IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 12, Issue 2, Ver. III (Mar.-Apr. 2017), PP 60-67
www.iosrjournals.org

Plagiarism Detection Between Theory And Practical Calculations

Intisar H. Albakush[1]
[1](*Electrical Engineering and computer science, Information Technology/ Singidunum Univeristy, Serbia)*

Abstract: *Plagiarism has been the problem of era in different acknowledge fields, particularly in the academic community, theses a battle between the plagiarism epidemic and detection create a rivalry between the machines and humans on both sides negative and positive, i.e.In plagiarisms cases and protecting and detecting the plagiarism. In this paper the work was on some human calculations for detecting the plagiarism and similarity in text documents and their relationship for automatically detecting the plagiarism mentioned tools and results of plagiarism detection tools for Arabic and English speeches, with calculate the precision and recall and F-meter for the mentioned software.*

Keywords: *Automatically plagiarism detection, calculate the similarity, plagiarism detector tool, Precision and recall values, tf-IDF.*

I. IntroductionAnd Related Work

The plagiarism, which defined as the routine of representing the creativity of someone else's thoughts, resolutions, or words without recognizing the original source has been dubbed as illegal quotation, theft, cheating, plagiarism, and likewise[1].

Plagiarism types had classified into main three types, extrinsic plagiarism, where the plagiarizer copy and paste the original information, in its own work, or changing some words or sentences with their synonyms or antonyms, this subtype called literal plagiarism and its easier to detect than other subtype, which called intelligent plagiarism, detecting the plagiarism in this type is really challenge, where plagiarizer paraphrasing or summarizing the whole information and the main idea of original works to pretend it as its own work [2].There were a lot of methods and calculations had found for detecting and limiting such kind of plagiarism, for example; Fingerprint-based method, where the query and suspicious documents marked with some fingers, or numbers, then comparing with each other for detecting the similarity and plagiarism. N-gram-based method; n could be character or word for detecting the exact string matching, Longest Common Sequence (LCS) combined with part-of-speech (POS) technique for detecting plagiarism [3] [4] [5] .Other methods had done by tokenizing the text, constructing word n-gram; and utilizing vector similarity asin [4][5][6] [7]. These methods used for detecting the plagiarism had done by rebuilding the sentences or phrase.For the same type of plagiarism and the next subtype aka intelligent plagiarism, where the plagiarizer paraphrases the text without citation, or summarizing the main idea of the text and rewrite it with their own words, in this case, fuzzy-based method and semantic-based method, are successful methods for detecting the similarity in query document[3].
In these methods the document reads segments and parses to know the main contents, then retrieved and ranked the resources of plagiarism, to come out with the percentage effect of piracy.These methods use, which known term frequency and inverse document frequency (tf-IDF), as we will explicate in the methods section.

In some other case of plagiarism called intrinsic plagiarism where detecting the plagiarism does not depend on outside sources for matching the words or sentences, or even ideas exactly, but the method here for detecting the plagiarism depending on the suspicious document by detecting the changing in style writing of the same source.The writing style can be analysed within the document and examination for incongruities can be performed, the complexity of styling analysis can be parsed according to some parameters as part of speech, syntactic feature, statistics text features.The primary method is detecting the changing in writing mode, there are some methods used character n-gram for distinguishing the main stylized writing of an author, then compare that with the document had created by that generator [6] it goes further in [7] [8].Intrinsic plagiarism detection, is still has many challenges that because detecting the plagiarism and similarity are done for literal writing than science writing, also it's more complicated in Arabic language where the text sometimes has some Quraan quotations, and the text somewhere needs to put diacritics to give precise meaning [9].

In cross-lingual plagiarism or translated language plagiarism. Detection methods had developed in recent few years, many researchers had adopted this method for detecting the plagiarism through the documents. All methods had put the translated machine or human translation as the first step after preprocessing stages.In some cross language detecting plagiarism cases, third language as a pivot between the primary pairs of languages, for example, Arabic language text has translated to Spanish language, the researcher had used the English words as a pivot in the transitive method, at that place are a set of researchers had adopted such method for detecting the plagiarism in text documents, as in[10], [11].Al-Johani et al [12] used the winnowing algorithm for detecting plagiarism across Arabian-English. Alzahrani et al [13], has made their experiments with short

phrases and sentences for detecting the plagiarism across Arabian-English using the semantic similarity methods.

