A Persian Fuzzy Plagiarism Detection Approach

Shima Rakian*
Department of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran
shima.rakian@yahoo.com
Faramarz Safi Esfahani
Department of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran
fsafi@iaun.ac.ir
Hamid Rastegari
Department of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran
rastegari@iaun.ac.ir

Received: 05/Dec/2014 Revised: 06/Feb/2015 Accepted: 18/Feb/2015

Abstract

Plagiarism is one of the common problems that is present in all organizations that deal with electronic content. At present, plagiarism detection tools, only detect word by word or exact copy phrases and paraphrasing is often mixed. One of the successful and applicable methods in paraphrasing detection is fuzzy method. In this study, a new fuzzy approach has been proposed to detect external plagiarism in Persian texts which is called Persian Fuzzy Plagiarism Detection (PFPD). The proposed approach compares paraphrased texts with the aim to recognize text similarities. External plagiarism detection, evaluates through a comparison between query document and a document collection. To avoid un-necessary comparisons this tool employs intelligent technology for comparing, suspicious documents, in different levels hierarchically. This method intends to conformed Fuzzy model to Persian language and improves previous methods to evaluate similarity degree between two sentences. Experiments on three corpora TMC, Irandoc and extracted corpus from prozhe.com, are performed to get confidence on proposed method performance. The obtained results showed that using proposed method in candidate documents retrieval, and in evaluating text similarity, increases the precision, recall and F measurement in comparing with one of the best previous fuzzy methods, respectively 22.41, 17.61, and 18.54 percent on the average.

Keywords: Text Retrieval; Plagiarism Detection; External Plagiarism Detection; Text Similarity; Fuzzy Similarity Detection.

1. Introduction

In this article, different types of plagiarism and their detection methods are studied. Also, a method based on fuzzy information retrieval is proposed to detect plagiarism.

Precision and recall are significant performance factors in plagiarism detection system. In this paper, we present an approach to external plagiarism detection in Persian texts, PFPD (Persian Fuzzy Plagiarism Detection). The aim of this tool is to make compatible fuzzy method in Persian language, thus the procedures put forward by [1], would be improved and the precision and recall increased.

The fuzzy statement here, is extracted from previous researches done in this field, and denotes the calculation of similarity between zero to one range.

The precision and recall in the tools for plagiarism detection is very important. The problem with previous systems in intelligent plagiarism detection is the embedding of plagiarized parts in varied sentence structures and synonym replacement [2].

The Language-independent tools may be inefficient for particular languages such as Arabic and Farsi [4].

The main problem is to present an approach to verify plagiarism through an efficient algorithm in order to recognize similarities and improve the precision and recall in obtained results in a timely manner. The other problem is prevention or minimizing the unnecessary repetitious operations [5]. The existing solutions are time-consuming [6].

In the phase of candidate document selections and plagiarism analysis, presented methods do not encompass adequate precision and recall to detect paraphrasing [2].

Therefore, the problem of this research is:

Increasing precision and recall in candidate documents retrieval in Persian language, through hierarchical methods, and in measuring the similarity of a Persian text by using fuzzy-methods.

We believe that the use of semantic methods, based on fuzzy IR for paraphrasing detection establishes more precision and recall, in comparison with other methods [2]. Using the retrieved candidate's documents of hierarchical methods at any phase, the program checks the documents with an increased possibility of plagiarism detection and prevents checking of unimportant documents [4].

For the design of this system, Apache nutch and Apache solr are applied. Nutch used for crawling while solr is for indexing and searching data of candidate documents retrieval. Experiments on three corpora TMC, Irandoc and extracted corpus from prozhe. com, are performed to get confidence on proposed method performance. The experiments denote that using proposed method in candidate documents retrieval, and in evaluating text similarity, increases the precision, recall and F measurement in comparing with one of the best previous fuzzy methods, respectively 22.41, 17.61, and 18.54 percent on the average.

* Corresponding Author

The organization of the paper is as follows. Section 2 describes related works to the paper and their limitations. The approach proposed here is described in Section 3. Section 4 states a description of test collection which is used for evaluation and presents the results of this evaluation. Finally, Conclusions are presented in Section 5.

2. Related Work

Depending on the language of compared texts, plagiarism detection, is classified into two groups as; mono-lingual or cross-lingual (Figure 1).

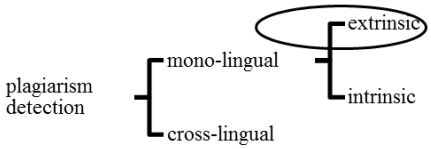

Fig. 1. Plagiarism Detection Techniques

For external plagiarism detection, a suspicious document is compared with one or more other documents. The operational platform for that is briefly outlined below [3]: (Figure 2)

