JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

Research Paper
J. of infosci. theory and practice 3(1): 24-39, 2015
http://dx.doi.org/10.1633/JISTaP.2015.3.1.2

Query Formulation for Heuristic Retrieval in Obfuscated and Translated Partially Derived Text

Aarti Kumar *
Department of Computer Applications
Maulana Azad National Institute of Technology
Bhopal, India
E-mail: aartikumar01@gmail.com

Sujoy Das
Department of Computer Applications
Maulana Azad National Institute of Technology
Bhopal, India
E-mail: sujdas@gmail.com

ABSTRACT

Pre-retrieval query formulation is an important step for identifying local text reuse. Local reuse with high obfuscation, paraphrasing, and translation poses a challenge of finding the reused text in a document. In this paper, three pre-retrieval query formulation strategies for heuristic retrieval in case of low obfuscated, high obfuscated, and translated text are studied. The strategies used are (a) Query formulation using proper nouns; (b) Query formulation using unique words (Hapax); and (c) Query formulation using most frequent words. Whereas in case of low and high obfuscation and simulated paraphrasing, keywords with Hapax proved to be slightly more efficient, initial results indicate that the simple strategy of query formulation using proper nouns gives promising results and may prove better in reducing the size of the corpus for post processing, for identifying local text reuse in case of obfuscated and translated text reuse.

Keywords: Heuristic, obfuscated, translated, simulated paraphrasing, retrieval, Hapax, query formulation, pre-retrieval

1. INTRODUCTION

Text reuse identification has become a challenging problem due to the presence of enormous amounts of digital data, more so because of obfuscated text reuse. The result of obfuscation is a modified version of the original text. The modification can be at the level of words, phrases, sentences, or even whole texts by ap-

Open Access

Accepted date: March 10, 2015
Received date: August 20, 2014

***Corresponding Author:** Aarti Kumar
Research Fellow
Department of Computer Applications
Maulana Azad National Institute of Technology, Bhopal, India
E-mail: aartikumar01@gmail.com

plying a random sequence of text operations such as change of tense, alteration of voice (active to passive, and vice versa), change in treatment of direct speech, abbreviations, shuffling a word or a group of words, deleting a word, inserting a word from an external source, or replacing a word with a synonym, antonym, hypernym, or hyponym. These alterations may or may not modify the original meaning of the source text.

Obfuscation in text reuse can be at different levels of degree. It may be from no obfuscation to a low or high level of obfuscation. The range from word to sentence level defines the level of obfuscation from low to high. Translation of text from the language of source document to another language of suspicious target document is also a kind of high level obfuscation which is extremely difficult to deal with, more so when the reuse is local in nature. Local reuse occurs when sentences, facts, or passages, rather than whole documents, are reused and modified.

Therefore, techniques that can be applied in order to identify the different levels of obfuscation and their local nature may also vary due to the complexity of the problem. In a large corpus analyzing complete sets of source documents for local text reuse is an expensive affair; therefore, it is better to retrieve a subset of documents and then at the time of post processing, a retrieved set of source documents may be analyzed for local text reuse. Initial filtering is known as heuristic retrieval (Barrón-Cedeño, 2010). Heuristic retrieval, also called pre-retrieval, can therefore be defined as the process of retrieving a small sub-set of a potential reused document for any particular source-document from a large set of documents (corpora), with a view to minimize time and space requirements.

Heuristic retrieval is important due to the (i) enormous size of a typical corpus; (ii) presence of large numbers of irrelevant documents for a particular set of suspicious documents; and (iii) the processing cost involved in processing the dataset. Also, while for small document collections it is practicable to perform a complete comparison against every document, this is obviously not possible when the collection size is enormous. So it is better to go for heuristic retrieval before post processing.

In this paper an attempt is made to find key terms from a target set of suspicious documents to retrieve an initial set of source documents for further post processing. In case of local obfuscated text reuse, generating keywords from the whole text can drift the query and may fetch many unwanted documents. The main focus is, therefore, to formulate effective queries to retrieve a subset of documents that bears more closeness to any given suspicious document.

In the proposed work three methods have been studied for formation of query, especially when the content of the documents are obfuscated or translated and the text reuse is local in nature. The results are derived from the PAN CLEF 2012 training corpus.

In Section 2 existing work in pre-retrieval query performance is discussed, Section 3 discusses proposed methodology, and Section 4 and 5 discuss experimental results and conclusions, respectively.

2. RELATED WORK

Hauff, Hiemstra, and Jong (2008) assessed the performance of 22 pre-retrieval predictors on three different TREC collections. As most predictors exploit inverse term/document frequencies in some way, they hypothesize that the amount of smoothing influences the quality of predictors.

Cummins, Jose, and O'Riordan (2011) developed a new predictor based on standard deviation of scores in a variable length ranked list, and showed that this new predictor outperforms state-of-the-art approaches without the need of tuning.

Possas et al. (2005) worked on TREC-8 test collection and proposed a technique for automatically structuring web queries as a set of smaller sub queries. To select representative sub queries, information of distributions is used and a concept of maximal term sets derived from formalism of association rules theory is used for modelling.

Many kinds of text reuse detection techniques have been proposed from time to time by different authors, including: Potthast et al. (2013); Gustafson et al. (2008); Mittelbach et al. (2010); Palkovskii, Muzyka, and Belov (2010); Seo and Croft (2008); Clough et al. (2002); Gupta and Rosso (2012); Bar, Zesch, and Gurevych (2012); and Potthast et al. (2013a, b).

Gipp et al. (2013) proposed Citation-based Plagiarism Detection. Compared to existing approaches, CbPD does not consider textual similarity alone, but uses the citation patterns within scientific documents as a unique, language independent fingerprint to identify semantic similarity.

Vogel, Hey, and Tillmann (1996) presented an HMM-based approach for modelling word alignments in parallel texts in English and French. The characteristic feature of this approach is to make the alignment probabilities explicitly dependent on the alignment position of the previous word. The HMM-based approach produces translation probabilities.

Barrón-Cedeño (2012) compared two recently proposed cross-language plagiarism detection methods: CL-CNG, based on character n-grams, and CL-ASA, based on statistical translation, to their new approach based on machine translation and monolingual similarity analysis (T+MA). Barrón-Cedeño explores the effectiveness of his approach for less related languages. CL-CNG is not appropriate for this kind of language pairs, whereas T+MA performs better than the previously proposed models. The study investigated Basque, a language where, due to lack of resources, cross language plagiarism is often committed from texts in Spanish and English.

Grozea and Popescu (2009) evaluated cross-language similarity among suspected and original documents using a statistical model which finds the relevance probability between suspected and source documents, regardless of the order in which the terms appear in the suspected and original documents. Their method is combined with a dictionary corpus of text in English and Spanish to detect similarity in cross language.

A plagiarism detection technique based on Semantic Role Labeling was introduced by Osmana et al. (2012). They improved the similarity measure using argument weighting with an aim to studying the argument behaviour and effect in plagiarism detection.

Pouliquen et al. (2003) have worked on European languages and have presented a working system that can identify translations and other very similar documents among a large number of candidates, by representing the document content with a vector of thesaurus terms from multilingual thesaurus, and then by measuring the semantic similarity between the vectors.

The approach used by Palkovskii and Belov (2011) implied the usage of automatic language translation (Google Translate web service) to normalize one of the input texts to the target comparison language, and applies a model that includes several filters, each of which adds ranking points to the final score.

Ghosh, Pal, and Bandyopadhyay (2011) treated cross-language text re-use detection as a problem of Information Retrieval, and it is solved with the help of Nutch, an open source Information Retrieval (IR) system. Their system contains three phases – knowledge preparation, candidate retrieval, and cross-language text reuse detection.

