Persian Plagiarism Detection Using Sentence Correlations

Muharram Mansoorizadeh
Department of Computer engineering
Bu-Ali Sina University, Hamedan, Iran
mansoorm@basu.ac.ir

Taher Rahgooy
Department of Computer engineering
Bu-Ali Sina University, Hamedan, Iran
taher.rahgooy@gmail.com

ABSTRACT
This report explains our Persian plagiarism detection system which we used to submit our run to Persian PlagDet competition at FIRE 2016. The system was constructed through four main stages. First is pre-processing and tokenization. Second is constructing a corpus of sentences from combination of source and suspicious document pair. Each sentence considered to be a document and represented as a tf-idf vector. Third step is to construct a similarity matrix between source and suspicious document. Finally the most similar documents which their similarity is higher than a specific threshold marked as plagiarized segments. Our performance measures on the training corpus were promising (precision=0.914, recall=0.848, granularity=3.85).

CCS Concepts
• Information systems→Information retrieval →Retrieval tasks and goals. Near-duplicate and plagiarism detection.

Keywords
Persian PlagDet, Plagiarism detection, document retrieval, tf-idf

1. INTRODUCTION
Plagiarism in academia is rising and multiple authors have worked to describe these phenomena [11]. As commented by Hunt in [7] , "Internet Plagiarism" is referred sometimes as a consequence of the "Information Technology revolution", as it proves to be a big problem in academia. According to Park [11], plagiarism is analyzed from various perspectives and considered as a problem that is growing over time. To tackle this problem, the most common approach so far is to detect plagiarism using automated algorithms based on text processing and string matching algorithms.

Two main strategies for plagiarism detection have been considered by researchers, namely, Intrinsic and external plagiarism detection [17] [5]. Intrinsic plagiarism detection aims at discovering plagiarism by examining only the input document, deciding whether parts of the input document are not from the same author. External plagiarism detection is the approach where suspicious documents are compared against a set of possible references. From exact document copy, to paraphrasing, different levels of plagiarism techniques can be used in several contexts, according to Meyer zu Eissen [17].

For external plagiarism detection Stein, Meyer zu Eissen, and Potthast [15] introduce a generic three-step retrieval process. The authors consider that the source of a plagiarism case may be hidden in a large reference collection, as well as that the detection results may not be perfectly accurate. Figure 1 illustrates this retrieval process. In fact, all detection approaches submitted by the competition participants can be explained in terms of these building blocks [5].

The process starts with a suspicious document d_q and a collection D of documents from which d_q's author may have plagiarized. Within a so-called heuristic retrieval step a small number of candidate documents, D_x, which are likely to be sources for plagiarism, are retrieved from D. Note that D can be as large as the entire web. Hence, it is impractical to compare d_q with all of its members. Usually an initial inspection is made to select a rough subset of D as prospective candidates. In a so-called detailed analysis step, d_q is compared section-wise with the retrieved candidates. All pairs of sections (s_q, s_x) with $s_q \in d_q$ and $s_x \in d_x$, $d_x \in D_x$, are to be retrieved such that s_q and s_x have a high similarity under some retrieval model. In a knowledge-based post-processing step those sections are filtered for which certain exclusion criteria hold, such as the use of proper citation or literal speech. The remaining suspicious sections are presented to a human, who may decide whether or not a plagiarism offense is given.

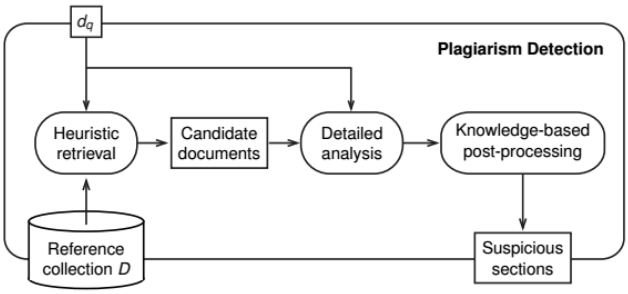

Figure 1: Generic retrieval process for external plagiarism detection [5].

Given a set of suspicious and source documents written in Persian, the plagiarism detection task of Persian PlagDetect competition [2] at FIRE 2016, as an external plagiarism detection setup, is to find all plagiarized sections in the suspicious documents and, if available, the corresponding source sections. The problem consists of a set of source and suspicious documents and a list of source-suspicious pairs to evaluate. The challenge is to find and locate plagiarized fragments in each suspicious and source document for each pair. We tackle this challenge by using a four-step method which is described in the following.

2. RELATED WORK
Text plagiarism is a long lasting game between authors and publishers [8]. Automatic plagiarism detection has a story as long as modern machine learning and text mining [15]. Plagiarism in

simplest case can be exact copying of others' published or unpublished works. More advanced cases try to paraphrase stolen ideas in different words [1]. Studies confirm that usually plagiarists copy and paste the original text or just paraphrase it keeping most of the text's keywords [16]. Text mining methods can reveal these simpler but widely used cases of plagiarism [10]. Deep semantic analysis of texts is studied for detecting more advanced paraphrasing cases [3].