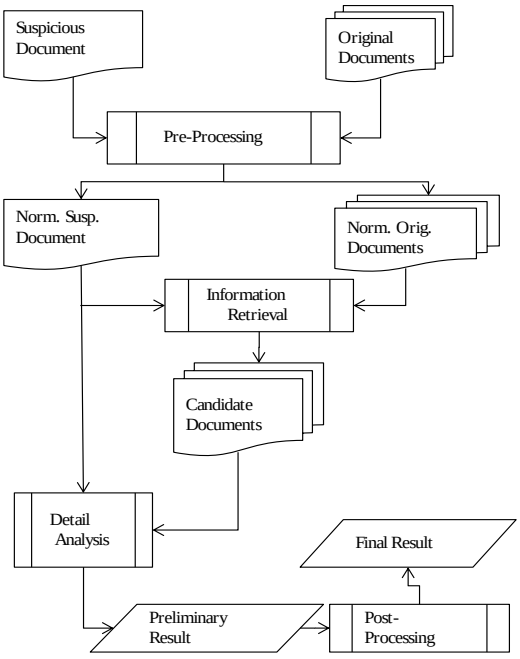

Fig. 2. Methodology of extrinsic plagiarism detection [2,3]

1. The suspicious document and original documents; including the sources that may contain plagiarism which are regarded as input.
2. Three main steps are required:
   a. The retrieval of a list of candidate documents, using the models of information retrieval.
   b. Comprehensive analysis to compare suspicious document with candidate documents.
   c. Post-processing to mix small detected units to be presented to the viewer. At this stage, it should be decided whether plagiarism exists or not.

Meyer zu Eissen, S et. al. [7] discuss the different types of plagiarism. This represented classification in [2], is a more comprehensive, where plagiarism is divided into two groups: i) Literal ii) Intelligent. Fig. 3 illustrates the different types of plagiarism. Hence, the focus of this paper is detecting paraphrased sentences.

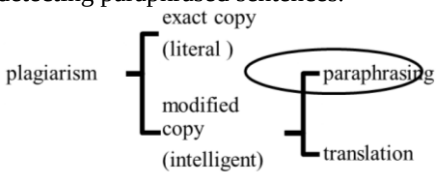

Fig. 3. Types of plagiarism

The two main stages in plagiarism detection are candidate document retrieval and detail analysis, to compare suspicious document with candidate documents. In candidate documents retrieval there is global similarity, but in detail analysis step, local similarity is considered. In local similarity, two documents are checked from a semantic point of view which forms the basis for plagiarism detection [8].

In this section only the most related works are mentioned. In research [9], to obtain semantic similarity, the depth and length of the shortest path to the word, in sequence of Synset in Wordnet, is applied. The structure of Wordnet changes this into a useful tool for a natural language processing.

Table 1. Most Related Work.

	Date	Specification	Advantage	Disadvantage
[5]	2005	Using the term-term correlation matrix in fuzzy information retrieval approach	Able to distinguish synonym replacement and structure change	High expenses to make and keep matrix
[6]	2006	Using the depth and length of the shortest path to the word, in sequence of Synset in Wordnet	Able to distinguish synonym replacement and structure change	Unable to be used in Persian language, because Farsnet is unable to calculate depth and length of the path to the word
[1]	2010	Fuzzy analysis for plagiarism detection	Because of its nature, Fuzzy analysis for paraphrasing detection is more effective than other approaches	Using shingling and Jaccard coefficient in the candidate selection step reduces the precision. In cases where sentences are long and complex, this procedure is of low precision and may not detect plagiarism.
[4]	2012	In candidate documents, documents	Performs the analysis on several levels,	Low precision, due to synonym replacement.

	Date	Specification	Advantage	Disadvantage
		selection, research in sequences of the documents, paragraphs and sentences through intersection rates, was carried out.	to avoid un-necessary comparisons.	To obtain intersection rates in three levels is costly. Using fingerprint in paraphrasing detection, is not suitable
Proposed Method (PFPD)	2014-2015	Fuzzy analysis for plagiarism detection and in candidate documents retrieval	Increasing precision and recall in candidate documents retrieval and in measuring the similarity. Avoid un-necessary comparisons	

In [1], proposed Web based plagiarism detection using fuzzy information retrieval, a mixture of fuzzy similarity model [10] and semantic similarity, obtained from lexical database [9] were employed. Similar to model [9], instead of using term-term correlation matrix, in model [10], the extracted Synset from Wordnet was employed. Thus Semantic observation of the text in which synonym of the word; was extracted using Wordnet database [11] was strongly increased. After that the similarity degree of two sentences was calculated. This system was proposed to detect external plagiarism, capable of paraphrasing detection in English language. Pre-processing of this system was done using different procedures including: tokenization, stemming, Stopword removal and candidate selection with Shingling and Jaccard coefficient. In the analysis phase of plagiarism detection, makes use of fuzzy plagiarism analysis. This system was tested using datasets PAN'09, PAN'10. The advantage of this system is that fuzzy analysis for paraphrasing detection is more effective than other approaches such as Shingling. The disadvantage of this system may be caused by imprecision of shingling and Jaccard coefficient in candidate selection. The goal in this fuzzy analysis for plagiarism detection was to detect paraphrasing, and in the case of long sentences, the already proposed fuzzy analysis proved to be imprecise due to addition of words and other sentences. The proposed tool aims to improve the plagiarism detection method using fuzzy IR.