Gupta and Singhal (2011) tried to see the impact of available resources like Bi-lingual Dictionary, WordNet, and Transliteration, mapping Hindi-English text reuse document pairs and using the Okapi BM25 model to calculate the similarity between document pairs.

The approach used by Aggarwal et al. (2012) in journalistic text reuse consists of two major steps, the reduction of search space by using publication date and vocabulary overlap, and then ranking of the news stories according to their relatedness scores. Their approach uses Wikipedia-based Cross-Lingual Explicit Semantic Analysis (CLESA) to calculate the semantic similarity and relatedness score between two news stories in different languages.

Arora, Foster, and Jones (2013) used an approach consisting of two steps: (1) the Lucene search engine was used with varied input query formulations using different features and heuristics designed to identify as many relevant documents as possible to improve recall; and (2) merging of document list and re-ranking was performed with the incorporation of a date feature.

Pal and Gillam (2013) converted English documents to Hindi using Google Translate and compared them to the potential Hindi sources based on five features of the documents: title, content of the article, unique words in content, frequent words in content, and publication date using Jaccard similarity. A weighted combination of the five individual similarity scores provides an overall value for similarity.

Tholpadi and Param (2013) describe a method that leverages the structure of news articles, especially the title, to achieve good performance on the focal news event linking task. They found that imposing date constraints did not improve precision.

IDF, Reference Monotony, and Extended Contextual N-grams were used by Torrejon and Ramos (2013) to link English and Hindi News.

Haiduc et al. (2013) have proposed a recommender, Refoqus, which automatically recommends a reformulation strategy for a given query to improve its retrieval performance in Text Retrieval. Refoqus is based on Machine Learning and its query reformulation strategy is based on the properties of the query.

Carmel et al. (2006), while trying to find a solution to the question "what makes a query difficult," have devised a model to predict query difficulty and number of topic aspects expected to be covered by the search results and to analyze the findability of a specific domain.

The Capacity Constrained Query Formulation method was devised by Hagen and Stein (2010). They focused on the query formulation problem as the crucial first step in the detection process and have presented this strategy, which achieves better results than maximal term set query formulation strategy.

3. PROPOSED METHODOLOGY

Most of the authors listed have worked upon whole corpora, which consume valuable resources with respect to space and time. Also, most query formulation and retrieval strategies fail in the case of highly obfuscated and translated reused documents. In such cases, even the most popular TF-IDF strategy is no exception. Use of thesaurus in query formulation aids in query drifting and results in fetching unwanted irrelevant documents. Devising a simple strategy which can reduce the size of the corpora by retrieving potential reused documents in the pre-retrieval stage, and before going for state-of-the-art text reuse or plagiarism detection techniques, can render the process more efficient in terms of system resources and would produce more accurate results. Our strategy is straightforward and simple and tackles this important pre-computation step that finds promising candidate documents for in-depth analysis.

In this preliminary study, an attempt has been made to analyse proposed strategies on the training corpus of PAN CLEF 2012 which is 1.29 GB in size and contains 12,024 files divided into 8 folders.

The suspicious documents that fall under the categories of low and high obfuscation, simulated paraphrasing, and translation are studied in this paper. Simulated paraphrasing is intentional obfuscation done by humans with an intention to hide plagiarism attempts.

The steps followed for formulating a query for heuristic retrieval are as follows:

3.1. Pre-Processing

As we are dealing with local text reuse and obfuscation which can be in any part of the document, the document is divided into units. The document unit that we have taken is 'paragraph,' with the assumption that even if sentences or a block is reused and obfuscated, that will most likely be a paragraph or a sentence within any paragraph. Therefore, the suspicious document is divided into paragraphs before tokenizing the text. The document is normalized by removing the punctuation. Stop-words, verbs, and adverbs are removed using a pre-compiled list of stop-words,[1] verbs, and adverbs[2] obtained from the web.

3.2. Query Formulation

Queries have been formulated paragraph-wise for each suspicious document. So the number of paragraphs in the suspicious document decides the number of queries for that document. Relying on a given document structure like paragraphs bears the risk of failing for some unseen documents that are not well formatted (Potthast et al., 2013); still, the idea behind generating queries for each paragraph is that if reuse is local, then at least keywords from the paragraph which have been reused will maximize the chance of fetching the required source document. The query is generated from

[1] List of stop-words available at:
http://www.ranks.nl/resources/stopwords.html
http://norm.al/2009/04/14/list-of-english-stop-words/
http://www.webconfs.com/stop-words.php
http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[2] List of verbs and adverbs available at:
http://www.englishclub.com/vocabulary/regular-verbs-list.htm

each paragraph by selecting a) proper nouns; b) unique words or Hapax; or c) most frequent words.

Along with this we have also tried many other strategies for heuristic retrieval in obfuscated corpora, none of which showed fruitful results. Therefore we are not discussing them in our work which we have presented here.

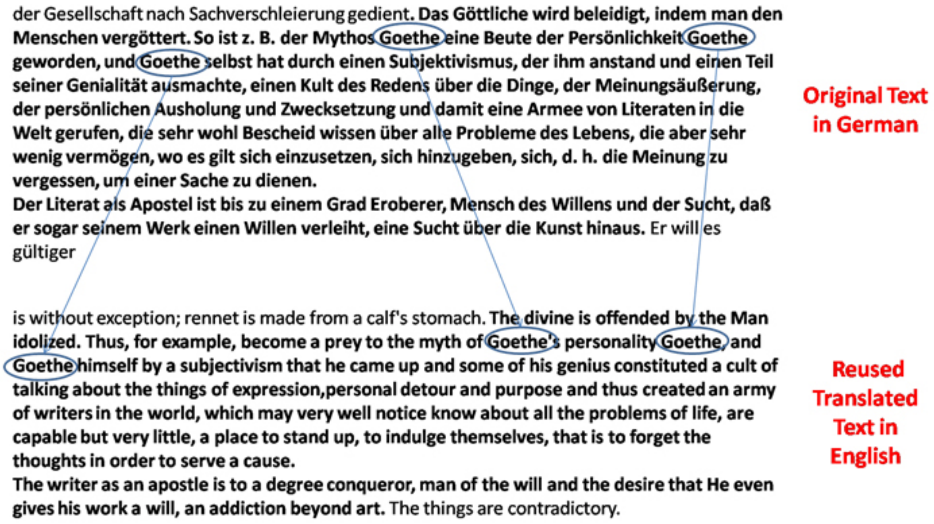
Three pre-retrieval strategies lead to the formulation of these three different runs:

3.2.1. Run 1: Query Formulation Using Proper Nouns

It has been observed that rarely are nouns or proper nouns ever changed while obfuscating or translating the text. This feature of obfuscation is highly visible in translated text reuse (Fig. 1) which shows only the unchanged proper noun "Goethe" when the whole text in German was translated into English for reuse (suspicious doc-id:01277, source doc-id:02637, susp_language="en" susp_offset="29575" susp_length="675" source_language="de" source_offset="176202" source_length="787" of PAN CLEF).

This prompted us to select proper nouns for formulating queries. The assumption that if the same scripts are being used, only the proper nouns—that is, the names of persons, locations, and organizations, etc.—do not change, led to the formulation of queries with proper nouns. The grammar rule that proper nouns begin with a capital letter has been used to identify them.

