Graph-based Approach to Text Alignment for Plagiarism Detection in Persian Documents

Mozhgan Momtaz
Faculty of New Sciences and Technologies
University of Tehran, Tehran
m.momtaz92@ut.ac.ir

Kayvan Bijari
Faculty of New Sciences and Technologies
University of Tehran, Tehran
kayvan.bijari@ut.ac.ir

Mostafa Salehi
Faculty of New Sciences and Technologies
University of Tehran, Tehran
mostafa_salehi@ut.ac.ir

Hadi Veisi
Faculty of New Sciences and Technologies
University of Tehran, Tehran
h.veisi@ut.ac.ir

ABSTRACT

This paper presents a new approach for Persian plagiarism detection. This approach uses a graph structure as well as one of the graph similarity methods (iterative methods) for similarity detection of two Persian documents. In this approach, documents are represented by a graph with specified length, then each part of suspicious document is compared to that of the source document. The graph is made if these parts have more common bigrams than a predefined threshold. Once graphs are made, an iterative method is used to find analogous nodes in graphs. Two graphs are marked as similar if they contain at least a certain number of similar nodes. In order to evaluate the proposed method, it was run on PAN2015 and PAN2016 Persian Text Alignment dataset. The Plagdet score is defined to evaluate plagiarism detection methods in PAN contest. The gained Plagdet of proposed method is 90% on PAN2015 and 87% on PAN2016.

CCS Concepts

• **Information systems→ Plagiarism Detection software**
• **Computing methodologies→ Graph-based**

Keywords

Plagiarism Detection, Text Based Graph Representation, Text Alignment.

1. INTRODUCTION

Nowadays, a large volume of information is a compound of different types of contextual data such as books, articles and other documents and this volume is growing increasingly. In many cases, we need to identify either the duplicated documents or the ones which are near-copy documents among the many cases. In this regard, Plagiarism detection in documents is one of the main topics which gained attention among researchers in the recent years. The act of plagiarism is to use other author's writing or ideas, without proper appreciation to the author or proper citation to the original source [3]. In the recent years, identifying plagiarism has become easier using different systems, but different types of plagiarism is still an issue. In some types of plagiarism, the structure of the document is changed by rearranging the words or using synonym words. Therefore, the results of basic plagiarism detection methods are not acceptable. So the need for more sophisticated methods for plagiarism detection is growing. Different kinds of plagiarism are shown in figure 1.

Categories of text alignment dataset are based on PAN Competitions [12].

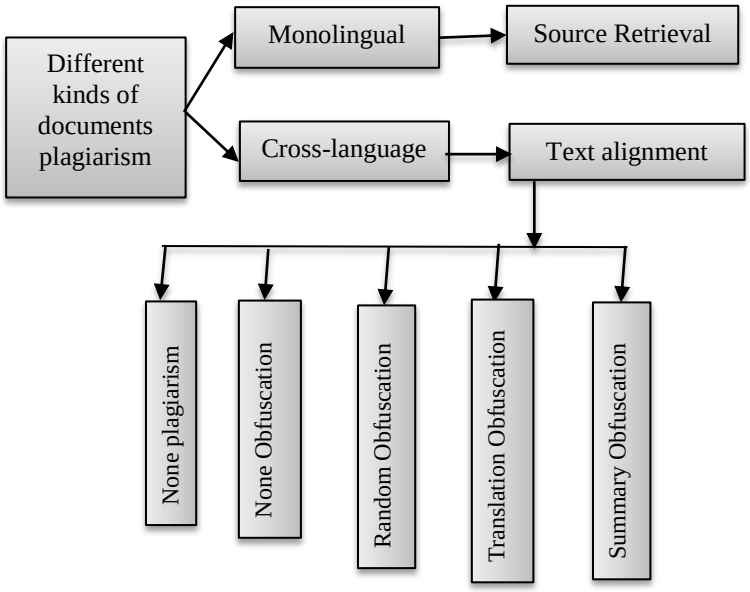

Figure 1- Different kinds of documents plagiarism

The proposed method, converts every document to a number of parts with specific length using the graph approach idea and then if necessary, it converts each part to a graph for precise plagiarism examination, which creates graph based on simultaneous occurrence of the words in fixed window size. After this step, the similarity between two graphs measured using node similarity measures, if the rate of similarity is more than the specified threshold, then that part is labeled as plagiarism. The proposed method was run on the PAN2015 dataset [6] and PAN2016 dataset (Persian Plagdet2016 contest [5]).