The paper had arranged as: section 2 explained with examples the theory and mathematics calculation of similarity in text documents, uses Euclidean, Jaccard distance, cosine angle calculating and ti-idf technique.

Section 3: displayed some useful tools used for detecting the similarity and plagiarism,with experiments and results applied to plagiarism detector software. Section 4: discussion of previous experiments. Section 5: Conclusion and future work.

II. The Theory Calculation For Detecting The Similarity And Plagiarism

1- Jaccarda Calculations.
2- Vector space similarity calculations [Cosine similarity detection]
3- Tf-idf similarity calculations

The variance in methods for detecting the plagiarism and calculating a similarity in text documents is primarily due to the type of documents, as follows below.

The most popular method for evaluating the similarity between documents, is assessing the distance between terms in the documents, this technique also employed in computer programming for detecting the meanings of words, duplicated of documents, for instance in Google searching, and so on [14].Thereis a lot of such type of measurements, for instance Dice, Euclidean, Jaccard distance. Will review in glance the mathematical computation of Euclidean and Jaccard similarity distance.

2.1 Euclidean similarity distance

$$d_E(u,v) = \|u - v\| = \sqrt{\sum_{i=1}^{d}(v_i - u_i)} \quad ^2 \qquad (1)$$

Where u,v are documents had represented as vectors for calculating the distance between both of them, i to d represented the list of objectives for document u and v.

Immediately use the equation would not be useful or clear, but should modify and documents also ordered for such measurements.

2.2 Jaccard similarity distanceIt is defined by following equation

$$J = \frac{|d_1 \cap d_2|}{|d_1 \cup d_2|} \quad (2)$$

Where d_1 and d_2 are two documents, in example d_1 and d_2Will represented as numbers, for a short explanation.

Suppose $d_1 = \{0,1,8,6\}, d_2 = \{1,2,5,8\}$

According to previous equation

$$J = \frac{\{5,8\}}{\{0,1,2,5,6,7,8\}} = \frac{2}{7} = 0.286$$

If detecting plagiarism and similarity is done by the interrogation, a query is restricted by length and key words, therefore the suspicious documents retrieved from the resources, “almost web resources”, will be restricted by ranking and by the precise matching of lyric.

2.3 Cosine Angel for measuring the similarity in text documents

The best method to find the similarity between query and another query, or query and documents, is to represent the query and the matching document, as vectors. And then count on the cosine of the angle between that vectors, every act in following example:

Supposing that d1 and d2 are two documents presented as vectors with the product dots $(t_{11}, t_{12}, t_{13}, \dots\dots\dots t_{1n})$ and $(t_{21}, t_{22}, t_{23}, \dots\dots\dots t_{2n})$. Where (t) entails that the terms in the documents have coordinates, aka the product dots (11,12,21,22, etc...),Thatrepresents the number of document and number of terms in that document to be represented as vectors. And at that place is also Vq, which interpret the query vector, had mapped as $(v_{q1,} v_{q2,}, v_{q3,}, \dots\dots v_{qn})$. For assessing the similarity between both documents as vectors and their query, we count on the cosine of the angle between vectors by following equation:

