Improving Synoptic Querying for Source Retrieval

Notebook for PAN at CLEF 2015

Šimon Suchomel and Michal Brandejs

Faculty of Informatics, Masaryk University
{suchomel,brandejs}@fi.muni.cz

Abstract Source retrieval is a part of a plagiarism discovery process, where only a selected set of candidate documents is retrieved from a large corpus of potential source documents and passed for detailed document comparison in order to highlight potential plagiarism. This paper describes a used methodology and the architecture of a source retrieval system, developed for PAN 2015 lab on uncovering plagiarism, authorship and social software misuse. The system is based on our previous systems used at PAN since 2012. The paper also discusses the queries performance and provides explanation for many implementation settings. The proposed methodology achieved the highest recall with usage of the least number of queries among other PAN 2015 softwares during the official test run. The source retrieval subsystem forms an integral part of a modern system for plagiarism discovery.

1 Introduction

In plagiarism detection, the source retrieval is a task for the automated system to retrieve candidate documents from large document collections, which may have served as a pattern for plagiarism, and pass the retrieved documents for further inspection [9]. The further inspection means to evaluate document similarities in detail, which, however, can be done only among documents that are known to the system. The desired similarities are mainly textual, thus the text of suspicious document is usually aligned with each document from the system's document base. If the collection of potential source documents is too large, it is infeasible to calculate detailed similarities among each document pairs from that collection, therefore, only a small selected subset of potential sources to each input suspicious documents is retrieved. The whole collection of possible source documents is an unknown environment for the plagiarism detection system, thus the document retrieval is carried out by utilization of a search engine which is capable of a document retrieval.

This paper describes the key aspects and the main changes in the source retrieval methodology from the system used at PAN 2014 [3], which is described in [10]. The queries performance analysis is also provided.

The task for the PAN lab was to retrieve all plagiarized documents, based on a suspicious document, from the reference document collection by utilizing a web search engine, while minimizing the retrieval costs.

For the source retrieval, based on each suspicious document the prepared queries were passed to search engines according to their type. The synoptic queries were passed

to ChatNoir [4] and the phrasal queries to Indri [8]. Both search engines index ClueWeb09[1] which constituted the corpus of potential source documents. Afterwards, the search engine results were examined and if a similar passage with the suspicious document was found, the result was reported as a candidate document for being a source of plagiarism.

2 Building of Queries

The system prepared several types of queries for each suspicious document, which we divide into two main groups: the whole document keywords-based queries and the paragraph-based queries. For queries construction, a weight w_i was assigned to each term k_i from input document d_j. A term is represented by a word, extracted from the input text using blank spaces as a separator between two words, and cleaned of all punctuation. The weights follow the TF-IDF weighting scheme [6]. As a reference corpus for the weights calculation, an English web-based corpus containing more then 4 billion tokens was used [2]. For each term, it's lemma l_i was also extracted using Python NLTK[3] lemmatizer, which was performed on the basis of extracted sentences. The ambiguous terms were let in the original form. Each document d_j was then represented as (k_i, w_i, l_i) where $i \in [1, |d_j|]$.

2.1 Keywords-based Queries

Keywords-based queries were six terms long. The first query – the pilot query, was extracted from the best scored keywords and passed to both search engines in order to aim for theme related documents and starting the synoptic search. For the purpose of the pilot query, the Indri was set to combine all the query tokens.

For each of those best six keywords, the most frequent in-sentence adjacent words were also extracted. In such a way, collocations were extracted, together with the keywords both three and two terms long. The main difference from the PAN 2014 approach [10] is in the combining of extracted collocations into the queries.

Three term long collocations posed individual phrasal queries[4]. Such short queries were executed via the Indri search engine with the setting of the proximity term number to 1, which denotes the search for exact occurrence. From the subsequent two-token long collocations, an additional 6 terms long queries were created, whilst having all terms in one query unique. They were aimed at the ChatNoir search engine. Finally, the remaining keywords were combined into additional keywords-based queries.