In [4], APLAG was proposed which was able to detect paraphrasing in the Arabic language. In this system, during the pre-processing phase, tokenization, Stopword removal, stemming and synonym replacement were used and in candidate documents selection, research in hierarchy of the documents, paragraphs and sentences through intersection rates, was done. In the analysis phase of plagiarism detection, this system made use of fingerprint techniques which was tested using three datasets provided by the writer. The advantage of this system is performing the analysis on several levels, to avoid un-necessary comparisons. The disadvantage of this system is its low precision, due to synonym replacement. To obtain intersection rates in three levels is costly. Additionally, the use of fingerprint in paraphrasing detection is not suitable. The idea of intersection blocks in this paper is the result of the aforementioned hierarchal procedure. In Table these approaches and their advantages and disadvantages are highlighted.

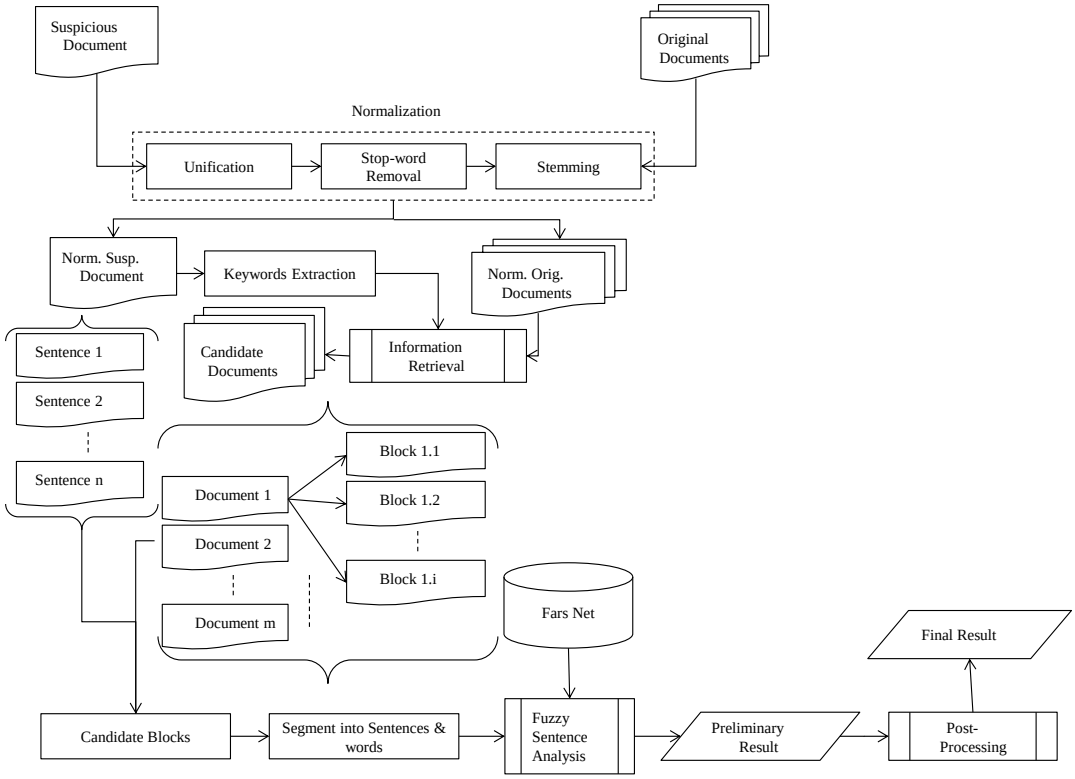

Fig. 4. PFPD Plagiarism Detection Steps

3. PFPD-Persian Fuzzy Plagiarism Detection

In this paper, a tool is proposed to detect external plagiarism in Persian texts, namely Persian Fuzzy Plagiarism Detection (PFPD). This tool is to compare Persian documents, using a fuzzy approach, to recognize similarities.

Compared with other methods, this tool makes use of intelligent technology for comparison of suspicious documents in different levels sequentially, in order to avoid un-necessary comparisons.

In the article, sentences-level representation is used at the analysis phase of plagiarism. Our aim is to adopt fuzzy model in Persian language and improve the methods offered in [1], to calculate the degree of similarity between two sentences. In this method, the degree of similarities between two sentences is calculated within 0-1. If this figure is larger than threshold, the two sentences are considered similar. The threshold is regarded 0.65, the same as [1].

To investigate the effectiveness of this method, three datasets were used. The performance of the system is measured with precision, recall, and F-measure metric.

3.1 PFPD Approach

We use an algorithm for mono-lingual external-paraphrasing detection. This method is particularly useful for recognition of different paraphrasing levels, by using semantic similarity based on fuzzy IR.

Fig. 4 illustrates the operational platform for this method. Input encompasses the suspicious document and source collection; including sources which may have plagiarism. In order to avoid the unnecessary steps in recognition of the plagiarized sentences, keywords from the suspicious texts are used as verifiers of the original documents. Afterwards, certain blocks are chosen in the original document based on the similarity of the sentences of the suspicious document and the blocks in original text.