t3

$$\cos\Theta = \frac{d1.d2}{|d1|.|d2|} \qquad (3)$$

$$|d1| = (t_{11} + t_{12} + t_{13} + \cdots + t_{1n})^{1/2}$$

$$|d2| = (t_{21} + t_{22} + t_{23} + \cdots + t_{2n})^{1/2}$$

$$d1.d2 = (t_{11}.t_{21} + t_{12}.t_{22} + t_{13}.t_{23+}..t_{1n}.t_{2n})$$

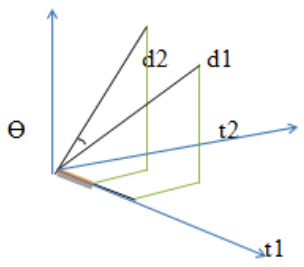


Fig.(1) Calculate cosine angle between two text files

As an instance, there are some facts for calculating the similarity in documents using such cosine angle measurements.
If the cosine value = 1, that means both documents are almost the same and the angle between them = 0º.
If the cosine value = 0, that means both documents are entirely different and the angle between them reaches 90º [15].
Example for computing the cosine similarity angle between the two documents and their interrogation.
Suppose that: d1 [We have a party Saturday night]
d2 [We did our party last Saturday at night]
And the Query
Q [Party on Saturday night]
d1 [we], [have],[a],[party], [Saturday], [night].
d2 [we], [did], [our], [party], [last], [Saturday], [at], [night].
Q [party], [on], [Saturday], [night]
Let see the relations between documents and query according to the following table

Table 1Summary of theory calculation uses cosine angle and tf-idf technique

	tf					
Term	Q	d1	d2	dfi	D/dfi	idf
a	0	1	0	1	3	0.4771
At	0	0	1	1	3	0.4771
did	0	0	1	1	3	0.4771
Have	0	1	0	1	3	0.4771
Last	0	0	1	1	3	0.4771
Night	1	1	1	3	1	0
On	1	0	0	1	3	0.4771
Our	0	0	1	1	3	0.4771
Party	1	1	1	3	1	0
Saturday	1	1	1	3	1	0
We	0	1	1	2	1.5	0.1761

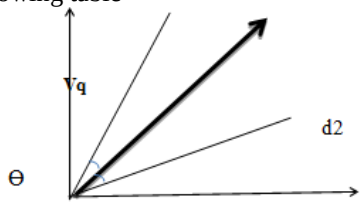


Fig.(2) Cosine angle similarity between query and documents

So for calculation let statistic the matching words between three vectors, for computing the cosine angle between d1 and vector query Vq, from equation (3)

$\cos\Theta=\frac{d1.d2}{|d1|.|d2|}$ (4)

$|d1|= (1+0+0+1+0+1+0+0+1+1+1)^{1/2}$=2.45

$|Vq|= (0+0+0+0+0+1+1+0+1+1+0)^{1/2}$=2

$d1.Vq$= (1.0+0.0+0.0+1.0+0.0+1.1+0.1+0.0+1.1+1.1+1.0)=3

Then CosinΘ=0. 6122 and Θ=52.25.

Same steps had done with Vq and d2 and the CosinΘ was 0.53 and Θ=10. 25.

III. TF-IDF Technique

The good way for theory calculation in the free text document is the one known as the tf - IDF, i.e. Term Frequency-Inverse Document Frequency. This calculation has a good opportunity to give an importance to each term in a document, as it is clear from the tf - IDF definition[16].Tf (term frequency) - is a number of repeating the term "word" in a document divided by the total number of words in that document.
If we have a document [d] with the total number of words [W_n] and frequent words appearing [i] times in that document, [w_i], And so:
Tf=w_i/W_n (5)
Tf = number of repeating the term "word" in a document / the total number of words in the document
IDF (inverse document frequency) - is a logarithm of the total number of text files in corpus divided by the number of documents having the condition
If we possess a corpus of (n) documents and just (i) documents have the term, then:
Idf= log(df/dfi) (6)
The grade or weight of tf-idf=tf. IDF (7)
The calculation example, using tf-idf technique is clear in table(1), previous section.

IV. Tools Used For Detecting The Plagiarism

There are tremendous numbers of software had created for detecting the plagiarism in almost all disciplines of knowledge we listed down some of them, then show some experiments had done with plagiarism detection software.