Before formulation of query, stop-words/function words were removed using a precompiled list of stop-words, verbs, and adverbs acquired from the web. Cases of the words were taken care of. Thus what were left were mainly nouns and adjectives. Mostly the adjectives do not start any sentence without the support of an article and therefore, adjectives could not have been the ones with a starting uppercase letter. Any noun may have contained a starting uppercase letter and could have become the only source of introducing noise and contributing to a few false positives.



der Gesellschaft nach Sachverschleierung gedient. **Das Göttliche wird beleidigt, indem man den Menschen vergöttert. So ist z. B. der Mythos Goethe eine Beute der Persönlichkeit Goethe geworden, und Goethe selbst hat durch einen Subjektivismus, der ihm anstand und einen Teil seiner Genialität ausmachte, einen Kult des Redens über die Dinge, der Meinungsäußerung, der persönlichen Ausholung und Zwecksetzung und damit eine Armee von Literaten in die Welt gerufen, die sehr wohl Bescheid wissen über alle Probleme des Lebens, die aber sehr wenig vermögen, wo es gilt sich einzusetzen, sich hinzugeben, sich, d. h. die Meinung zu vergessen, um einer Sache zu dienen.**
Der Literat als Apostel ist bis zu einem Grad Eroberer, Mensch des Willens und der Sucht, daß er sogar seinem Werk einen Willen verleiht, eine Sucht über die Kunst hinaus. Er will es gültiger

is without exception; rennet is made from a calf's stomach. **The divine is offended by the Man idolized. Thus, for example, become a prey to the myth of Goethe's personality Goethe, and Goethe himself by a subjectivism that he came up and some of his genius constituted a cult of talking about the things of expression,personal detour and purpose and thus created an army of writers in the world, which may very well notice know about all the problems of life, are capable but very little, a place to stand up, to indulge themselves, that is to forget the thoughts in order to serve a cause.**
The writer as an apostle is to a degree conqueror, man of the will and the desire that He even gives his work a will, an addiction beyond art. The things are contradictory.

Fig. 1 Text reuse in cross language documents

http://www.momswhothink.com/reading/list-of-verbs.html
http://www.linguanaut.com/verbs.htm
http://www.acme2k.co.uk/acme/3star%20verbs.htm
http://www.enchantedlearning.com/wordlist/verbs.shtml
http://www.enchantedlearning.com/wordlist/adverbs.shtml

3.2.2. Run 2: Query Formulation Using Unique Words (Hapax)

The terms which appear only once in the document are also known as *Hapax legomenon* or Hapax. Hapax is a term that occurs only once within a context, either in the written record of an entire language, in the works of an author, or in a single text. In run 2 we used such terms from each paragraph for formulating the query.

The reason behind taking Hapax for query formulation is that even if one or two words are left out at the time of obfuscation, then these words shall help in identifying the local text reuse.

3.2.3. Run 3: Query Formulation Using Most Frequent Words

In this run the terms which were most frequent in the pre-processed paragraphs were used for query formulation. This is the most common strategy used in query formulation. In our approach the most frequent words for query formulation have been included just for comparison purposes and for analysing the efficiency of this strategy in obfuscated and translated locally reused text.

All of the above-mentioned pre-retrieval query formulation strategies are prompted by a set of source documents retrieved during initial experimental investigations. Query words in italics are proper nouns and underlined query words are Hapax under different obfuscation (Table 1).

3.3. Indexing and Retrieval

The corpus of the source documents is indexed using the Indri retrieval engine.[3] Retrieval on the indexed corpus is also done using Indri, which is based on the Inquery query language and uses an inference network (also known as a Bayesian network). Java platform (jdk1.7.0_07) using Indri was used for testing our algorithm and for retrieval of source documents.

3.4. Results and Evaluation

The performance of the two strategies of query formulation for a) proper nouns, b) unique words or Hapax, is analysed against the performance of the most commonly used method, i.e. queries formed using, c) most frequent words.

The complete process of heuristic retrieval is shown in Fig. 2.

4. EXPERIMENT

In this preliminary study, an attempt has been made to analyse proposed strategies on the training corpus of PAN CLEF 2012, which is 1.29 GB in size and contains 12,024 files divided into 8 sub-folders.

The dataset is comprised of six different sub-sets contained in folders named no-plagiarism, no-obfuscation, artificial-low, artificial-high, translation, and simulated paraphrasing. We dealt with the later four categories. We tried to formulate simple querying strategies with a view to retrieving a subset of potential documents. Formulating any querying strategies from the terms of English target documents with an aim to retrieve the original German and Spanish documents would fail, because although the English translated documents use the same script they have an altogether different vocabulary when compared to German or Spanish. This would require translation of both texts into one single language which is a tedious task for large corpora. In the proposed work we are not dealing with the complete task of text reuse detection. Our aim is only to reduce the size of the corpora on which state-of-the-art techniques can be used for retrieving the reused portions with further refined processes.

The PAN-CLEF training corpus was divided into 8 sub-folders (Fig. 3):

The corpus consists of:

/susp: suspicious documents as plain text.
/src : source documents as plain text.

The suspicious documents contain passages ‘plagiarized’ from the source documents, obfuscated with one of five different obfuscation techniques.

Furthermore, the corpus contains 6,000 XML files each of which report, for a pair of suspicious and source documents, the exact locations of the plagiarized passages (Fig. 4). The XML files are split into six datasets:

[3] Indri retrieval engine available at: http://sourceforge.net/projects/lemur/files/lemur/indri-5.4/

Table 1. Query Formulation using Proper Nouns and Hapax under Different Kinds of Obfuscation

SN	Reuse Type	Source docid	Reused portion in Source Document	Query Formulated Using Proper Noun	Query Formulated Using Hapax	Suspicious docid / Reused Text in Suspicious Document
1	Low Obfuscation	0180	Her singular story excited a considerable share of public attention; and she was engaged to sing, and perform the military exercises at various places of public entertainment: soon afterwards she married one *Eyles*, a carpenter at *Newbury*.	Eyles Newbury	singular story excited considerable exclaimed attention engaged military exercises various places entertainment subsequently married Eyles carpenter Newbury	00015/ “Her singular story excited a considerable share of public attention; and she was engaged to sing, and perform the military exercises of various places at public entertainment: soon subsequently us married one *Eyles*, a carpenter at *Newbury*.”
2	High Obfuscation	02320	An Epistle to Sir *Robert Walpole*. Three Poems; I. On the death of the late king; II. On the Accession of his present majesty. III. On the Queen. On the arrival of *Prince* Frederic. The origin of the Knights of the *Bath*, inscribed to the Duke of *Cumberland*.	*Enemy Robert Walpole Prince Antonio Soldiers City Mile Cumberland*	Enemy Robert Walpole king determined Iii Bath city Prince Antonio Soldiers City inscribed Mile Cumberland	00294/ “Been that Enemy *Robert* and *Walpole*. One Words; i did. nothing of king; three. By S is determined about present. Iii. On *Bath*. On city under *Prince Antonio*. The lake on the-- *Soldiers* up *City*, could be inscribed in *Mile* of *Cumberland*.”
3	Simulated Paraphrasing	03261	Fourteen ditches lined with sword-blades and poisoned chevaux-de-frise, fourteen walls bristling with innumerable artillery and as smooth as looking-glasses, were in turn triumphantly passed by that enterprising officer. His course was to be traced by the heaps of slaughtered enemies lying thick upon the platforms; and alas! by the corpses of most of the gallant men who followed him!--when at length he effected his lodgment, and the dastardly enemy, who dared not to confront him with arms, let loose upon him the tigers and lions of *Scindiah’s* menagerie. This meritorious officer destroyed, with his own hand, four of the largest and most ferocious animals, and the rest, awed by the indomitable majesty of BRITISH VALOR, shrank back to their dens.	*Figurehting Scindiah’s British Valor*	Figurehting trenches paced swords spears perched walls glowered guns path took fallen comrades heaped bodies officer dead came quarters attempted ovwerwhelm menagerie loosed strength arms slew wildests beast Seeing rest creatures amazed British Valor fled dens howling despair	01657/ “The reviewing officer paced victoriously by the fourteen Figurehting trenches lined with swords and spears; above them, perched upon fourteen walls glowered the guns. The path he took was lined with his fallen comrades, and with the heaped bodies of the enemy dead. It was when he came to his quarters that the enemy attempted one last time to ovwerwhelm him. The wild beasts of *Scindiah’s* menagerie were loosed upon him. By the strength of his arms alone he slew four of the wildests beast. Seeing this, the rest of the creatures, amazed by his *British Valor*, fled back to their dens in howling despair.”
4	Translation	03036	Diese wunderbare Einladung schreckte uns nicht ab, ihnen zu folgen. Zuerst ging der Stieg durch abgestuerzte *Kalkfelsenstuecke* hinauf, die durch die Zeit vor die steile Felswand aufgestufet worden und mit Hasel- und *Buchenbueschen* durchwachsen sind.	Kalkfelsenstuecke, Buchenbueschen	wonderful invitation scared staircase Kalkfelsenstuecke crashed aufgestufet period steep hazel Buchenbueschen mixed	00452/ “This wonderful invitation to us scared not depend, to follow them. First, the staircase went up by *Kalkfelsenstuecke* crashed, the aufgestufet been through the period before the steep rock face with hazel and *Buchenbueschen* are mixed.”