After pre-processing operations, a list of candidate documents, related to suspicious documents are retrieved and the sentences are checked for plagiarism in details. The method of semantic similarity based on fuzzy IR was used. The fuzzy IR method investigates the suspicious blocks in details. To detect exact copy, 100% similarity should be reached. However, since the system is intended to detect paraphrasing, if the compatibility is more than a certain level, we regard the texts as similar. The detail of the illustrated steps are described below.

3.2 Phase I: Pre-processing

The process begins with eliminating excess data in the original and suspicious documents. The text should be pre-processed, in advance. Pre-processing includes: text unification, Stopword removal, and stemming. Pre-processing is a key stage in obtaining satisfactory results when facing the difficulties in natural language processing. Stopword removal leads to non-sense words. Stemming algorithm is also used to eliminate prefix and postfix to establish word roots. To do that, we produced Aria Package[1], which performs the required operations for Persian language processing.

3.3 Phase II: Candidate Document Retrieval

In this phase keywords in query document and their Synsets are extracted and among original documents, those which include these words, are retrieved. The retrieved and query document are inspected and for each sentence of query document, related blocks with higher intersection are retrieved.

Operations are done before sending a paragraph to the plagiarism detection system to retrieve candidate documents.

Note that using this method, the number of candidate for each suspicious document is dynamic and may be small which could reduce calculation time, while having a fixed number of candidate.

3.4 Phase III: Plagiarism Analysis

After above stage is complete, main operations of analysis are started. In this stage, Candidate blocks with suspicious document are analyzed, sentence by sentence to have a precise study.

At first, blocks are broken by punctuation. At this stage, the values below, for each sentence of the suspicious document and sentence of the blocks of original documents are calculated. The operation unit at this stage is "word". The reason not to use word n-gram, is that in the text paraphrasing, there is the possibility of word order changes. At first similarity of the words in the two sentences (δ) is obtained (1). Synonym words are obtained using Farsnet [12], that is the first Persian Wordnet.

$$\delta = \#\text{Synonym words} * 0.5 + \#\text{exact words} * 1 \quad (1)$$

The intersection ratio of two sentences to the first sentence (α) and the intersection ratio of two sentences to the second sentence are calculated with (2) and (3).

$$\alpha = \frac{\delta}{|S_1|} \quad (2)$$

$$\beta = \frac{\delta}{|S_2|} \quad (3)$$

Intersection to union of two sentence ratio (γ) is calculated with (4).

$$\gamma = \frac{\delta}{(|S_1| + |S_2|) - \delta} \quad (4)$$

Afterwards, the similarity ratio (sim(s1,s2)) is calculated with (5).

$$sim(S_1, S_2) = \frac{A * \min(\alpha, \beta) + B * \max(\alpha, \beta) + C\gamma}{A + B + C} \quad (5)$$

In this similarity metrics order of the words does not affect the degree of final similarity between them. The

[1] www.aria-ware.com

values of coefficients A, B, and C are considered 20, 8, and 3 according to tests.

Finally if $sim(S_1,S_2)$ is greater than the threshold value (T), the sentence is marked as plagiarism, otherwise it is considered to be not plagiarized (6). Following tests, appropriate value for T is obtained to be 0.65, similar to [1].

$$\mathrm{EQ}(S_q, S_x) = \begin{cases} 1 & \text{If } \mathrm{sim}(S_1, S_2) \geq T \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

3.5 Phase IV: Post-processing

Finally, the output of this algorithm is a list of sentences, indicated as similar/plagiarized. Since the sentence is a comparison unit, they are combined in paragraphs.

4. Experimental Evaluation

In text paraphrasing, two sentences may have the same meaning, but different structures; e.g. replacing synonym and adding or shortening the sentences.

For the design of this system, Apache nutch and Apache solr are applied. Nutch used for crawling while solr is for indexing and searching data.

Considering that this method is to improve the method proposed in [1], a comparison between the two was made.

4.1 Test Collections

During the evaluation of the method, we used three different corpora:

1) The Tehran Monolingual Corpus (TMC)[1]: TMC, established by Tehran University, was employed, and after pre-processing, the proposed method was performed. The dataset included the news that was extracted from Hamshahri Corpus and ISNA (Iran students' broadcasting) news agency website. TMC is a large-scale Persian monolingual corpus. From TMC, 1000 documents were achieved as source documents, from which 400 suspicious documents were produced.
2) The IRANDOC Test Collection: Iranian Research Institute for Information Science and Technology (IRANDOC) [2] previously known as Iranian Research Institute for Scientific Information and Documentation is an Iranian research center with a national mission to meet the country's needs in the field of information science and technology. IRANDOC provided 230 documents from which 220 suspicious documents were produced.
3) The prozhe.com Test Collection[3]: The prozhe.com is a website presents student research documents and articles, from which 440 documents are extracted as source documents, and 160 suspicious documents were produced from these.

[1] Available from http://ece.ut.ac.ir/nlp/resources.htm
[2] http://www.irandoc.ac.ir
[3] http://www.prozhe.com

4.2 Query Collections

Two query collections for each corpus, is established. Each collection includes four types of query that is produced artificially.