Software useful for detecting the plagiarism in natural language with almost more than 2languages, where most of the support more than 30 languagesand trusted from many universities and institutes around the world, all of them are not free for the whole applications. And have a good opportunity for checking anti-plagiarism, for instance database, internet checking, etc. as the following tools:

1. Turnitin/Turnitout
2. QARNET
3. iThenticaten
4. Plagiarism detector.

4.1 Plagiarism detector tool

It is a desktop installation used for the textual plagiarism detection with good options for finding a similarity and plagiarism in textual documents.It enables users to correspond on the Internet a single document or a set of documents in folder against a database and obtain good detailed reports displaying the percentage of plagiarism, as we will discover in our experimental section.The plagiarism detector and its database are not wholly without paying.After paying, we can produce an advanced report and store the relevant text file in the database, for checking the plagiarism in query document. The threshold of plagiarism is 10%. The software deals with virtually all types of documents, and we utilized it for detecting the plagiarism in Arabic and English languages.

The plagiarism detector software uses a tf - IDF technique to calculate a similarity in textual documents.

We have performed the experiments using a plagiarism detector software. We have worked with two types of corpus as in [17]. First corpus was developed to test detection of the plagiarism with intrinsic approach, and the second for testing the plagiarism detection with the extrinsic plagiarism approach in Arabic speech, while the other corpus is for detecting the plagiarism in extrinsic plagiarism approach, with English language.

Our experiment processed sixty documents of different volume and topics. For instance, a corpus comprising medical documents, literary subjects, portraying famous people from different domains, describing animal's life, etc.

The results of our experiments are listed in the accompanying tables.

Note: we have listed just 5 documents for every corpus as an exemplar.

4.1.1Extrinsic copies of Arabic text files

4.1.1.1 Big documents (Baseline)

Table 2 plagiarism detector result for extrinsic corpus with big documents.

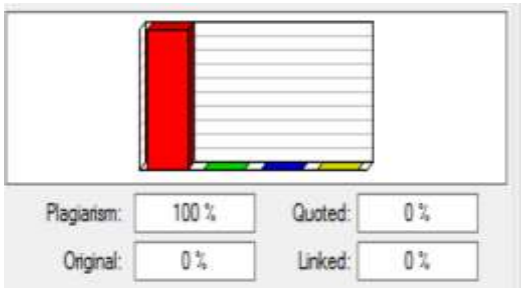
Ser. No.	Doc. Length	Plagiarism %	Quotation %	Originality %
1	4221	96	0	4
2	3334	98	0	2
3	2892	100	0	0
4	8149	99	0	1
5	3764	97	0	3

4.1.1.2 Short documents (summary) of the same previous documents

Table 3 plagiarism detector result for extrinsic corpus with short documents.

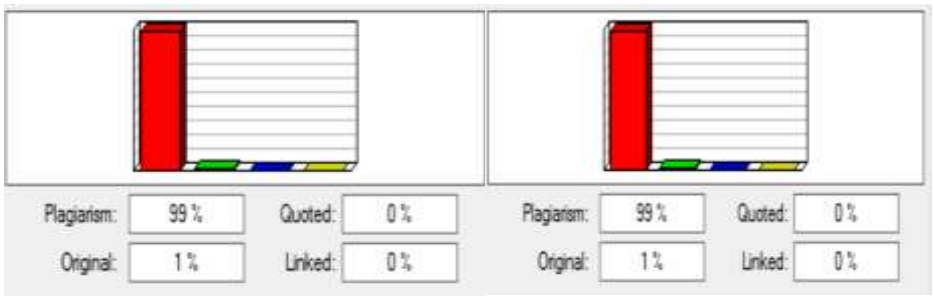
Ser. No.	Doc. Length	Plagiarism %	Quotation %	Originality %
1	244	99	0	1
2	330	99	0	1
3	216	99	0	1
4	262	99	0	1
5	191	99	0	1

(a) (b)

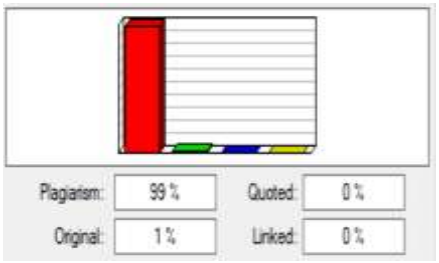


(c)

Fig. (3a,b,& c) statistical of plagiarism detector system for Arabic documents with big size.