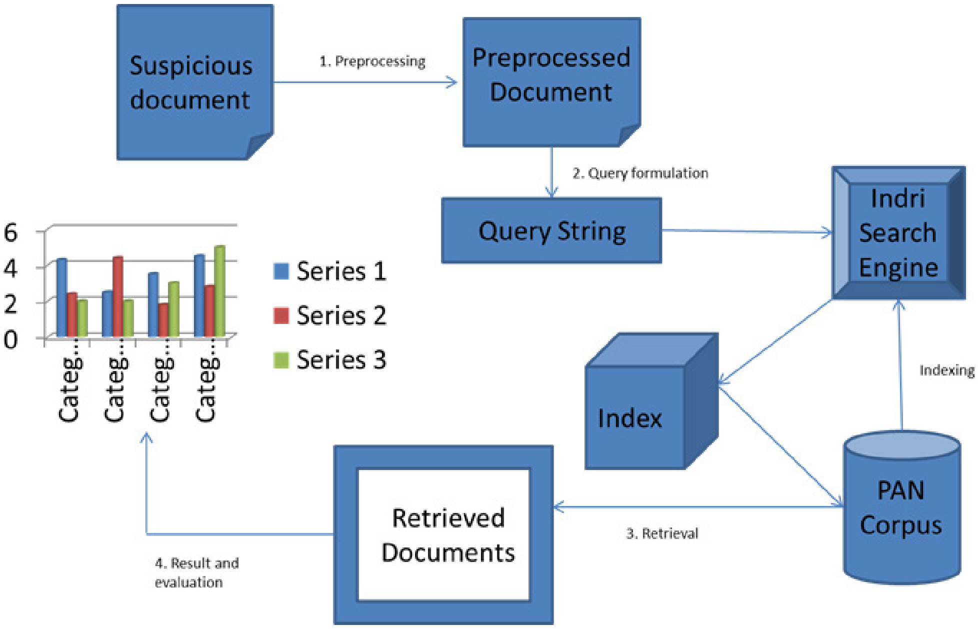
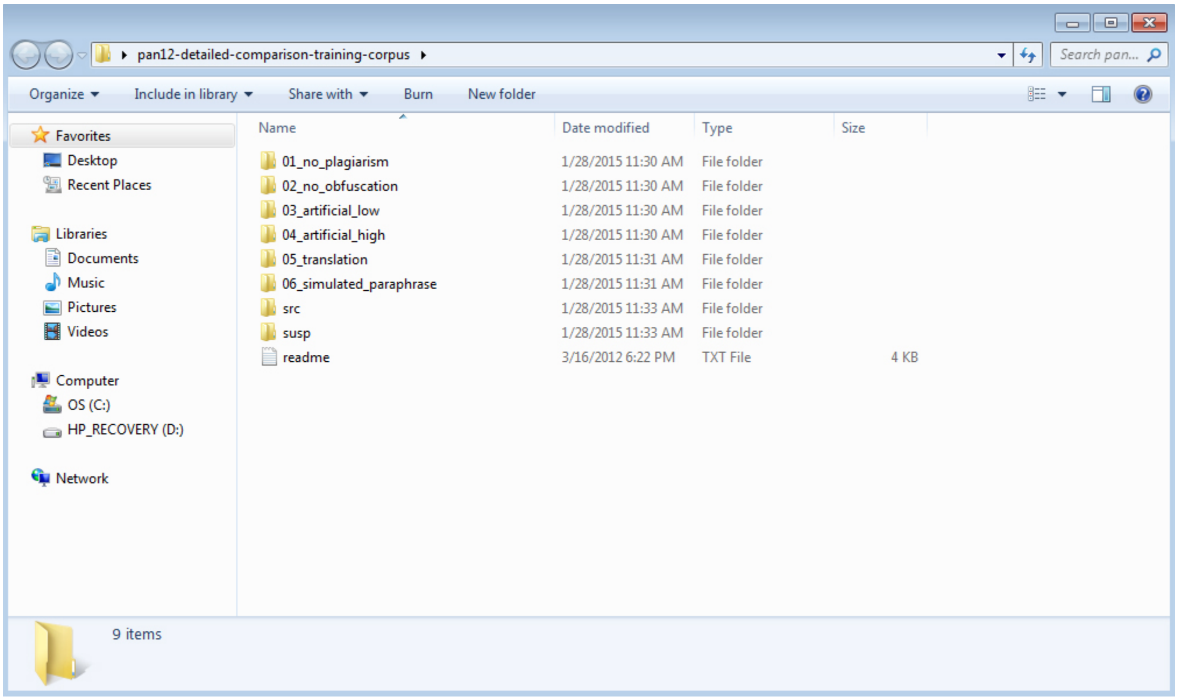