The four types include the cases below:

1) Synonym replacement of 50% of the words in each sentence.
2) Sentence structure changes with an increase in 45% in the number of words in each sentence.
3) A combination of points 1 and 2, in a sentence.
4) A combination of points 1 and 2, in different sentences of document.

In order to have a query, 50% of the sentences, depending on the type of query (which maybe of the four cases above) are replaced. The difference between first and second series is in the replacement of the second 50% of the document.

In the queries of the first set, 25% of the sentences with exact copy, and 25% as non-copied sentences are replaced (the rest set with ratio of 1 to 1).

In the queries of the second set, 12.5% of the sentences with exact copy, and 37.5% as non-copied sentences are replaced (the rest set with ratio of 1 to 3).

For IRANDOC corpus, one new query was added, which includes 100 paraphrased sentences that are created manually.

The reason for establishing two query collection is to highlight the faults in [1], and ratification of these problems with the present method. Two main fault in [1] are:

1) Lack of precision in the recognition of plagiarism and inability to distinguish paraphrased cases, in which the length of the sentences are increased by adding the words. This phenomena is due to poor of fuzzy formula in [1].
2) Inability to detect plagiarism, due to the approach of weak candidate retrieval. In [1], the Jaccard coefficient is used which is operation heavy. In addition, when the plagiarized text of less volume in comparison with original documents, the Jaccard coefficient is less than 0.1, that document is not selected as candidate to follow other stages. For example, assume of a thesis with 5 chapters, just one is copied, considering the above mentioned problem, research [1] is not able to detect that copy.

The first query collection, was used to highlight the fault of the fuzzy method applied in [1] and a solution was in proposed method. The second query collection was used to represent the fault of second case.

The description of applied corpora and their test results are presented in the following section.

4.3 Evaluation

Each sentence of the suspicious document is compared with original documents. Assuming S_1 to have a bigger sentence length and S_2 as the second sentence length, then $1 \leq S_2 \leq S_1$. In Fig. 5 and Table similarity values are presented for $9 \leq S_1 \leq 12$.

Obtained results indicated that, when lengths S_1 and S2 are close to each other, results of PFPD and method [1] are closely in compatibility. But when S_1 and S_2 are far from each other and small sentence is extracted from the bigger sentence (i.e. more than 85% of bigger sentence), PFPD is able to detect plagiarism, while the method outlined in [1] is not able to detect these cases.

As you see in Table , there is only one row in $9 \leq S_1 \leq 12$, which is detected by the method [1], but the present method is not able to detect that. In which $\gamma \leq 0.5$, indicating ratio of intersection on union is less than 0.5, and the method [1] detected that wrongly.

Fig. 6 illustrates the percent of precision, recall and overlap calculated correctly, For $1 \leq S_1 \leq 30$, in PFPD, and the method [1].

Studying the results indicated that the average amount of precision and recall in PFPD respectively are 100%, 99.97%, but in [1] are 99.10%, 85.05% respectively. Additionally, the PFPD is able to detect 99.97% of the correct cases detected in [1] on average.