The rest of paper is organized as follows: Section 2 is devoted to related works; Section 3 presents the graph-based methodology for plagiarism detection. In Section 4, experimental evaluation of the proposed method is given, and finally Section 5 ends the paper with conclusion and future works.

2. RELATED WORK

In order to detect monolingual plagiarism, various methods have been proposed. In this section, each of these methods is explained briefly.

Character-based methods: The most important method is the fingerprint method. Fingerprint algorithms [20] consider the text as a series of characters and then they divide the characters in n-series groups, the most important algorithms include 16-gram, 8-gram and 5-gram methods. In this method, the degree of similarity depends on the number of similar characters in a string. Although this method ends up with a good result in detecting plagiarism, but when plagiarism has some paraphrasing or modified words, this method does not act in an effective way to detect plagiarism.

Structural-based methods: In the previous method, the only attention is on words as features of the document. However, in the structural-based methods [9], pays attention to the titles, paragraphs, sections and resources. One of the most famous structural-based methods is tree-based method, which gained much attention in the recent years. In the tree-structured model, a two-layer model is defined that the top layer is for retrieving documents and the bottom layer is considered for detecting plagiarism between retrieved documents using the methods of similarity detection such as cosine similarity method.

Classification-based methods: In this method the documents are classified based on specific words (or keywords) [19]. The primary goal in this approach is to retrieve similar documents and to speed up the process of plagiarism detection.

Semantic-Based methods: This method uses lexical network in order to find semantic similarity for plagiarism detection [17]. The most famous lexical network in English language is WordNet [8]. By means of WordNet, it is possible to achieve more information about a special word. This method is effective when plagiarism is done using synonym words. FarsNet [15] the Persian equivalent of WordNet, is also proposed for Persian language.

Graph-based methods: In this method, the text will be converted into a graph. Nodes in the graph can be words, phrases or even sentences of the text. Edges in the graph represent the link or relation between nodes and they can show the semantic link between words or the simultaneous occurrence in one sentence [16]. This method will be discussed more in the proposed method section. Converting a text into a graph, enables us to detect plagiarism using the advantages of graph similarity algorithms [7].

In this paper, the proposed method is a combination of structural-based method and graph-based method. Paying attention to the structure of the text leads to detection of plagiarism in the document even if the plagiarism is the type in which the structure of the text is changed.

One common and standard approach to model text document is bag-of-words. This model is suitable for capturing word frequency [16]. Assuming that order of the word's occurrence has nothing to do with its meaning; this model has a proper result in information retrieval. The drawback of this model is when it tries to find the reused text and plagiarism between different parts of the text, if a reused text is occurring by using synonym words then this model could not properly detect the reused text. Furthermore, this model doesn't express the meaning and the structure of the text [16]. However, Graph representation is mathematical constructs and can model different word's relation and textual structure of the documents [16]. Some issues of the bag-of-words model and the solutions based on graph model are summarized in table 1.

Table1- Some issues of the bag-of-words model and the solutions based on graph-based model for plagiarism detection applications

Issues of bag-of-words methods to model text for plagiarism detection applications	Graph-based solution
Ignoring order of the words	Using directed graph (step 3 of proposed method)
Ignoring the structure of the text	Considering the whole sentence as a graph (step 2 and 3 of proposed method)
Neglecting word synonyms	Ability to add synonyms to the graph corresponding text (step 3 of proposed method)

3. METHODOLOGY

Any textual document can be presented via a corresponding graph. Graph based representation of text is important because it enables us to turn an unstructured text into a structural text, and then the advantageous of graph representation can be applied to text summarization, identifying similarities of the documents, and applications of text mining. For natural language processing applications text graph of documents should be built. In a text graph, nodes represent words of the document, and the edges present the relation between different words. The relation of words can vary from application to application. The proposed plagiarism detection method is consist of 5 steps that will be discussed further in the following.

Step 1. preprocessing: normalization is one of the basic steps in text mining and text processing. In the normalization process, punctuations and stop words are removed. In this paper, we have used Hazm package [10] for Persian text normalization.

Step 2. turn text to set of clauses: suspicious document and the reference document are divided to a series of sentences. Each sentence of the suspected document will be compared to all the sentences of the reference document. In this step, a filtering will be done on sentences in order to reduce runtime. Finally, if the two sentences at least have the cosine similarity of 0.4, then they go to the step of graph making process (this value is obtained experimentally), otherwise the comparison will continue to other sentences of the reference document.