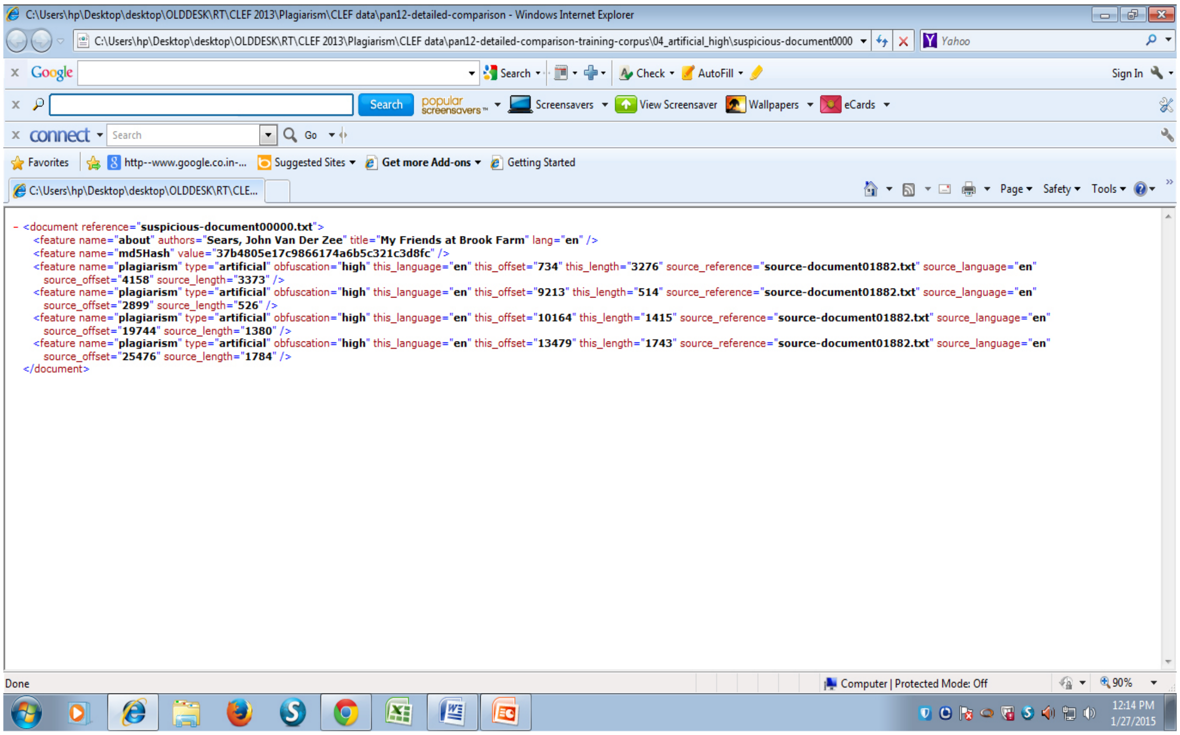
Fig. 2 Heuristic retrieval process



Name	Date modified	Type	Size
01_no_plagiarism	1/28/2015 11:30 AM	File folder	
02_no_obfuscation	1/28/2015 11:30 AM	File folder	
03_artificial_low	1/28/2015 11:30 AM	File folder	
04_artificial_high	1/28/2015 11:30 AM	File folder	
05_translation	1/28/2015 11:31 AM	File folder	
06_simulated_paraphrase	1/28/2015 11:31 AM	File folder	
src	1/28/2015 11:33 AM	File folder	
susp	1/28/2015 11:33 AM	File folder	
readme	3/16/2012 6:22 PM	TXT File	4 KB

9 items

Fig. 3 Sub folders of PAN-CLEF training corpus


C:\Users\hp\Desktop\desktop\OLDDESK\RT\CLEF 2013\Plagiarism\CLEF data\pan12-detailed-comparison - Windows Internet Explorer
C:\Users\hp\Desktop\desktop\OLDDESK\RT\CLEF 2013\Plagiarism\CLEF data\pan12-detailed-comparison-training-corpus\04_artificial_high\suspicious-document0000
Yahoo
Google
Search
Check
AutoFill
Sign In
Search
popular screensavers
Screensavers
View Screensaver
Wallpapers
eCards
connect
Search
Go
Favorites
http--www.google.co.in-...
Suggested Sites
Get more Add-ons
Getting Started
C:\Users\hp\Desktop\desktop\OLDDESK\RT\CLE...
Page
Safety
Tools
Done
Computer | Protected Mode: Off
90%
12:14 PM
1/27/2015

```
- <document reference="suspicious-document00000.txt">
    <feature name="about" authors="Sears, John Van Der Zee" title="My Friends at Brook Farm" lang="en" />
    <feature name="md5Hash" value="37b4805e17c9866174a6b5c321c3d8fc" />
    <feature name="plagiarism" type="artificial" obfuscation="high" this_language="en" this_offset="734" this_length="3276" source_reference="source-document01882.txt" source_language="en"
      source_offset="4158" source_length="3373" />
    <feature name="plagiarism" type="artificial" obfuscation="high" this_language="en" this_offset="9213" this_length="514" source_reference="source-document01882.txt" source_language="en"
      source_offset="2899" source_length="526" />
    <feature name="plagiarism" type="artificial" obfuscation="high" this_language="en" this_offset="10164" this_length="1415" source_reference="source-document01882.txt" source_language="en"
      source_offset="19744" source_length="1380" />
    <feature name="plagiarism" type="artificial" obfuscation="high" this_language="en" this_offset="13479" this_length="1743" source_reference="source-document01882.txt" source_language="en"
      source_offset="25476" source_length="1784" />
  </document>
```

Fig. 4 Example XML file of PAN-CLEF training corpus with report for a pair of suspicious and source documents and the exact locations of the plagiarized passages

/01_no_plagiarism: XML files for 1,000 document pairs without any plagiarism.

/02_no_obfuscation: XML files for 1,000 document pairs where the suspicious document contains exact copies of passages in the source document.

/03_artificial_low: XML files for 1,000 document pairs where the plagiarized passages are obfuscated by the means of moderate word shuffling.

/04_artificial_high: XML files for 1,000 document pairs where the plagiarized passages are obfuscated by the means of not-so moderate word shuffling.

/05_translation: XML files for 1,000 document pairs where the plagiarized passages are obfuscated by translation into a different language.

/06_simulated_paraphrase: XML files for 1,000 document pairs where the plagiarized passages are obfuscated by humans via Amazon Mechanical Turk.

The experiments were performed on low obfuscated, high obfuscated, simulated paraphrasing, and translation corpus texts.

It was observed that reused text in low obfuscation showed only minor changes in the text, and only a few of the words, mainly adverbs, were replaced in suspicious text. The reused text with high obfuscation had a comparatively larger number of words and even phrases replaced by synonyms, antonyms, and other similar phrases.

Simulated paraphrasing, although comparable, still was a difficult case to deal with, as whole texts were completely and intentionally paraphrased to hide the signs of reuse.

None of the authors are native speakers of any of the languages used in reused translated texts like German, Spanish, etc. but as the script was the same, the only action authors could do is to observe proper nouns appearing in both source and reused text. A snapshot of different kinds of obfuscation in the PAN CLEF 2012 training data set is shown in Table 2.

Table 2. Snapshot of Different Kinds of Obfuscation in PAN CLEF 2012 Training Data Set