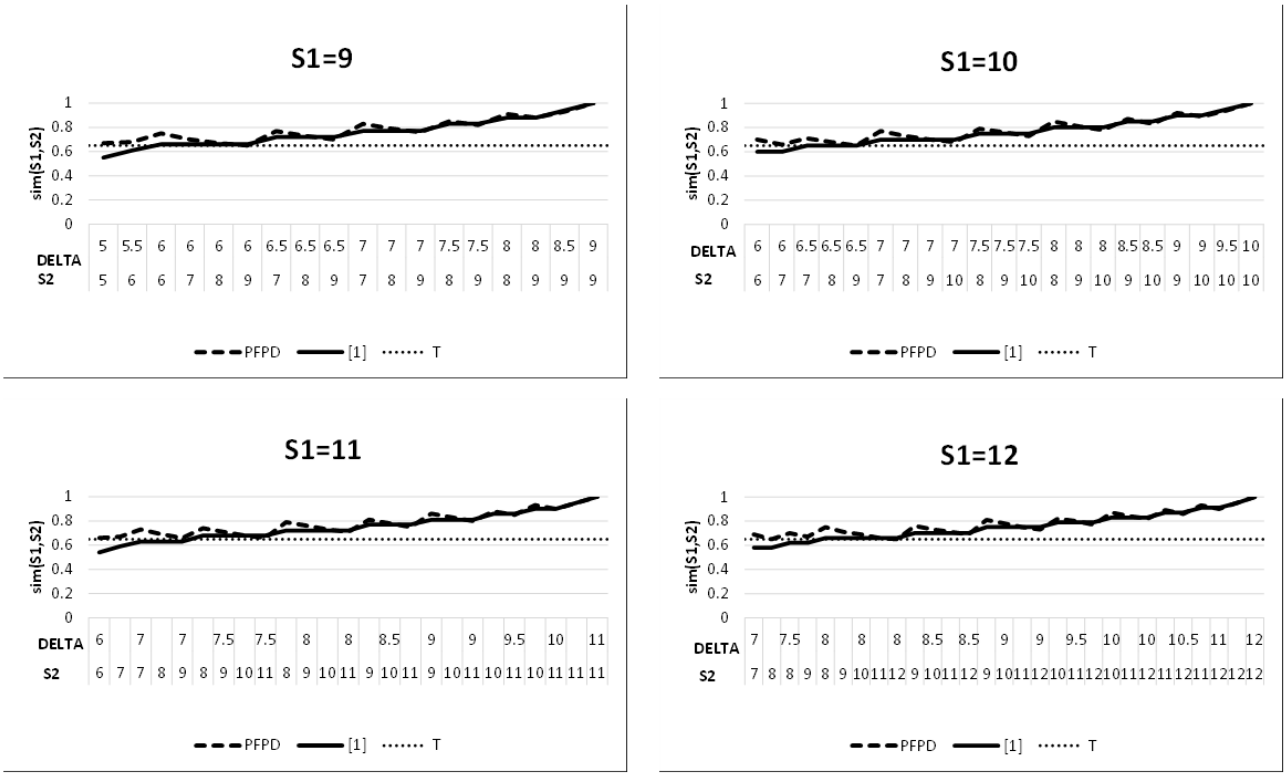

Fig. 5. Similarity ratio calculated for $9 \leq S1 \leq 12$ and $PFPDSim(S1,S2) \geq T$ when T=0.65

Table 2. similarity ratio calculated in $9 \leq S1 \leq 12$ for the cases where ($PFPDSim(S1,S2) \geq T$ or $[1]Sim(S1,S2) < T$) and ($PFPDSim(S1,S2) < T$ or $[1]Sim(S1,S2) \geq T$)

S_1	S_2	□	□	□	□	PFPD	[1]
9	5	5	0.55	1	0.55	0.67	0.55
	6	5.5	0.61	0.91	0.57	0.68	0.61
10	6	6	0.6	1	0.6	0.7	0.6
	7	6	0.6	0.85	0.54	0.66	0.6
	10	6.5	0.65	0.65	0.48	0.63	0.65
11	6	6	0.54	1	0.54	0.66	0.54
	7	6.5	0.59	0.92	0.56	0.67	0.59
	7	7	0.63	1	0.63	0.73	0.63
	8	7	0.63	0.87	0.58	0.69	0.63
	9	7	0.63	0.77	0.53	0.66	0.63
12	7	7	0.58	1	0.58	0.69	0.58
	8	7	0.58	0.87	0.53	0.65	0.58
	8	7.5	0.62	0.93	0.6	0.7	0.62
	9	7.5	0.62	0.83	0.55	0.67	0.62