Step 3. creating corresponding graph: in graph creation step, each sentence will be converted to a graph. The nodes of the graph are main and unique words, and in this graph, an edge will be established between a specific word and 4 words before and after it in the document. Igraph package [11] is used for graph creation in this paper.

Step 4. plagiarism detection: when graph creation is complete, we are looking for nodes in the suspected document that is common with the node of the reference document. An iterative method based on simple idea indicates that the two graphs are similar when they have similar nodes, and nodes are similar if they have analogous neighbors [18]. We use this method for our specific graph. Then we find their similarity using equation 1:

$$similarity(A,B) = \frac{A \cap B}{\max(len(A), len(B))} \quad (1)$$

Where B is primary neighbors of the common node in the graph of reference document. And A is primary neighbors of the common node in the graph of suspicious document.

If the similarities between two nodes is greater than the threshold α (α = 0.4), then that node is selected as the similar node. Finally, if a sentence has more similar nodes than the threshold β (β=1/3(the number of key words in suspicious document)), that statement is labeled as one of the sentences which plagiarism has occurred. α and β are thresholds that are achieved experimentally and they are based on performances of the system.

Step 5. Integrate plagiarism labeled Sentences: In this step, we integrate sentences with plagiarism label (output of step 4 of the algorithm) based on start and end offset of sentences in text. This step important for granularity measure[1] [2]. If there exist no labeled sentences, we can assume that no plagiarism is occurred.

4. EXPERIMENTAL EVALUATION

In this section the results of the implementation of the proposed method on the plagiarism data sets are given. Moreover, in the following we are going to focus on analyzing results. The two data sets that were used for analyzing the proposed method are as follows.

1. Persian Text alignment dataset PAN2015: This dataset is published on the website of PAN contest.

2. Persian Text Alignment dataset PAN2016: This dataset contains 2749 training and testing documents and are related to Persian Plagdet2016 international contest [14, 4, 13], which was organized by the Institute of Information and Communication Technology (ICTRC) and contest results are available at the contest site.

4.1 Experimental Results

Table 2 shows the results of the implementation of the proposed method on the datasets.

Table2- Experimental Results on Persian document dataset

dataset	Precision	Recall	Palgdet
PAN2015	0.91	0.89	0.90
PAN2016(training)	0.90	0.89	0.89
PAN2016(contest)	0.89	0.85	0.87

As shown in the results, according to the evaluation criteria, the graph-based method has achieved favorable results without using linguistic corpora and only due to the structure of the text. Graph approach has unique features to detect similar documents. Among these features, one can mention paying attention to non-adjacent words in a sentence. This feature makes plagiarism detection easy, because plagiarism is done by rearranging words, but in the character-based methods attention is just on the relationship between adjacent words. Furthermore, another feature of the graph is considering the minimum threshold of similarity between the two nodes. In this case, if plagiarism is done by add and removal of the words, by considering minimum similarity threshold, these changes do not have much negative impact on plagiarism detection. The result of Persian Plagdet contest is also reported in [1].