Type of Reuse	Susp Doc id	Reused Text	Source Doc id	Original Text
Low Obfuscation	15	Her singular story excited a considerable share of public attention; and she was engaged to sing, and perform the military exercises of various places at public entertainment: soon subsequently us married one Eyles, a carpenter at Newbury.	1802	Her singular story excited a considerable share of public attention; and she was engaged to sing, and perform the military exercises at various places of public entertainment: soon afterwards she married one Eyles, a carpenter at Newbury
High Obfuscation	14	"That affair shall be shockingly," observed Life, "unless as short in we ourselves are beautiful, and ebony amigo have now done." I had observed that no first two or-- big knots occupying most intervals of face-scenes were evidently interested debate, and was being sorrow and is disappointed offspring. I should have liked to have put them all into rear, and had so have move to them did, could one have insured his not being intruded on black-man.	2151	"That ceremony will be quite superfluous," observed I, "unless as far as we ourselves are concerned,and our sable friends here."I had observed that the two or three little knots occupying the intervals of the side-scenes were evidently interested observers of our debate, and grieved and disappointed by the result.I should have liked to have put them all into the front, and then have acted to them, could one have insured their not being intruded on by any stray white-man.
Simulated paraphrasing	1661	A 23-year old widow was reportedly in Eastern Brooklyn, was starving herself but later was somehow given a power which by she could predict the future, read minds, and see happenings of different times and space, through penetration and other means. "She could well be the next Spiderman of Brooklyn, as far as this town knows." A local barbershop owner stated, noting the high crime raise since the market downfall. On October 20th, 1878, "Life without Food" was headed in a New York Newspaper.	1067	THE BROOKLYN CASE. For several years past there have been rumors more or less definite in character that a young lady in Brooklyn was not only living without food, but was possessed of some mysterious faculty by which she could foretell events, read communications without the aid of the eyes, and accurately describe occurrences in distant places, through clairvoyance or whatever other name may be applied to the influence. Finally, in the New York Herald of October 20th, 1878, appeared an account, headed "Life without Food
Translation	1277	The divine is offended by the Man idolized. Thus, for example, become a prey to the myth of Goethe's personality Goethe, and Goethe himself by a subjectivism that he came up and some of his genius constituted a cult of talking about the things of expression, personal detour and purpose and thus created an army of writers in the world, which may very well notice know about all the problems of life, are capable but very little, a place to stand up, to indulge themselves, that is to forget the thoughts in order to serve a cause. The writer as an apostle is to a degree conqueror, man of the will and the desire that He even gives his work a will, an addiction beyond art.	2637	der Gesellschaft nach Sachverschleierung gedient. Das Göttliche wird beleidigt, indem man den Menschen vergöttert. So ist z. B. der Mythos Goethe eine Beute der Persönlichkeit Goethe geworden, und Goethe selbst hat durch einen Subjektivismus, der ihm anstand und einen Teil seiner Genialität ausmachte, einen Kult des Redens über die Dinge, der Meinungsäußerung, der persönlichen Ausholung und Zwecksetzung und damit eine Armee von Literaten in die Welt gerufen, die sehr wohl Bescheid wissen über alle Probleme des Lebens, die aber sehr wenig vermögen, wo es gilt sich einzusetzen, sich hinzugeben, sich, d. h. die Meinung zu vergessen, um einer Sache zu dienen. Der Literat als Apostel ist bis zu einem Grad Eroberer, Mensch des Willens und der Sucht, daß er sogar seinem Werk einen Willen verleiht, eine Sucht über die Kunst hinaus.

It is observed that independent authors can create the same short sentences rather than long, similar ones, which are less likely to be similar by chance (Gustafson et al., 2008). Therefore, to avoid false positives queries were formed and posed to retrieval engine only if three or more than three words of each type were extracted from each paragraph.

The comparison was made based on two kinds of result sets of retrieval: 1) when the number of retrieved documents per query is 5; and 2) when the number of retrieved documents per query is 1.

5. RESULTS AND DISCUSSION

In the training corpus of PAN-CLEF, in addition to the XML files, each folder contains a text file called 'pairs.' For the 1,000 document-pairs (XML files) in the folder, this file lists the filename of the suspicious and the source document in a row, separated by a blank (Fig. 5):

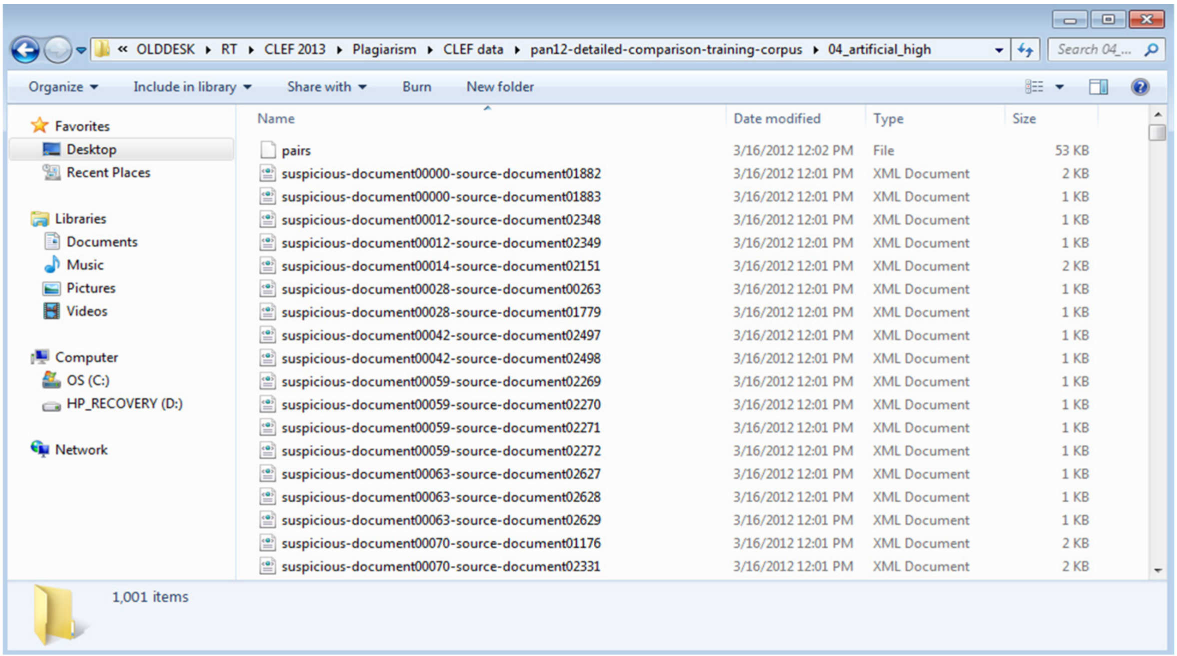
e.g.

suspicious-document00086.txt source-document00171.txt

A list of all of the relevant documents for each category was compiled using the relevance data provided by CLEF and the result of the experimentation was compared against this. As far as translated texts are concerned, they were in either German or Spanish, and none of the authors(s) are either native speakers of these languages or familiar with the vocabulary of these languages. So for the analysis of our results we only relied on the relevant source document list provided by CLEF for each suspicious document and have made that list our judgment criteria for pre-retrieval. Table 3 shows one such list for suspicious documents and source documents under translation dataset.

The average retrieval percentage summary of pre-retrieval query performance on the PAN CLEF 2012 training corpus is depicted in Table 5 and Figs. 6 and 7.

When query is formulated using proper nouns and results are recorded for five documents, the per query percentage of correct source documents retrieved is 88.34%, 71.50%, 52.85%, and 79.17%, respectively, in the cases of low, high obfuscation, simulated paraphrasing, and translation. In case of Hapax the percentage is 92.21%, 84.17%, 59.68%, and 52.50% (Table 5 and Fig. 6).



Name	Date modified	Type	Size
pairs	3/16/2012 12:02 PM	File	53 KB
suspicious-document00000-source-document01882	3/16/2012 12:01 PM	XML Document	2 KB
suspicious-document00000-source-document01883	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00012-source-document02348	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00012-source-document02349	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00014-source-document02151	3/16/2012 12:01 PM	XML Document	2 KB
suspicious-document00028-source-document00263	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00028-source-document01779	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00042-source-document02497	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00042-source-document02498	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00059-source-document02269	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00059-source-document02270	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00059-source-document02271	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00059-source-document02272	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00063-source-document02627	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00063-source-document02628	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00063-source-document02629	3/16/2012 12:01 PM	XML Document	1 KB
suspicious-document00070-source-document01176	3/16/2012 12:01 PM	XML Document	2 KB
suspicious-document00070-source-document02331	3/16/2012 12:01 PM	XML Document	2 KB

1,001 items

Fig. 5 Document-pairs (XML files): List of filenames of the suspicious and the source document in a row, separated by a blank

In the case of one document retrieved per query, the percentage is 81.55%, 66%, 36.18%, and 77.50%, respectively, for low obfuscation, high obfuscation, simulated paraphrasing, and translation when query is formulated using proper nouns, and the same is 86.43%, 76.83%, 41.24%, and 42.92%, respectively, for Hapax (Table 5 and Fig. 7).