3. PROPOSED ALGORITHM

In the following we describe the steps we took in order to find plagiarized segments of the document pairs.

3.1 Motivation

In her MSc thesis [16], Zahra Taheri has investigated many scientific plagiarism cases. Most of the cases were near-copy or slightly modified versions of the original sentences. Any scientific field has a common and widely accepted terminology that the related community adopts in writing papers, books and other types of publications. For example in machine learning literature the term *feature* is used for describing attributes of objects. As a dictionary term *property* is synonymous to *feature* but no one uses *property* and *property extraction* instead of *feature* and *feature extraction* in a scientific writing that discusses classification and pattern recognition. Hence, a plagiarized text inherits many of the main terms of the original text.

Original sentence
نقطه جوش آب صد درجه سانتیگراد است
Boiling Point of Water is one hundred Degrees.
Plagiarized sentences
نقطه جوش آب صد درجه سلسیوس است
Boiling Point of Water is one hundred Celsius degrees.
نقطه جوش آب ۱۰۰ درجه سانتیگراد است
Boiling Point of Water is 100 degrees.
آب در صد درجه سانتیگراد به جوش می آید
Water boils at one hundred degrees
اگر دمای آب به صد درجه سانتیگراد برسد، می جوشد
If temperature of water reaches one hundred degrees, it will boil
اگر آب را تا صد درجه حرارت دهید می جوشد
If you heat water up to one hundred degrees, it boils.

Figure 2 a syntactic example of a persian sentence and its possible plagiarized versions

A syntactic example is demonstrated in Figure 2. Key terms in the original sentence are *water, boiling*, and *one-hundred*. To communicate the same concept, plagiarized versions of the sentence had to use these terms with re-ordering and /or changing verb tenses. This fact narrows down the task of plagiarism detection to classic document matching and retrieval.

3.2 Preprocessing and Feature Extraction

Document retrieval task is defined as: [9]

Given a document collection D and a query document, q, find it's most similar documents in D.

The task is usually solved by representing documents in some *vector space* and exploiting similarities between the query document and the collection members in this space. Bag of words representation of documents is a classic, yet powerful method for document indexing and retrieval. In this method a document is assumed to be the set of its terms, ignoring their relative positions in paragraphs and sentences. A dictionary is the set of all the distinct terms in the collection. A document, *d,* is represented by a one-dimensional vector, *v,* in which v_i denotes relative importance of the i-th dictionary term in *d.* Using this representation, similarities between documents can be estimated by well-known techniques such as Euclidean distance or cosine of their respective vectors. Usually v_i is defined as function of appearances of the word in the document and the whole collection. Term frequency (TF) of term in a document is the relative frequency of the term in the document.

$$\boldsymbol{TF_i = \frac{F_i}{|d| + 1}} \qquad \textbf{Eq 1}$$

In this equation F_i is the frequency of i-th term in the document and |.| denotes document length; i.e. total number of terms in the document. Assuming that there totally N documents in the collection and N_i of them contains i-th term, inverse document frequency (IDF) is defined by equation Eq 2

$$\boldsymbol{IDFi = \log(\frac{N}{N_i + 1})} \qquad \text{Eq 2}$$

Finally, as an importance measure of the word in a document, tf-idf is defined as:

$$\boldsymbol{TFIDF_i = TF_i . IDF_i} \qquad \text{Eq 3}$$

tf-idf favors terms with high frequency in the document but low frequency in the collection. Common terms such as *it, is,* and *what* that appear in most of the documents will have a low tf-idf value. These so-called *stop words* cannot distinguish the documents efficiently; hence usually are removed in pre-processing steps of NLP tasks.

We adopt tf-idf for representation and retrieval of matching documents. Each source and suspicious document is firstly tokenized and then split to sentences. This step is performed using NLTK 3.0 toolkit [4]. Then, each sentence is treated as an individual document. All the sentences of source and suspicious documents constitute document collection.

3.3 Similarity matrix construction

In this step, the similarity matrix between source sentences and suspicious sentences is calculated. If there are N source sentences and M suspicious sentences, the similarity matrix *S,* has N×M elements in which $S_{i,j}$ denotes the similarity between sentence i from source document and sentence j from suspicious document.

There are many different similarity and distance measures. For example an option is to use Euclidian distance, which measures the distance between two vectors *u* and *v* in the encompassing Euclidian space. In order to convert the distance to a similarity measure we can use:

$$sim_{Euclidian}(u, v) = \frac{1}{\|u - v\| + 1} \quad \text{Eq 4}$$

Another choice that we used is cosine similarity:

$$\cos(u, v) = \frac{u.v}{\|u\|\|v\|} \quad \text{Eq 5}$$

This measure can range from -1 to 1. When the value is 1, it means that *u* and *v* are in the same direction and when it is -1, it means that *u* and *v* are in the opposite directions. We use cosine similarity in subsequent steps. Initial inspection and evaluation confirmed that it has slightly better results than Euclidean distance.

3.4 Finding plagiarized fragments

We used pairs of sentences which their similarity was greater than a pre-specified threshold as plagiarized fragments for both source and suspicious documents. The value we used as a threshold was 0.4, which is obtained through cross validation.