1 The logarithm of the granularity to smooth its influence on the overall score.

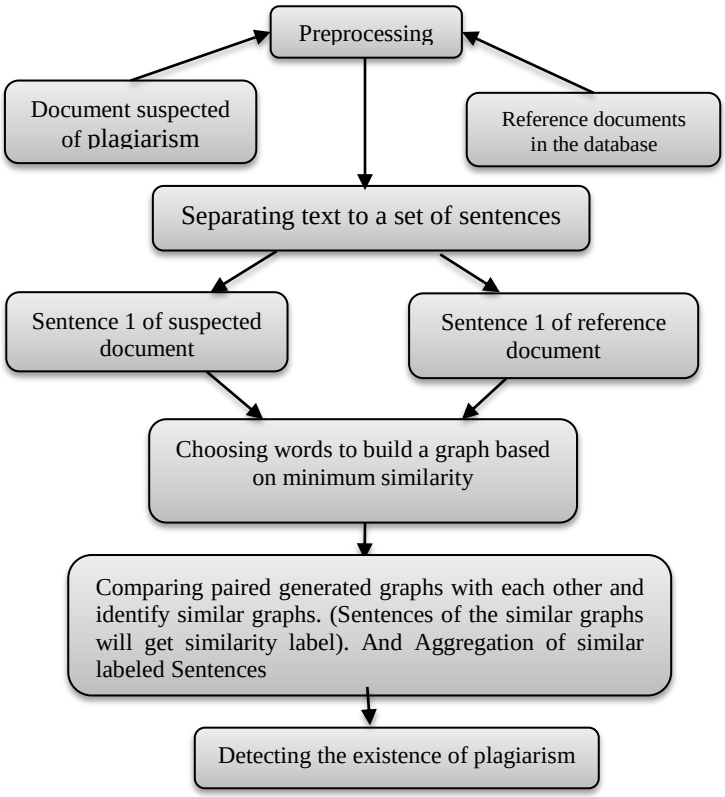

Figure 2- diagram of detecting plagiarism

5. CONCLUSION AND FUTURE WORK

Using a graph structure, we proposed a method to convert unstructured text into graphs, Graph-based approach provides the ability that takes advantage of the benefits of graph algorithms and use them in natural language processing algorithms. In this paper, we discussed and analyzed the results of the generalized method to detect plagiarism on the inner levels (Text Alignment). By achieving the Benchmark of Plagdet 87% without using linguistic corpora and grammatical rules, it is expected that more works in graph based approaches achieve better results in plagiarism detection. Furthermore, being independent from rules and corpora enables this method to detect plagiarism in other languages. As a future work we want to increase the accuracy of the algorithm to detect semantic plagiarism using FarsNet lexical network. Since it is possible to add word's synonyms to the corresponding graph of the document, by adding synonym words, the accuracy of detecting semantic plagiarism is increased. Another important category of modern plagiarism, is plagiarism on summary of a text. Due to the flexibility of graph approach in detecting plagiarism, graph approach is also efficient in detecting plagiarism on summary of a text.

5. REFERENCES

[1] 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.

[2] Barrón-Cedeño, A., Potthast, M., Rosso, P., Stein, B., and Eiselt, A. 2010. Corpus and Evaluation Measures for Automatic Plagiarism Detection, In *LREC*.

[3] Flowerdew, J and Li, Y. 2007. Plagiarism and second language writing in an electronic age, *Annual Review of Applied Linguistics*, vol. 27, pp. 161--183.

[4] 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.

[5] ICT Research Institute, ACECR ,Iran, 2016. [Online]. Available: http://ictrc.ac.ir/plagdet/.

[6] Khoshnavataher, K., Zarrabi, V., Mohtaj, S., and Asghari, H. 2015. Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation. *Notebook for PAN at CLEF2015*, 8-11 September, Toulouse, France.

[7] Kumar, N. 2014.A Graph Based Automatic Plagiarism Detection Technique to Handle Artificial Word Reordering and Paraphrasing, *Computational Linguistics and Intelligent Text Processing*, pp. 481--494.

[8] Leacock, C., Miller, G.A. and Chodorow, M., 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1), pp.147-165.

[9] Leilei, K., Haoliang, Q., Shuai, W., Cuixia, D., Suhong, W., and Yong, H. 2012. Approaches for candidate document retrieval and detailed comparison of plagiarism detection, in *Notebook for PAN at CLEF 2012*.

[10] [Online]. Available: hazm package, http://www.sobhe.ir/hazm. [Accessed 2015-04-30].

[11] [Online]. Available: igraph package, http://igraph.org/python/. [Accessed 2015-03-15].

[12] Potthast, M and Hagen, M ., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., and Stein, B. 2013. Overview of the 5th international competition on plagiarism detection, in *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, CELCT.

[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] Shamsfard, M., 2008. Developing FarsNet: A lexical ontology for Persian. In *4th Global WordNet Conference, Szeged*, Hungary.

[16] Sonawane, S.S. and Kulkarni, P.A., 2014. Graph based Representation and Analysis of Text Document: A Survey of Techniques. International Journal of Computer Applications, 96(19).

[17] Torres, S. and Gelbukh, A., 2009. Comparing similarity measures for original WSD lesk algorithm. Research In *Computing Science*, 43, pp.155-166.

[18] Zager, L., 2005. Graph similarity and matching (Doctoral dissertation, Massachusetts Institute of Technology).

[19] Zini ,M. 2006 Plagiarism Detection through Multilevel Text Comparison, In *Automated Production of Cross Media Content for Multi-Channel Distribution,Second International Conference*.

[20] Zini, M., Fabbri, M., Moneglia, M., and Panunzi. 2006. Alessandro, Plagiarism detection through multilevel text comparison, In *Second International Conference, IEEE*.