Plagiarism Detection in MEDLINE Using Multiple Query Expansion and Approximate Phrase Match Techniques

Vijay S Mhaskar, Prof. Thombre B. H

Department of Computer Engineering,
Shree Ramchandra College of Engineering, Lonikand Pune
Savitribai Phule Pune University, Maharashtra.
vijaymhaskar@gmail.com, babanthombre@gmail.com

Abstract—Plagiarism is using others work and using them off as one's own without appropriately acknowledging the source. Identifying duplicated and plagiarized passages of text is becoming popular area of research. Plagiarism is an important issue in every academic and research institutes and this situation is becoming worse with easily available online resources. MEDLINE has a dataset of more than 26 million publications from 5639 various publications in the area of medicine and related fields. New publications are getting added continuously which makes difficult to keep track of information contained within it. In this paper we have proposed method for plagiarism detection that will identify most relevant sources of plagiarism from MEDLINE. We have used set of document matching rules based on query expansion techniques and approximate phrase match techniques are used to match relevant documents. Also the system that we are proposing for candidate document selection is built on highly scalable open source technologies. finally, when ranking suspected documents, we carefully rank those documents based on combined weight of various match types. We have evaluated our proposed approach with recently available MEDLINE corpus which gave us very promising results.

Index Terms—Candidate Document Retrieval, Information Retrieval, Plagiarism, Plagiarism Detection, Query Expansion, Search Engine, Semantic Search

I. INTRODUCTION

Internet is the biggest source of information. Now, people can easily search, get access and browse the web to get the information they need. Just imagine how difficult it would be to do scientific research without the Internet and web space. Educational institutions and academic conferences wish to ensure that the work products submitted to them have not been plagiarized from another source. Furthermore, due to the size and digital structure of the Internet, it is easy to illegally use someone else work now. The problem of plagiarism has a direct relationship with the growing scientific community. The most common plagiarism is written text document which is formed by copying some or all parts of the original document, sometimes with some modifications[3]. Identification of documents which are copied is a slow process to humans because of large number of documents which have to be analysed. The documents in digital format make the process of plagiarism quite simple, it means that we can trace such cases automatically.

A. Problem Definition

The taxonomy of plagiarism is developed from intelligent identification of intra-textual manipulations. Alzahrani et al[3], presented a very good model that is based upon qualitative analysis of various findings during the review of student submissions. Here they have presented two dominant cases of plagiarism, the literal and the intelligent. Literal plagiarism is copying or manipulation of textual outputs either in whole or in part without providing due credit to the originator. Within the framework of literal plagiarism, plagiarists are unlikely to spend significant time attempting to hide their academic crime. Intelligent plagiarism, on the other hand, is much more difficult to detect and involves the manipulation of text, translation of foreign copy, or the adoption and ownership of others ideas, theories, or concepts.

Platform scalability problem needs to be addressed in any plagiarism detection system. The system that we are using for to detect suspicious documents should be enough to handle large document collection as well as complex query operations. Existing systems performs poorly when it comes to multiple complex matching operations on large document collections such as MEDLINE. Also while searching for suspicious documents in MEDLINE corpus if only exact phrase match technique is used then It may fail to identify the document similarity if original text has been rewritten which most of the plagiarist attempt. This needs to handled using multiple match types along with phrase match. Also While generating final ranked list of documents existing system[1] uses combSUM[12] as a rank fusion technique which will not have influence of match type of search query.

B. Motivation

The Motivation of this research work is to improve the accuracy and efficiency of existing plagiarism detection technique. Educational institutions and academic conferences wish to ensure that the work products submitted to them have not been
plagiarized from another source. MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database of life sciences and biomedical information. In this approach of candidate document selection, we use scalable IR-based approach which incorporates multiple query expansions and approximate phrase match to identify obfuscated documents. The proposed approach has not been previously used for retrieving candidate documents from MEDLINE.

C. Objectives and Goals

Objective for this project is to use multiple query expansions from UML resources to search documents. Also Various approximate match techniques also used to search for candidate documents. We will build information retrieval system based on popular open source search platform Apache Solr[8] which is tested regularly at scale to handle billions of documents with very high query volumes. Apache Solr is based on popular Apache-Lucene search library[9].

- Objective of this project is to develop highly scalable system to identify potential sources of plagiarism from MEDLINE.
- Improve the accuracy of plagiarism detection using multiple query expansion techniques and approximate phrase match technique.
- Investigate rank fusion technique for fair ranking between multiple results sets of suspicious documents.

II. REVIEW OF LITERATURE

In plagiarism detection systems, mainly two important problems are considered: the problem of candidate documents selection that are globally similar to a document which is under investigation, and the problem of comparing of document under investigation and its candidates to pinpoint plagiarized fragments in detail.