Queries using most frequent words showed the worst performance in all cases. In the case of five documents retrieved per query it is 55.45%, 30.17%, 18.45%, and 36.25% for low obfuscation, high obfuscation, simulated paraphrasing, and translation, respectively (Table 5 and Fig. 6), and in the case of one document per query performance is further degraded with only 47.35%, 19.17%, 6.14%, and 31.67% retrieval of source documents, respectively, for different obfuscation levels (Table 5 and Fig. 7).

The Indri retrieval results show that queries formed with proper nouns and Hapax outperformed the most frequent words query formulation strategy (Table 4). In the cases of low and high obfuscation and simulated paraphrasing, keywords with Hapax proved to be slight-

Table 3. Compiled List of Source Documents for Each Target Document under Translation Dataset

Translation					
Susp-doc-id	**No. of Source documents**				
11	2635	2636			2
64	2659	2764	2871		3
104	686	2656	2742	2773	4
152	2924				1
182	2917				1
203	2717				1
264	2855				1
280	2718	2772	2990		3
303	2660				1
372	2720	2740	2835		3
396	2672	2916	2988	2989	4
452	2878	3035	3036	3037	4
459	2691	2861	2862		3
1276	2634				1
1277	2637	2638			2
1278	80	2639	2640		3
1279	2641	2642			2
1280	2643				1
1281	2644	2645	2646	2647	4
1282	2648	2649			2

ly more efficient. Queries formed using proper nouns performed exceptionally well in the case of Translated Local Text Reuse. In some of the cases the latter strategy retrieved all documents whereas the former could not retrieve even a single document. Similarity scores of proper noun queries were higher than those of the other two methods in most of the cases, even when the source document ranking was the same (Table 4). Most of the words of proper noun queries are also found in Hapax queries for the same paragraph in a query (Table 4). Proper noun queries are crisp and concise whereas Hapax queries are long (Table 4). The criterion that any

Table 4. Comparison of Queries Formulated and used (Suspicious Document id:00000 Source Document ids: 01882, 01883)

SN	Category	Queries	Rank-wise Result	Score and File size in bytes
1	Proper noun	James Simpson Antiquaries	source-document01882.txt source-document01953.txt source-document01270.txt source-document00851.txt source-document01050.txt	-6.354200718010045 6441 -7.6762179564722155 4277 -8.53381409288211 5099 -8.59352578372967 19299 -8.610674504938267 13541
	Hapax	prosecute James Simpson element attention unfinished argumentation papers yard observation Antiquaries	source-document01882.txt source-document01953.txt source-document01959.txt source-document03635.txt source-document01298.txt	-9.028233506404897 6441 -9.678877861391424 4277 -9.7500059759488 1045 -9.828335246625025 10850 -9.828335246625025 10850
	Most frequent	occasionally	Not used	
2	Proper noun	Vecta Victuarii Bede Hengist Vetta Baeda Kent Islet Cantuarii Lessons Pot Person	source-document01882.txt source-document01887.txt source-document00354.txt source-document00088.txt source-document02260.txt	-9.750870421016652 6441 -12.303098381165826 65750 -13.125661628692807 5721 -13.486182164022381 699 -13.493166903777693 596
	Hapax	1 trusty lasting mt 2 labors smudge stock possibly atom context validation assisted hall nonaccomplishment disapproval abstractor happening section Bede siemens etymology Hengist history kin chagrin held demo course Vetta relation alter Baeda wanted amp babu Kent Islet Cantuarii defamation Lessons enchantment Pot Person antioxidant asiatic monad author sentiment immature reshuffle paper problematical given	source-document00495.txt source-document00712.txt source-document03622.txt source-document00014.txt source-document03466.txt	-11.790930339629977 568 -11.808559151551538 491 -11.87914428216741 26 -11.88033122592693 29 -11.88712259002039 652
	Most frequent	indicate descend Vecta import horsa state Victuarii	source-document01882.txt source-document01887.txt source-document02537.txt source-document01757.txt source-document00026.txt	-8.1154956919508 6441 -10.587797646913668 65750 -11.70450589357424 1852 -11.718357995820861 8126 -11.779193051404548 704
3	Proper noun	Ven Gott Saturdays Iodine	source-document01883.txt source-document03068.txt source-document03073.txt source-document02740.txt source-document02105.txt	-9.411390970245929 7333 -10.677457576722492 531 -10.716320677272764 614 -10.718245605682348 620 -10.7709554797222 1406
	Hapax	clink Ven Gott occupy prize Saturdays community acid regretful large boat conflagration Iodine ship 131 vay wednesdays seasons cohort shown ammunition moder	source-document01883.txt source-document04112.txt source-document01180.txt source-document01658.txt source-document01606.txt	-10.51488499328055 7333 -10.705161013504394 2145 -10.746800582155643 1781 -10.785107898485926 683 -10.806178830219498 484
	Most frequent	state	Not used	

word which starts with an uppercase letter is considered to be a proper noun has been applied, so a few other terms, mainly nouns which start a sentence, are mistakenly included in the query.

Query performances are comparable in the cases of proper nouns and Hapax (Figs. 6 and 7), and retrieval scores are higher in the case of proper noun queries; therefore proper noun queries may be preferred over Hapax queries as those formed using proper nouns also reduce the chances of getting query drift.

Table 5. Table Showing Average Retrieval Percentage Summary

Method used→	Documents per query (Document=5)			Documents per query (Document=1)		
Type of reuse↓	Proper nouns	Hapax (frequency=1)	Most frequent	Proper nouns	Hapax (frequency=1)	Most frequent
Low Obfuscation	88.34%	92.21%	55.45%	81.55%	86.43%	47.35%
High Obfuscation	71.50%	84.17%	30.17%	66.00%	76.83%	19.17%
Simulated Paraphrasing	52.85%	59.68%	18.45%	36.18%	41.24%	6.14%
Translation	79.17%	52.50%	36.25%	77.50%	42.92%	31.67%

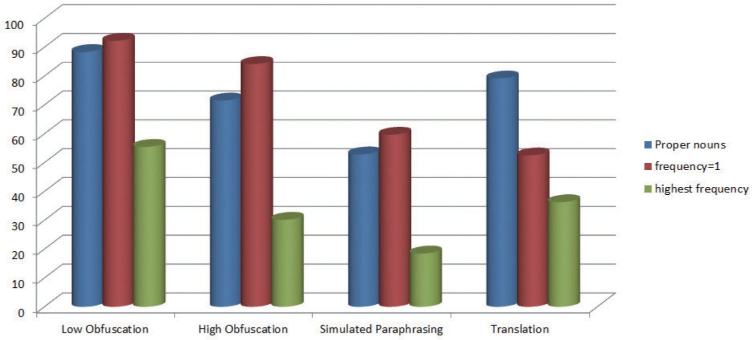


Fig. 6 Graph showing query performances (as average percentage) when the result considered per query is 5 documents

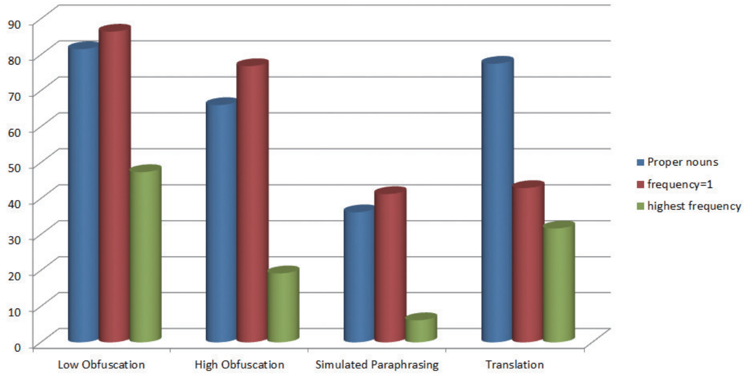


Fig. 7 Graph showing query performances (as average percentage) when the result considered per query is 1 document

6. CONCLUSIONS

The results of these experiments show that a pre-retrieval strategy of proper nouns and Hapax outperformed a most frequent words strategy. Initial study reveals that level of obfuscation may also have an influence on pre-retrieval strategy. Whereas Hapax was observed to be slightly more efficient than other strategies in the cases of low obfuscation, high obfuscation, and simulated paraphrasing, queries formulated using proper nouns were definitely the most efficient in the case of heuristic retrieval for local reuse in translated texts. The Heuristic retrieval strategies that were somewhat more efficient require further study on mono- and cross-lingual text reuse with different scripts. Further work is in progress as these are intermediate results of the experiments performed on the PAN CLEF 2012 training data set.