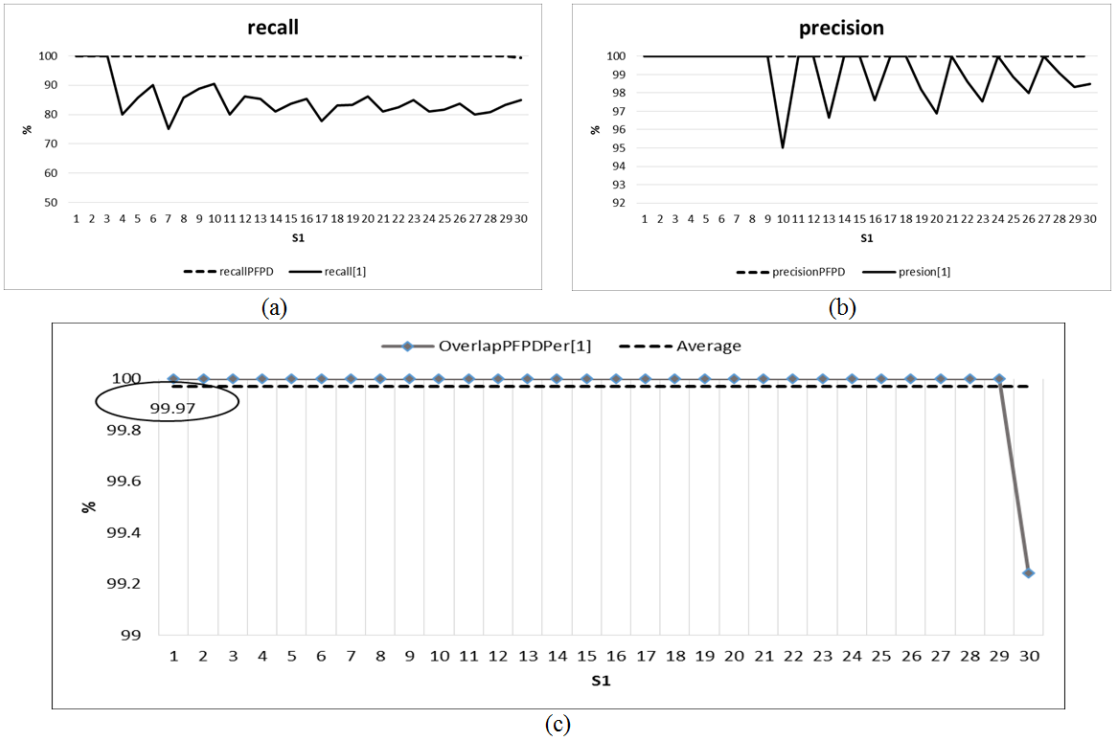

Fig. 6. (a) Recall (b) Precision (c) overlap PFPD per [1] for 1≤S1≤30

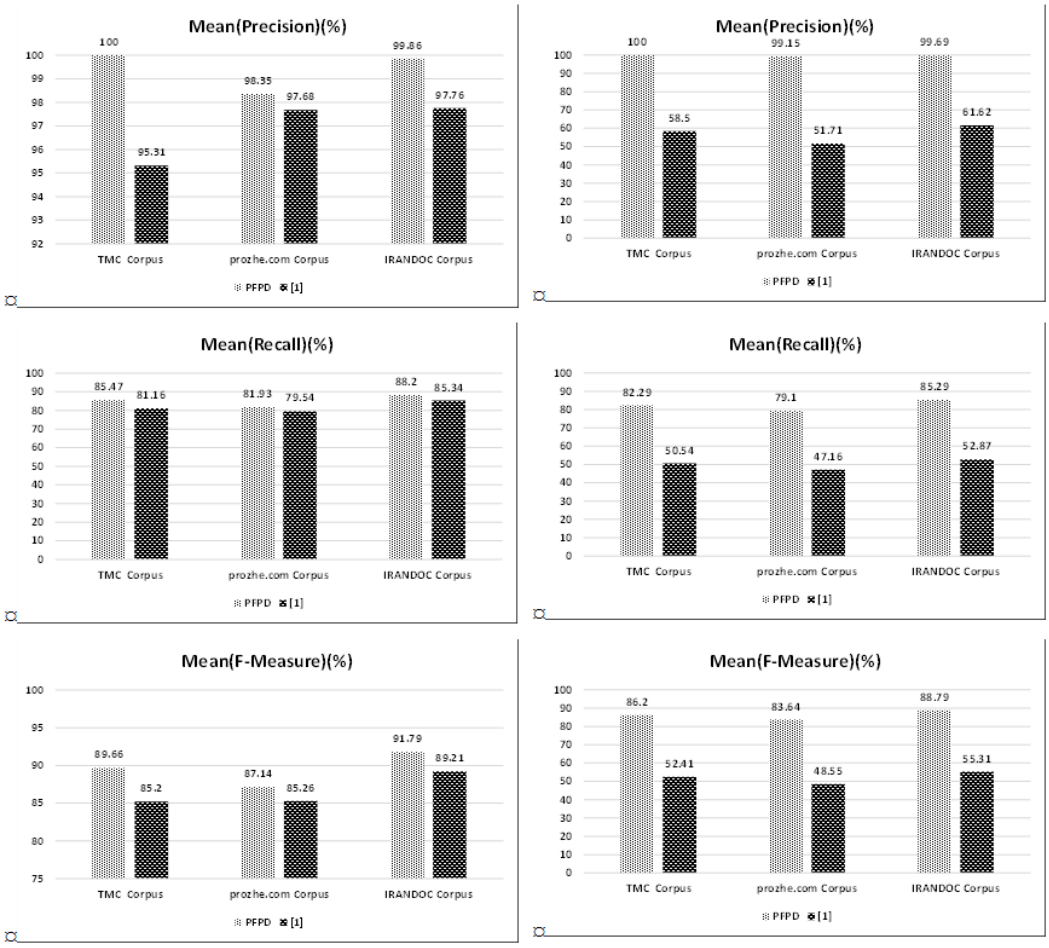

Fig. 7. Precision, Recall and F-Measurement (a) first experiment series (b) second experiment series.

The three datasets described in Section 4.1, and their queries were input in to the system. The results are presented in Fig. 7.

Based on these results, all cases picked up by method [1], could be also picked up by PFPD.

In the first experiment series used for test fuzzy formula, in four types of query increase of precision, recall and F-measurement are obtained respectively 2.49, 3.19 and 2.97. In the second experiment series which were used for test candidate document retrieval stage, it was found that precision and recall in method [1] is strongly dependent on volume of the copied document. If it is of low value in the candidate retrieval stage, this document would not be selected, and this method would not progress to the next stage and the plagiarized text would not be detected. PFPD method could obtained a high degree of precision and recall by improvement to the candidate retrieval stage. Increase of precision, recall and F-measurement are obtained respectively 42.34, 32.04 and 34.12.

Regarding the first and second experiments, increase of the precision, recall and F measurement are obtained respectively 22.41, 17.61, and 18.54 percent on the average.

4.4 Time Complexity

According to the investigations done, the number of source documents have no effect on precision and recall. Because according to the suggested algorithm, all the source documents including one of the keywords in suspected document, are selected and checked as candidate for next stage. Therefore the number of these documents is of no effect on the precision and recall. But influencing on algorithm speed which are examined in the following.

