed to query formulation step.

As a result, the implemented system uses two different scenarios applied to sentences for keyword extraction. After dividing a suspicious document into some chunks, the following scenarios are used based on operations depicted in Table 1.

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

Table 1. Multiple Operations on sentences in keywords 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 tf.idf1 and tf.idf2 values)

The outputs of these scenarios are also passed through a filter that removes some terms with low weight to reach the word count limitation of ChatNoir API. The remaining terms formulate a query and in the next step we can select top weighted sentence for query formulation.

2.3 Query Formulation

For top sentences selected from previous step, the extracted keywords are simply placed next to each other based on their order in sentence and are passed to next step as a query. According to ChatNoir limitations, the threshold for the number of words in each query is limited to 10.

2.4 Search Control

In this step, we filtered the constructed queries based on the possibly previous downloaded documents. The input query is compared against the downloaded documents that are gathered from previous rounds of source retrieval steps. If there is at least one downloaded document that contains at least 60% words of the query, then the query is dropped from passing to the next step. This threshold was achieved based on experiments.

2.5 Document Filtering and Downloading

We have used ChatNoir API for applying the input queries into the search engine. Then 14 top ranked results returned for each query. Input query is divided into two sub-queries and for each of them, one snippet with the length of 500 characters is extracted per returned document. These snippets are combined with one another and make a passage. If the resulted passage contains at least 50% words of the query, then the related document is downloaded and maintained for search control operation. This threshold was achieved based on experiments.

3 Evaluations and Discussion

We have implemented our approach using python programming language and NLTK package for text processing operations [7]. At first, the prepared software for source retrieval was run on training dataset [8] and after getting feedback from the results, the following parameters were optimized:

- Chunk length
- Number of queries in each chunk
- Returned results for each query
- Similarity threshold between a query and resulted snippets
- Similarity threshold between a query and downloaded documents

Then, the software was placed on a dedicated virtual machine and was run on test dataset through TIRA [9]. Table 1 shows achieved results of our software on the PAN 2015 test dataset. The bold cases show highest rank for our software between all participants. According to "No Detection" score, our software has achieved highest rank in this measure. In other words, for only one plagiarized document, the "no true positive detection" was made. However, the number of downloads is relatively high. One of our main objectives was to deploy a method for query building with special keywords to get a high recall. Our algorithm has reached second highest rank in recall score among all participants. In download filtering step, the software gathers a complete set of suitable source documents and uses them with a simple query filtering method. As a result, the number of queries that we use as input to APIs search engine in our software is lowest among other participants. Since that we used a simple approach in download filtering, so we achieved the best rank in software runtime measure among the participants.

Table 2. Performance of our approach on source retrieval 2015

Downloads	F1	No Detection	Precision	Queries	Recall	Runtime
183.3	0.1154	**1**	0.07539	**43.5**	0.41381	**8:32:37**

4 Conclusion and Future Works

In this paper, we have described an approach for source retrieval task of PAN competition. This process has achieved second highest rank in recall and first in "No Detection" score. Because of high detection power of the system, the collected documents cover most of the relevant sources. Extra queries have been filtered using a simple method by making a union between query terms and collected documents. As a result we have achieved the highest rank in terms of 'number of queries'. Moreover, we also achieved the first rank in 'runtime'.

For future works, we will try to decrease the number of downloaded source documents while keeping the complete set of related documents for query filtering.

Acknowledgement

This work has been accomplished in ICT research Institute, ACECR, under the support of Vice Presidency for Science and Technology of Iran - grant No. 1164331. The authors gratefully acknowledge the support of aforementioned organizations. Special thanks go to the members of ITBM research group for their valuable collaboration.

References

1. Potthast, Martin, Tim Gollub, Matthias Hagen, Jan Graßegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, Alberto Barrón-Cedeño, Parth Gupta, Paolo Rosso, and Benno Stein. Overview of the 4th International Competition on Plagiarism Detection. In Working Notes Papers of the CLEF 2012 Evaluation Labs, September 2012. ISBN 978-88-904810-3-1. ISSN 2038-4963.

2. Potthast, Martin, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. "ChatNoir: a search engine for the ClueWeb09 corpus." In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 1004-1004. ACM, 2012.

3. Strohman, Trevor, Donald Metzler, Howard Turtle, and W. Bruce Croft. "Indri: A language model-based search engine for complex queries." In *Proceedings of the International Conference on Intelligent Analysis*, vol. 2, no. 6, pp. 2-6. 2005.

4. Matsuo, Yutaka, and Mitsuru Ishizuka. "Keyword extraction from a single document using word co-occurrence statistical information." *International Journal on Artificial Intelligence Tools* 13, no. 01 (2004): 157-169.

5. Witten, Ian H., Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. "KEA: Practical automatic keyphrase extraction." In *Proceedings of the fourth ACM conference on Digital libraries*, pp. 254-255. ACM, 1999.

6. Abney, Steven Paul. "The English noun phrase in its sentential aspect." PhD diss., Massachusetts Institute of Technology, 1987.

7. Bird, Steven. "NLTK: the natural language toolkit." In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 69-72. Association for Computational Linguistics, 2006.

8. Potthast, Martin, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. "An evaluation framework for plagiarism detection." In *Proceedings of the 23rd international conference on computational linguistics: Posters*, pp. 997-1005. Association for Computational Linguistics, 2010.

9. Gollub, Tim, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. "Recent trends in digital text forensics and its evaluation." In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pp. 282-302. Springer Berlin Heidelberg, 2013.