2.2 Paragraph-based Queries

The scored suspicious document was also divided into paragraphs-like chunks using an empty new line (occurring in its plaintext format) as a chunks' separator. From each

[1] http://www.lemurproject.org/clueweb09.php/
[2] http://www.sketchengine.co.uk/documentation/wiki/Corpora/TenTen/enTenTen
[3] http://www.nltk.org/
[4] Phrasal queries posed an exception in a query length.

chunk c_i a single query was prepared. Let $\mathrm{beg}(c_i)$ denote the position of the first word of the chunk c_i in the input file and let $\mathrm{end}(c_i)$ denote the position of the last word of the chunk c_i in the input file. The paragraph based query from c_i comprised from 10 words $k_i \in s_i$, with maximal $\sum_{j=1}^{10} w_{ij}$, where s_i denotes the interval $[\mathrm{beg}(c_i), \mathrm{end}(c_i)]$. Ten tokens is the maximum length of a query for the ChatNoir search engine. The maximum length was chosen in order to produce the most specific query for the given paragraph, which should maximize the probability of retrieving texts containing similar paragraphs. The query was constructed from tokens which might be scattered over the whole chunk, therefore it cannot be used as a phrase query. On the contrary, due to its specificity, it is hardly usable as a synoptic query or a theme-related keywords-based query.

Paragraph-based queries were passed to ChatNoir. The interval s_i was associated to all paragraph-based queries, which denotes the file position of the query. The query is said to characterize the text within its interval.

2.3 Queries Scheduling

In order to acquire maximum information from the top-scored keywords, they were combined into several different types of queries. They appeared in the pilot query and in the collocational queries and they may have appeared in paragraph-based queries. Apart from this distinct appearance of the top scored keywords in different formulated queries, no further query reformulation, such as reformulation based on the results, was applied.

Queries were scheduled for execution sorted by their priority, starting with the pilot query, next the collocational phrase queries, the collocational synoptic queries, afterwards the queries constructed from remaining keywords if any, and lastly all the paragraph-based queries.

The paragraph-based queries were executed on demand according to their position, if and only if there was no intersection of the query position interval with any of the intervals from all the so far found similarities.

3 Results Downloading and Assessing

A maximum of 100 results obtained from search engines were processed based on each query. Only selected results were downloaded and textually aligned with the suspicious document. The decision whether to download a result was made based on the result's 500 characters long snippets, which were generated for each token from the query and concatenated into one text chunk. If this chunk showed promising similarity with the suspicious document, the result was downloaded. This decision making was adopted from our previous years' implementations, for more information see [10,12].

Each downloaded document was thoroughly compared with its suspicious document by calculating *common features* [11,12] based on word *n*-grams and stop-word *m*-grams [7]. The common features formed valid intervals, which were covered "densely enough" by the features. Two valid intervals were merged if they were closer than 81

Table 1. Query type scope

Query type	#Queries	#URLs retrieved	Scope Usage	Top Retrieval	Zero Retrieval
Pilot	183	16341	89.3%	83.6%	1.1%
Collocational Phrasal	520	34095	65.6%	56.7%	12.3%
Collocational	311	23188	74.6%	64.3%	2.9%
Other Keywords-based	101	5367	53.1%	38.6%	8.9%
Paragraph-based	2109	81788	38.8%	26.8%	18.5%

characters, which estimates the length of a text line. Each resulting valid interval represented one plagiarism case. Such an interval s_{res} was marked in the suspicious document and all the waiting paragraph-based queries for which $s_i \cap s_{res} \neq \emptyset$ were excluded from the queue of prepared queries.

4 Method Assessment

As a training corpus for the source retrieval task, the task organizers provided 98 English-written documents, which contained plagiarized passages from web pages retrieved from the ClueWeb09 document collection. The documents were mostly highly plagiarized, only one document from the corpus was plagiarism free, and this was a short document containing only a single paragraph of 204 words. Each document was about a specific topic and the documents were created manually [5]. The plagiarism cases for each document were annotated. The size of the plaintexts were 30 KB on average and each document contained around five thousand words on average.