In [1] Rao Muhammad Adeel Nawab, Mark Stevenson and Paul Clough proposed an approach to find plagiarism in MEDLINE using query expansion techniques. Query expansion is performed using the ULMS Metathesaurus to deal with situations in which original documents are obfuscated. It mainly focuses on cases where plagiarised text has been highly obfuscated which presents major challenge to automated plagiarism detection systems. Evaluation was carried out using the MEDLINE Corpus, which contains potential real cases of plagiarism. Results show that the IR-based approach using query expansion outperforms a state-of-the-art approach. The IR-based approach proposed here achieves higher results than the Kullback-Leibler Distance approach. As this system is based on Lucene[11] Although it is expected that performance will drop when the entire MEDLINE database is used.

In [2] Martin Potthast, Alberto Barn-Cedeo, Benno Stein, Paolo Rosso described approaches for producing exact and modified copies. Detecting plagiarism which involves little or no modifications of the original document is straightforward. However real examples where source text was rewritten using Anti Anti Plagiarism systems. This paraphrased passage was analysed by two well-known commercial plagiarism detection services and both are failed to detect plagiarism.

In [3], Alzahrani, Salma M., Naomie Salim, and Ajith Abraham proposed detail taxonomy, approaches used and semantic framework of the tool. After detailed Analysis of various methods, authors suggest that semantic and Fuzzy Based method can provide better results. Both semantic and Fuzzy are challenging areas. Since no standard Fuzzy dataset is available [5] to find fuzzy words whereas in case of semantic detection it is difficult to represent semantic of sentence. They also explained various other approaches which are used to detect plagiarism. It covers simple lexical methods to complex semantic methods. He also concluded that current plagiarism tools for educational institutions, academics and publishers limited to word to word plagiarism and only some instances of it. They do not cover adapting ideas from others. Now plagiarism has become more sophisticated, idea plagiarism is a key academic problem and should be addressed in future research[4]. He also proposed structural features and contextual information with efficient STRUC-based methods to detect section based importance and context based adaption idea plagiarism.

In [6] Alberto Barn-Cedeo, Paolo Rosso, Jos-Miguel Bened proposed an approach which retrieves candidate documents using the Kullback-Leibler Symmetric Distance method. Documents are modelled as probability distributions and compared using KL. Documents are converted into probability distributions by removing stop words, stemming and then computing tf.idf weights for the remaining word unigrams. Results showed that the overall accuracy and speed of the plagiarism detection system improved by applying the Kullback-Leibler Symmetric Distance to reduce the plagiarism detection search space.

III. SYSTEM OVERVIEW

A. Candidate Retrieval Process

In the proposed system IR-based approach is used to identify candidate source documents The source collection is indexed with an IR system. Indexing source collection in IR system is an offline job. In the IR-based framework, the candidate retrieval process can be divided following steps,

1) Pre-processing: In this phase each input suspicious document is split into sentences. NLTK is used for tokenizing input text into sentences. Then in the analysis phase all terms converted to lower case, stop words and punctuation marks are removed, then remaining terms are stemmed using porter stemming algorithm before adding them to index.

2) Query Formulation: We are using sentences from the suspicious document to create multiple queries. While creating these queries various search types like partial phrase match, minimum term match are used to match sentences in index. This will make sure that we are not missing any important documents.

3) Query Expansion: The Unified Medical Language System (UMLS), a set of tools and resources to assist with
the development of biomedical text processing systems, is used to carry out query expansion. Our approach uses two main UMLS resources the Metathesaurus and MetaMap[11]. Also along with these resources user can configure his own dictionary which can be used for query expansion.

4) **Retrieval:** Terms are weighted using the tf.idf weighting scheme and then input text which is supplied as a query is used to retrieve similar documents from the index. For each query result documents are ranked as per their similarity score.

5) **Results Merging:** Result set from each query type will be combined considering weight of each query type. The top N documents returned against multiple queries are merged to generate a final ranked list of source documents. Merging results from multiple queries is likely to perform well perform well in situations when large portion of text is altered. This stage is important in obtaining effective and balanced results.

**Fig. 1. Proposed System**

**B. Mathematical Model of Proposed System**

Let $S$ be the set of Inputs, Functions and Outputs $S = \{I, F, O\}$ Where, $I$ represent Input Dataset, $I = \{I_1, I_2\}$ $F$ represents the set of functions that are performed on the input dataset, $F = \{F_I, F_Q\}$

$F_I = \{F_1, F_2, F_3, F_4, F_5, F_9\}$

$F_Q = \{F_3, F_4, F_7, F_8, F_9\}$

$O$ represents output of system $O = \{O_1, O_2\}$

**Inputs:**

$I_1 = $ Input Dataset

$I_2 = $ Input Text

**Functions:**

$F_1 = $ Document Fetching from MEDLINE Repository.

$F_2 = $ Parsing Datasets.

$F_3 = $ Preprocessing Datasets.

$F_4 = $ Linguistic Analysis of Text.

$F_5 = $ Indexing of Text in Solr.

$F_6 = $ Searching Query.

$F_7 = $ Ranking Results.

$F_8 = $ Merging Results.