4. EVALUATION

The training results were obtained after we run our application on TIRA platform [6] [13]. The evaluation measures used for this task are precision, recall, and granularity. Another measure called *plagdet* used to combine the aforementioned measures in order to enable us to sort the results for all algorithms and compare them in more objective way [12] [14].

The detection granularity of detected documents collection *R* under source documents collection *S* is defined as:

$$gran(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s| \quad \text{Eq 6}$$

Where $S_R \subseteq S$ are cases detected by detections in R, and $R_s \subseteq R$ are the detections of a given s:

$$S_R = \{s | s \in S \wedge \exists r \in R : r \ detects \ s\}$$

$$R_s = \{r | r \in R \wedge r \ detects \ s\}$$

The domain of gran(S, R) is [1, |R|], with 1 indicating the desired one-to-one correspondence and |R| indicating the extreme case, where a single s ∈ S is detected multiple times.

Precision, recall, and granularity allow for a partial ordering among plagiarism detection algorithms. To obtain an absolute order they must be combined to an overall score:

$$plagdet(S, R) = \frac{F}{\log_2(1 + gran(S, R))} \quad \text{Eq 7}$$

Where F denotes the F-Measure, i.e., the weighted harmonic mean of precision and recall. Log of granularity decreases its impact on the overall score [14]. The results obtained by our algorithm are showed in table 1.

Table 1. Evaluation results for proposed algorithm

Threshold	Precision	Recall	granularity	plagdet
0.4	0.914	0.848	3.859	0.385
0.5	0.821	0.926	4.481	0.354

5. CONCLUSION

In this paper, we proposed a sentence-level algorithm based on tf-idf features for plagiarism detection task, which shows competitive results on the training data.

The algorithm works is designed for *near-copy* and *paraphrasing* types of plagiarism. It relies on the fact that a plagiarist willing to publish an article in a scientific field must use popular terminology of the field. Obviously sophisticated cases such as cross language plagiarism or grabbing another ones idea and discussing it in one's own words would hard to be detected by this algorithm. In the future works we will consider improving the feature vector of the sentences by incorporating more features and also to use a method to combine overlapping fragments. Furthermore we will study language modeling and semantic text normalization for detecting harder cases of plagiarism.

6. REFERENCES

[1] Alzahrani, S.M., Salim, N. and Abraham, A., 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(2), pp.133-149.

[2] Asghari, H., Mohtaj, S., Fatemi, O., Faili, H., Rosso, P., and Potthast, M., 2016. Algorithms and Corpora for Persian Plagiarism Detection: Overview of PAN at FIRE 2016. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[3] Barrón-Cedeño, A., Vila, M., Martí, M.A. and Rosso, P., 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, *39*(4), pp.917-947.

[4] Bird, S., 2006, July. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 69-72). Association for Computational Linguistics.

[5] Potthast, M., Stein, B., Eiselt, A., Barron, Cedeno, A., and Rosso, P., 2009. Overview of the 1st international competition on plagiarism detection. In *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse* (p. 1)

[6] Gollub, T., Stein, B. and Burrows, S., 2012, August. Ousting ivory tower research: towards a web framework for providing experiments as a service. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 1125-1126). ACM..

[7] Hunt, R., 2003. Let's hear it for internet plagiarism. *Teaching Learning Bridges*, *2*(3), pp.2-5.

[8] Mali, Yaser. 2011. Crib from the novelist Or thirteen best practices for how to do a clean plagiarism (in Persian).

[9] Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich. *Introduction to information retrieval*. Cambridge University Press, 2008.

[10] Oberreuter, G. and VeláSquez, J.D., 2013. Text mining applied to plagiarism detection: The use of words for

detecting deviations in the writing style. *Expert Systems with Applications*, *40*(9), pp.3756-3763.

[11] Park, C., 2003. In other (people's) words: Plagiarism by university students--literature and lessons. *Assessment & evaluation in higher education*, *28*(5), pp.471-488.

[12] Potthast, M., Eiselt, A., Barrón Cedeño, L.A., Stein, B. and Rosso, P., 2011. Overview of the 3rd international competition on plagiarism detection. In *CEUR Workshop Proceedings*. CEUR Workshop Proceedings.

[13] Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E. and Stein, B., 2014, September. Improving the Reproducibility of PAN's Shared Tasks. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 268-299). Springer International Publishing.

[14] Potthast, M., Stein, B., Barrón-Cedeño, A. and Rosso, P., 2010, August. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 997-1005). Association for Computational Linguistics.

[15] Stein, B., zu Eissen, S.M. and Potthast, M., 2007, July. Strategies for retrieving plagiarized documents. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 825-826). ACM.

[16] Taheri, Zahra. 2012. *Plagiarism Detection in Scientific Texts using Structural and Semantic Relations*.

[17] Zu Eissen, S.M., Stein, B. and Kulig, M., 2007. Plagiarism detection without reference collections. In *Advances in data analysis* (pp. 359-366). Springer Berlin Heidelberg.