During the training phase evaluation, for the whole 98 documents in the training corpus, 32.9 queries per document on average were executed, from which 18.8% were directed to Indri and 81.2% to ChatNoir. In total, 134247 unique URLs were retrieved, provided that each query asked the search engine for 100 results.

The assumption for preparing queries was that the paragraph-based queries were more specific, thus leading to less number of returned URLs. On the other hand, synoptic queries, such as the pilot query and other keywords-based queries, should retrieve more results provided the corpus is large enough. We can affirm this assumption by measuring the number of results returned from the search engine per query type. However, the query generality is limited by the maximum number of results, for which we asked the search engine. Each type of query's specificity and generality is shown in Tab. 1. The column *Scope Usage* shows the average from all queries of one type, expressed as a percentage of the potential maximum number of retrieved results. The table also shows the portion of queries based on which the search engines retrieved the maximum allowed number of results – *Top Retrieval*, which indicates they were general enough under the given conditions. The last column shows portions of queries for which the search engine returned zero answers. Table 1 supports the assumption that paragraph-based queries were more biased in retrieval towards their paragraph text.

The number of results for general queries has little information value. If the query is too general the search engine returns a huge number of results. Therefore, the synoptic

Table 2. Query type performance

Query type	#Queries	#Relevant URLs	Theoretical Portion	Hits per Query
Pilot	183	2815	44.0%	15.4
Collocational Phrasal	520	2974	46.5%	5.7
Collocational	311	1730	27.1%	5.6
Other Keywords-based	101	401	6.3%	4.0
Paragraph-based	2109	2713	42.4%	1.3

query construction must lead to the generation of large number of relevant results. Relevant results can be defined by their purpose, such as, for example, documents following a specific topic or similar to the suspicious document in some extend. We define a result relevant if its text alignment with the suspicious document produces one or more valid intervals, meaning the two documents contain a textually similar passage.

From all of the retrieved results, 6392 were found to be relevant. Please note that not all results were downloaded, therefore, some of the similarities might have been missed during the download decision. From all of the discovered URLs, 32538 were actually downloaded and textually aligned with its suspicious document. For all except one suspicious document that contained plagiarism, some relevant documents were found.

Table 2 shows the performance of queries by their type[5]. The third and fourth columns show the number of successful results and the coverage of results of current query type respectively. One relevant URL was counted into the total number of successful results only once, but some queries led to the retrieval of already discovered results. Therefore, in order to make an unbiased evaluation of the coverage, in terms of query execution sequence, each successful result is credited with all queries which retrieved that result. Table 2 shows that the portion of retrieved relevant URLs for the pilot, phrasal and paragraph-based queries is getting closer to nearly a half of all relevant retrieved URLs, but phrasal and paragraph-based queries needed nearly 3 times and 11 times more searches than the pilot queries respectively. The average number of hits per query depicts the fifth column of Tab. 2, which supports the assumption that the pilot query is the most important and it is the best choice for the synoptic search to start with, in order to cover the majority of plagiarism as quickly as possible.

The paragraph-based queries have relatively low yield of relevant results per one query, which is due to their specificity (Tab. 1), but they can cover a large portion of successful results. Therefore, it may be beneficial to skip these queries for some parts of input documents, e.g. parts where plagiarism was already discovered, and use them in order to aim the search for more suspicious parts, for example, parts selected using intrinsic plagiarism methods. Both tables show that 2109 paragraph-based queries were executed, however, the total number of prepared paragraph-based queries was 6693,

[5] For all the 98 suspicious documents, there were only 183 pilot based queries executed, despite the fact, that the pilot query should have been processed by both ChatNoir and Indri search engines. For 13 suspicious document, the pilot query was processed by only one search engine. There were missed 12 queries in Indri and one query in ChatNoir because of the timeout. The search engines were utilized during a standard operation over the network with timeout set to 8 minutes.