ACKNOWLEDGEMENTS

The authors are grateful to the organisers of PAN CLEF for allowing them to work on the PAN CLEF 2012 training corpus.

One of the authors, Aarti Kumar, is thankful to the Maulana Azad National Institute of Technology, Bhopal, India for providing her with the financial support to pursue her doctoral work as a full-time research scholar.

The authors are also thankful to the Indri group for providing us with the software to perform our experiments.

REFERENCES

Aggarwal, N., Asooja, K., Buitelaar, P., Polajnar, T.,& Gracia, J. (2012). Cross-lingual linking of news stories using ESA. In FIRE 2012Working Notes for *CL!NSS, FIRE* ISI, Kolkata, India(2012)

Arora, P., Jones, J., & Jones, G.J.F. (2013). DCU at FIRE 2013. Cross-Language Indian news story search. In *FIRE* 2013 Working Notes.

Bar, D, Zesch, T., and Gurevych, I. (2012, December). Text reuse detection using a composition of text similarity measures. *Proceedings of COLING 2012: Technical Papers* (pp. 167-184), COLING 2012, Mumbai.

Barrón-Cedeño, A. (2010, July). On the mono- and cross-language. Detection of text re-use and plagiarism. Paper presented at *SIGIR'10*, Geneva, Switzerland. ACM 978-1-60558-896-4/10/07.

Barrón-Cedeño, A. (2012). On the mono- and cross-language detection of text re-use and plagiarism. Ph.D. thesis. Universitat Politecnica de Valencia, Spain.

Clough, P.D., Gaizauskas, R., Piao, S.S.L., et al. (2002, July). METER: Measuring text reuse. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL) (pp. 152-159), Philadelphia.

Clough, P.D. (2001). Measuring text reuse in journalistic domain. *Proeeedings of the 4th CLUK Colloquium*. (pp. 53–63), UK.

Cummins, R., Jose, J., & O'Riordan, C. (2011, July). Improved query performance prediction using standard deviation. Paper presented at *SIGIR'11*, Beijing, China. ACM 978-1-4503-0757-4/11/07.

Ghosh, A., Pal, S., & Bandyopadhyay, S. (2011). Cross-language text re-use detection using information retrieval. In *FIRE* 2011 Working Notes.

Gipp, B., et al. (2013, July-August). Demonstration of citation pattern analysis for plagiarism detection. Paper presented at *SIGIR'13*, Dublin, Ireland. ACM 978-1-4503-2034-4/13/07

Grozea, C., & Popescu, M. (2009). ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, & E. Agirre (Eds.), *PAN'09* (pp. 10-18), Donostia, Spain.

Gupta, P., & Rosso, P. (2012, July). Text reuse with ACL (Upward) trends. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* (pp. 76–82), Jeju, Korea.

Gupta, P., & Singhal, K. (2011). Mapping Hindi-English text re-use document pairs. In *FIRE 2011* Working Notes.

Gustafson, N., & Soledad, M., et al. (2008). Nowhere to hide: Finding plagiarized documents based on sentence similarity. Paper presented at *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Provo, Utah, USA. 978-0-7695-3496-1/08 IEEE, DOI 10.1109/WIIAT.2008.16.

Hagen, M., & Stein, B. (2010). Candidate document retrieval for web-scale text reuse detection? Extended version of an ECDL 2010 poster paper. M. Hagen & B. Stein. Capacity-constrained query formulation. *Proc. of ECDL 2010* (posters) (pp. 384–388).

Haiduc, S., et al. (2013). Automatic query reformulations for text retrieval in software engineering. Paper presented at *2013 IEEE ICSE 2013*, San Francisco, CA, USA.

Carmel, D., et al. (2006, August). What makes a query difficult? Paper presented at *SIGIR'06*, Seattle, Washington, USA.

Hauff, C., Hiemstra, D., & Jong, F. (2008, October). A survey of pre-retrieval query performance predictors. Paper presented at *CIKM'08*, Napa Valley, CA, USA. ACM 978-1-59593-991-3/08/10.

Mittelbach, A., Lehmann, L., Rensing, C., et al. (2010, September). Automatic detection of local reuse. *Proceedings of the 5th European Conference on Technology Enhanced Learning no. LNCS 6383* (pp. 229-244). Berlin Heidelberg: Springer-Verlag.

Osman, A.H., et al. (2012). An improved plagiarism detection scheme based on semantic role labeling. *Journal of Applied Soft Computing, 12*, 1493-1502. doi:10.1016/j.asoc.2011.12.021

Pal, A., & Gillam, L. Set-based similarity measurement and ranking model to identify cases of journalistic text reuse. In *FIRE 2013* Working Notes.

Palkovskii, Y., Muzyka, I., & Belov, A. (2012). Detecting text reuse with ranged windowed TF-IDF analysis method. Retrieved from http://www.plagiarismadvice.org/research-papers/item/detecting-text-reuse-with-ranged-windowed-tf-idf-analysis-method

Palkovskii, Y., & Belov, A. (2011). Using TF-IDF weight ranking model in CLINSS as effective similarity measure to identify cases of journalistic text reuse. Berlin Heidelberg: Springer-Verlag.

Possas, B., Ziviani, N., Ribeiro-Neto, B., et al. (2005, October-November). Maximal termsets as a query structuring mechanism. Paper presented at *CIKM'05*, Bremen, Germany. ACM 1595931406/05/0010.

Potthast, M., Hagen, M., Gollub, T., et al. (2013, September). Overview of the 5th International Competition on plagiarism detection. Working notes paper presented at *CLEF 2013 Evaluation Labs and Workshop*, Valencia, Spain.

Potthast, M., Hagen, M., Völske, M., et al. (2013, August). Crowdsourcing interaction logs to understand text reuse from the web. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers) (pp. 1212-1221).

Pouliquen, B., et. al. (2003). Automatic identification of document translations in large multilingual document collections. *Proc. International Conference Recent Advances in Natural Language Processing (RANLP '03)*, pp. 401-408.

Seo, J., & Croft, W.B. (2008, July). Local text reuse detection. Paper presented at *SIGIR'08*, Singapore. ACM 978-1-60558-164-4/08/07.

Tholpadi, G., & Param, A. (2013). Leveraging article titles for cross-lingual linking of focal news events. In *FIRE 2013* Working Notes.

Torrejon, D.A.R., & Ramos, J.M.M. (2013). Linking English and Hindi news by IDF, reference monotony and extended contextual N-grams IR engine. In *FIRE 2013* Working Notes.

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-based word alignment in statistical translation. Proc. *16th conference on Computational linguistics (COLING '96), Association for Computational Linguistics* (vol. 2, pp. 836-841). doi:10.3115/993268.993313