(a) (b)



(c)

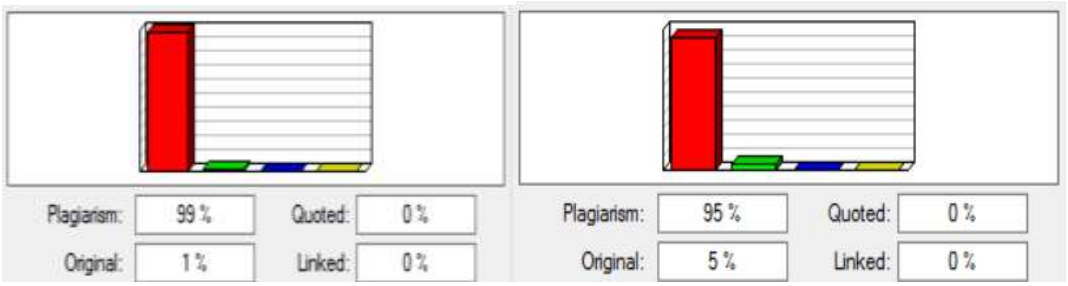
Fig.(4 a,b,& c) statistical of plagiarism detector system for Arabic documents with small size.

4.1.2 Extrinsic corpus for English documents

4.1.2.1 Long documents (Baseline)

Table 4 plagiarism detector result for extrinsic corpus of English language with big documents.

Ser.no.	Doc. Length	Plagiarism %	Quotation %	Originality %
1	3182	99	0	1
2	6321	95	0	5
3	4679	99	0	1
4	3396	96	0	4
5	4617	100	0	0



(a) (b)

Fig. (5 a,b) plagiarism detector statistic for long documents in English language

4.1.2.2 Short documents (summary) of pervious long documents

Table 6 plagiarism detector result for extrinsic corpus of English with short documents.

Sr. no.	Doc. Length	Plagiarism %	Quotation%	Originality %
1	297	99	0	1
2	300	100	0	0
3	263	99	0	1
4	288	100	0	0
5	304	100	0	0

(a) (b)

Fig.(6 a,b) plagiarism detector statistic for short documents in English language

2.3 Intrinsic copiesof Arabic documents

Table 7 plagiarism detector result for intrinsic corpus

Ser.no.	Doc. Length	Plagiarism %	Quotation %	Originality %
1	11195	99	0	1
2	10693	94	0	6
3	929	98	0	2
4	1428	97	0	3
5	1003	90	0	10

4.3 tf-idf calculation and plagiarism dectector tool, results comparing

The accompanying table compares the theoretical and practical calculation of plagiarism detector tool using tf-IDF. In our experiments, we tested three documents of different lengths against single document by using the plagiarism detector tool, then estimated the share of plagiarism using (tf-IDF) technique, as we will see in following table 8.

Tf(w) = number of similar words (wi) / number of whole words in a document (wn) x 100%

Table 8 The result of comparing between theory tf-IDF technique and plagiarism detector tool uses tf-IDF technique

Document	No. similar words (wi)	No. whole doc words (wn)	Tf(w)%	Plagiarism detector software%
D1	14	102	13.7	12.38
D2	625	1461	42.8	41.45
D3	337	4536	7.42	7.24

4.4 Precision and Recall evaluations

Precision and Recall for some documents had checked with a plagiarism detection tool

As we know Precision=$\frac{TP}{TP+FP}$ (8)

And

Recall = $\frac{TP}{TP+FN}$(9)

Where TP is a true positive of information, FP is called false positive, FN is false negative, and there is which called FP false positive. In following table calculation of some documents from different corpus of Arabic and English languages.