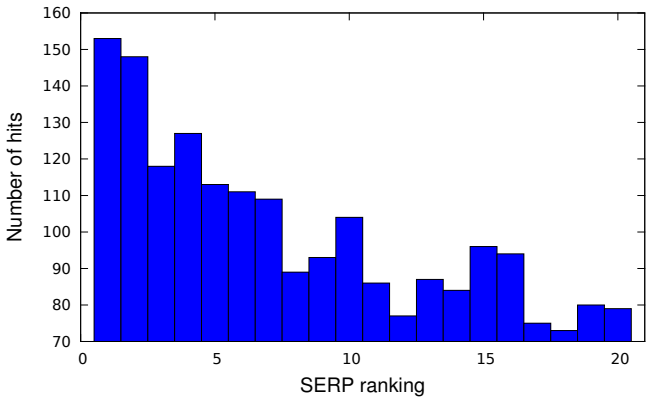

Figure 1. Number of relevant results per SERP rank.

meaning that there was 68.5% of such queries omitted due to their position inside an already discovered interval of textual similarity.

Since the pilot query yields 15.4 relevant results per query, it is clear that a search engine should be asked to retrieve at least tens of results based on this query type. However, the number of results influences not only recall of the system, but also time and space requirements of the system. Relevant URLs were also retrieved from very high sequence numbers of ranking in the Search Engine Result Page (SERP). Similarities were found even among higher than 100th result based on one query. The limit 100 results was set due to the time consumption of URLs checking. One of the master hits (see further) was retrieved from SERP's 100th position. Figure 1 depicts the total number of relevant URLs retrieved at first 20 positions[6] of SERP based on all queries.

Surprisingly poor performance compared to other types, can be observed at the other keywords-based queries (see both Tab. 1 and Tab. 2), which indicates that extraction of less than 10 quality keywords is sufficient for such texts. Table 1 shows that their scope was around half of asked results, despite the fact that they were aimed for synoptic theme related searches. Table 2 indicates that they covered only 6.3% of all discovered similarities; on the other hand, in terms of number, they represented the smallest type of queries.

The discovered relevant URLs contained some portion of similar text with the suspicious document. In the ClueWeb09 corpus, many texts are reused like in a real web, therefore, many web pages may be classified as near-duplicates [2] and many documents just contain smaller or larger passages identical to other web page. The retrieved relevant results are among those cases. However, the source of each plagiarism in the corpus of suspicious documents were annotated with the original web page from which the text was reused. We call the retrieval of such an original document a master hit.

[6] 20 best ranked URLs for each query.

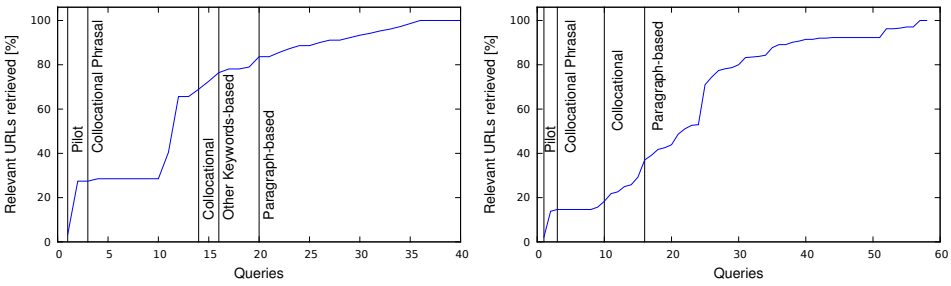

Figure 2. Relevant URLs retrieval progress of two selected suspicious documents.