$F_9 = $ Logging.

**Output:**

$O_1 = $ Set of Indexed Documents

$O_2 = $ Set of Suspicious Documents

**Fig. 2. Total documents available for search in index**

Also, In practical scenario total number of indexed documents will be subset of total number of fetched documents from MEDLINE, And Total number of fetched documents will be subset of total corpus size of Medline. In an ideal case these three will exactly same. But some documents will be discarded at fetching and parsing level due to various fetching, parsing or data issues, All these can be represented as below,

Where,

$U = $ Available corpus size at Medline

$F = $ Total number of fetched documents

$I = $ Total number of indexed documents

$S = $ Detected suspicious documents for given query.

So, Total Documents available for search in Index = $I \cap F$
IV. System Analysis

A. Retrieval

For any input text the results that we get are ranked using Lucene scoring formula. It used combination of the Vector Space Model (VSM) and Boolean model to determine document relevance. It uses the Boolean model to first narrow down the result set based on Boolean logic in users query and then uses tf.idf factors to score documents. Below is Lucene’s practical scoring formula[9].

\[
\text{score}(q, d) = \text{coord}(q, d) \times \text{queryNorm}(q) \times \sum_{t \in q} (t \times \text{idf}(t))^2 \times t \times \text{getBoost()} \times \text{norm}(t, d)
\]  

(1)

B. Result Merging

For result merging module \( S_{\text{finalScore}} \), is calculated by adding the similarity scores of source documents obtained against each query \( q \). Also we are considering weight of each query type while calculating score.

\[
S_{\text{finalScore}} = \sum_{q=1}^{N_q} S_q(d) \times W_q
\]  

(2)

where \( N_q \) is the total number of queries to be combined and \( S_q(d) \) is the similarity score of a source document \( d \) for a query \( q \). \( W_q \) is a weight of query type.

C. Result Evaluation

Candidate source document retrieval is carried out for each suspicious document, here focus is on returning less incorrect documents. If we miss important documents at the top of the list then they will not be identified in the later stages of plagiarism detection. Here we are using recall for top \( K \) documents across multiple queries. For a single query the Recall at \( K \) \( (R@K) \) is 1 if the source document appears in the top \( K \) documents retrieved by the query, and 0 otherwise. For a set of \( N \) queries, the averaged recall at \( K \) score is calculated as:

\[
R@K_{\text{avg}} = \frac{1}{|N|} \sum_{i=1}^{N} R@K_i
\]  

(3)

D. System Requirements

For this evaluation we have used below configurations. Considering the Medline data size we will need below hardware configurations to run our system efficiently.

1) Software Specification:

- Java version 1.8+ (JDK and JRE)
- Apache Solr 5.2.1
- Operating System: Microsoft Windows 8
- Integrated Development environment (IDE) : Eclipse Java EE IDE
- Application Server: Apache Tomcat 8

2) Hardware Specification:

- Hard Disk: 1 TB
- RAM: 8 GB
- Dual-core CPU, preferably above 2GHz

Also for these evaluation, the source collection is formed from the 2015 MEDLINE/PubMed Baseline Repository. The collection of suspicious documents contains 260 citations that were manually verified as duplicates. All the selected citation pairs do not have a common author so making them potential cases of plagiarism [10].

<table>
<thead>
<tr>
<th>Approach</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Query Expansion</td>
<td>0.858</td>
<td>0.875</td>
<td>0.925</td>
<td>0.928</td>
<td>0.928</td>
</tr>
<tr>
<td>Multiple Query Expansion</td>
<td>0.915</td>
<td>0.956</td>
<td>0.96</td>
<td>0.958</td>
<td>0.958</td>
</tr>
</tbody>
</table>

Table I shows the results of current implementation. We have calculated recall value at top 1, 5, 10, 15 and 20 candidate source documents. Our proposed approach for retrieving candidate documents performs better in identifying real cases of plagiarism. It is also expected that performance is improve when we apply Query expansion with WSD.

V. Conclusion

The digital document is being replicated across the server. Many times it happens that people makes the near copies of the original document. This paper explains and evaluates a new query expansion approach to the problem of candidate document selection for extrinsic plagiarism detection. Main focus of this paper is on scalability and cases when the plagiarised version has been highly obfuscated as this presents the greatest challenge to automated plagiarism detection systems. Evaluation was carried out using the MEDLINE Corpus which has potential real cases of plagiarism. Results show that this approach using multiple query expansions is more scalable than traditional approaches and also performs better than traditional approaches for candidate document retrieval task.

ACKNOWLEDGMENT

I am profoundly grateful to Prof. Baban Thombre for his guidance and continuous encouragement throughout to see that this paper rights its target since its commencement to its completion. I would like to express deepest appreciation towards his invaluable guidance and support me in completing this paper. Also I must express my sincere gratitude to all the staff members of Computer Engineering Department, SRCOE Lonikand, Pune who helped me directly or indirectly during this course of work.
REFERENCES