The experiments were done on a HP-Pavilion dv4 laptop. In these experiments, the volume of the input document was 3 KB. At first the source documents were 50, and in each experiment, 50 documents were added. The time of each experiment was measured. Fig. 8 demonstrates the comparison between PFPD and [1] in terms of time required for Plagiarism Detection. The results from the comparison show that the proposed method achieved better results in terms of time required for Detection.

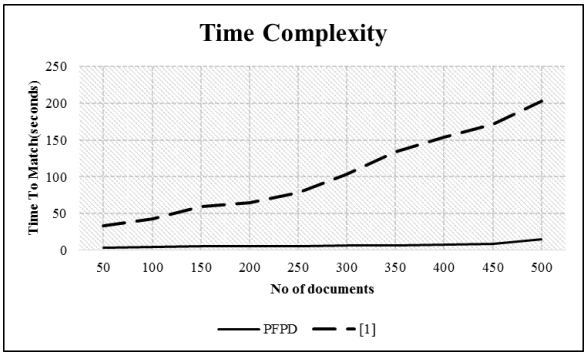

Fig. 8. Time Complexity

The method [1] investigates the whole content of the candidate documents, to obtain Jaccard coefficient. While the suggested method compares suspicious documents, in different levels hierarchically. This leads to reduction of the operations and increase of speed.

5. Conclusions

To identify paraphrasing based on sentences, fuzzy method is effective, as it has the capacity to distinguish similar sentences, based on the similarity among a set of synonym words. In this article, PFPD to detect external mono-lingual plagiarism was performed. This semantic fuzzy method is designed to detect different degrees of paraphrasing.

The obtained results showed that using proposed method in candidate documents retrieval, and in evaluating text similarity, increases the precision, recall and F measurement in comparing with one of the best previous fuzzy methods, respectively 22.41, 17.61, and 18.54 percent on the average. Also the results from the comparison show that the proposed method achieved better results in terms of time required for Detection.

Previous presented fuzzy methods are unable to distinguish paraphrased cases, in which the length of the sentences are increased by adding the words. While PFPD is able to detect such cases, it is also capable of picking up cases where the length of the plagiarized sentences are close to each other.

References

[1] S. M. Alzahrani and N. Salim, "Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection," Braschler and Harman, 2010.

[2] S. M. Alzahrani, et al., "Understanding Plagiarism linguistic patterns, textual features, and detection methods," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 42, pp. 133-149, 2012.

[3] B. Stein, et al., "Strategies for retrieving plagiarized documents," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 825-826.

[4] M. E. B. Menai, "Detection of Plagiarism in Arabic Documents," International Journal of Information Technology, vol. 4, 2012.

[5] A. H. Osman, et al., "CONCEPTUAL SIMILARITY AND GRAPH-BASED METHOD FOR PLAGIARISM

DETECTION," Journal of Theoretical and Applied Information Technology, vol. 32, pp. 135-145, 2011.
[6] D. Ceglarek and K. Haniewicz, "Fast Plagiarism Detection by Sentence Hashing," in Artificial Intelligence and Soft Computing, 2012, pp. 30-37.
[7] S. Meyer zu Eissen, et al., "Plagiarism detection without reference collections," Advances in data analysis, pp. 359-366, 2007.
[8] S. M. Alzahrani and N. Salim, "Plagiarism detection techniques," 2008.
[9] Y. Li, et al., "Sentence similarity based on semantic nets and corpus statistics," Knowledge and Data Engineering, IEEE Transactions on, vol. 18, pp. 1138-1150, 2006.
[10] R. Yerra and Y. K. Ng, "A sentence-based copy detection approach for web documents," Fuzzy systems and knowledge discovery, pp. 481-482, 2005.
[11] G. A. Miller, "WordNet: a lexical database for English," Communications of the ACM, vol. 38, pp. 39-41, 1995.
[12] M. Shamsfard, et al., "Semi automatic development of farsnet; the persian wordnet," in Proceedings of 5th Global WordNet Conference, 2010.

Shima Rakian was born in Iran. She has recently received her master's degree in software engineering from Islamic Azad University, Najafabad Branch, Iran. Her current research interests include Data Mining and Text Retrieval.

Faramarz Safi Esfahani was born in Iran. He got his Ph.D. in Intelligent Computing from University of Putra Malaysia in 2011. He is currently on faculty at Department of Computer Engineering, Islamic Azad University, Najafabad branch, Iran, since 2002. His research interests include intelligent computing, Cloud Computing, Autonomic Computing, and Bio-inspired Computing.

Hamid Rastegari was born in Iran. He got his Ph.D. in Computer Science – Soft Computing from University of UTM Malaysia in 2011. He is currently Assistant Professor on faculty at Department of Computer Engineering, Islamic Azad University, Najafabad branch, Iran. Experiences 1-Title: Head of Computer Department- Organization: University of Applied Science and Technology from 2002 to 2007 2-Title: Head of Postgraduate Department- Organization: IAUN from 2013 to Pres. 3-Title: Coordinator of 1st National Conference on Computer Science ament- Organization: IAUN from 2013 to 2013.