Taking into account only master hits[7], for the whole input corpus, the overall recall was 0.45 with 5 document having 100% recall and 12 documents without a master hit. We consider this performance as a very good result providing that no near-duplicates were taken into account. Figure 2 shows the progress of detection during the scheduled querying of two selected[8] documents. The *y* axis shows the percentage of retrieved relevant documents. The portion of queries covered by specific query type is distinguished with the type-labelled vertical line separators. The number of queries to first detection is also evaluated in PAN, this is the job for pilot queries, which should lead to positive results using the very first two queries. From the right plot of Fig. 2, it can be seen that paragraph-based queries can highly support the detection, if fewer similarities were discovered using previous types of queries. In a real-world situation, while expecting the documents to contain less plagiarism, we would try to lower the number of executed paragraph-based queries with methods detecting suspicious parts of the input documents, and schedule the paragraph-based queries located only in those parts. The left plot of Fig. 2 shows pilot and phrasal queries as the most profitable, which was in most cases.

5 Conclusions

This paper described an architecture of PAN 2015 software for source retrieval in a plagiarism detection task. The key settings were discussed and analyses of the settings provided. The software was based on previous versions used in PAN since 2012 [11,12,10], this paper also described changes made for 2015 lab at PAN.

For the source retrieval, based on each suspicious document, queries of several types were prepared: keywords-based; divided further into the pilot, phrasal collocations, collocations, and other keywords-based; and the paragraph-based queries which were associated with the position in the suspicious document of the paragraph they characterized. Queries were executed sequentially and all results from each query were evaluated, in

[7] The master hits analysis is included because of the low precision, which the system achieved in the test phase of the lab. In the real-world, the ani plagiarism system must provide the user the possibility of examination of relatively small textual similarities.

[8] For those documents, the highest number of distinct relevant URLs was retrieved.

order to skip some of the paragraph-based queries for whose paragraphs a similarity was already detected. Final results containing the valid intervals with the suspicious document, were reported. The pilot queries proved to be the best choice for synoptic search and the paragraph-based queries manage to perform well in the positional retrieval, which is biased towards searching for specific short texts.

The retrieval recall in the lab's official test run, compared to the previous year, has increased, but so has the total number of used queries. However, the proposed methodology achieved the highest recall with usage of the least number of queries among the PAN 2015 softwares during the official test run. The discussion and evaluation of PAN can be found in the lab overview paper by the lab organizers [1].

References

1. Hagen, M., Potthast, M., Stein, B.: Source Retrieval for Plagiarism Detection from Large Web Corpora: Recent Approaches. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2015)
2. Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th International Competition on Plagiarism Detection. In: CLEF 2012 Evaluation Labs and Workshop (2012)
3. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B.: Overview of the 6th International Competition on Plagiarism Detection. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 845–876 (2014)
4. Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., Welsch, C.: ChatNoir: A Search Engine for the ClueWeb09 Corpus. In: Hersh, B., Callan, J., Maarek, Y., Sanderson, M. (eds.) 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12). p. 1004. ACM (Aug 2012)
5. Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: ACL (1). pp. 1212–1221. The Association for Computer Linguistics (2013)
6. Salton, G., Yang, C.: On the Specification of Term Values in Automatic Indexing. Tech. Rep. TR-73-173, Cornell University (Ithaca, NY US) (1973)
7. Stamatatos, E.: Plagiarism detection using stopword *n*-grams. JASIST 62(12), 2512–2527 (2011)
8. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A Language-model Based Search Engine for Complex Queries. Tech. rep., in Proceedings of the International Conference on Intelligent Analysis (2005)
9. Suchomel, Š., Brandejs, M.: Approaches for Candidate Document Retrieval. In: Information and Communication Systems (ICICS), 2014 5th International Conference on. pp. 1–6. IEEE, Irbid (2014)
10. Suchomel, Š., Brandejs, M.: Heterogeneous Queries for Synoptic and Phrasal Search. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 1017–1020 (2014)
11. Suchomel, S., Kasprzak, J., Brandejs, M.: Three way search engine queries with multi-feature document comparison for plagiarism detection. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012 (2012)
12. Suchomel, S., Kasprzak, J., Brandejs, M.: Diverse Queries and Feature Type Selection for Plagiarism Discovery. In: Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013. (2013)