Table 9. comparing the precision, recall and F-measure for different corpus tested using plagiarism detector tool

Corpus type	Doc. size	Recall	Precision	F-measure%	Plagiarism%
Extrinsic	4221	0.66	0.35	47	96
Extrinsic	244	0.71	0.45	54	99
Intrinsic	11195	0.78	0.33	32	99
Extrinsic	6321	0.78	0.53	63	95
Extrinsic	300	0.76	0.33	46	100

V. Discussion

Previous equations and calculation showed that it is useful and helpful to make such mathematical calculations to be as small introduction to applied with some practical tools for detecting the plagiarism. In cosine angle for calculating the similarity the results said that d_2 is more similar to the query vector V_q thand_1, because the angle Θ between d_1 and V_q was smaller than between d_2 and V_q. It was just 52.25° between d_1 and V_q while it was 58° between d_2 and V_q.

Commonly, most of previous methods used the words bag for calculating the similarity, and that affected the accuracy of similarity calculations.

On the practical applications, our boards and shapes show a difficult endeavour of software in detecting the similarity.For instance, collecting the resources from the Internet with the "internet check" option in order to call back a vast figure of the most matching resources for computation of the share of plagiarism in documents by using a tf - IDF technique, as observed earlier.

The plagiarism detector tool is a very good fast and helpful, but it is more useful in the case of extrinsic plagiarism detection, while it needs to be more precise in the intrinsic plagiarism detection.

In long documents the plagiarism detector tools had retrieved less information resources than in case of short text files, where retrieved resources are reach hundred resources, which effected the precision and recall of the arrangement as we mention above.

Note: this software has wined in PAN competition2015 [18], where it considers as the best tool had used in that contest.

From comparing table 8 we found that the theory calculations of tf-IDF and plagiarism detector tool processing get almost the same.And it is normal because the system employs the same technique for observing the similarity and plagiarism in suspicious texts.

In precision and recall calculations for extrinsic and intrinsic, with Arabic and English languages, the result showed that, the system has a little bit high Recall, and almost similar, in all types of corpus. While precision graduated between low to medium percentage.

These results are not so accurate so, we can go further with more experiments

VI. Conclusion And Future Work

For observing the similarity and plagiarism in text documents or whatever subject of cognition, it is very useful to recognize the relationship between mathematics calculations and practical measurements and tools used for detecting the problem, that retards the whole between theory and practical works, also names the programmer, or IT researchers, and mathematician corporate and sharing their knowledge for treating and solving such problems.In improver, it will be easier to break off on exactly the pros and cons of whatever methods or principles had created to extend further and further in short time. Our future work adopted new method for detecting the plagiarism cross Arabic and English languages using Braille Language as a pivot.

References

[1]. Albakush, Intisar. "Plagiarism epidemic and plagiarism detecting state-of-art system". 2nd international conference on networking and computer application. Bankok. ISBN. 9788193137352. July. 2016: 170-176.

[2]. Alzahrani, Salha M., Naomie Salim, and Ajith Abraham. "Understanding plagiarism linguistic patterns, textual features, and detection methods." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.2 (2012): 133-149.

[3]. Alzahrani, Salha, et al. "iPlag: intelligent plagiarism reasoner in scientific publications." *Information and Communication Technologies (WICT), 2011 World Congress on*. IEEE, 2011.

[4]. M. Murugesan, W. Jiang, C. Clifton, L. Si, and J. Vaidya, "Efficient Privacy-preserving similar document detection," *VLDB Journal*, 2010.

[5]. C. Lyon, J. A. Malcolm, and R. G. Dickerson, "Detecting short passages of Similar text in large document collections," in *Proc. Conf. EmpiricalMethods Natural Language Processing*, 2001.

[6]. Zu Eissen, Sven Meyer, Benno Stein, and Marion Kulig. "Plagiarism detection without reference collections." *Advances in data analysis*. Springer Berlin Heidelberg, 2007. 359-366.

[7]. Stamatatos, Efstathios. "Intrinsic plagiarism detection using character n-gram profiles." *threshold* 2.1,500 (2009).

[8]. Oberreuter, Gabriel, et al. "Approaches for intrinsic and external plagiarism detection." *Proceedings of the PAN* (2011).

[9]. Bensalem, Imene, Paolo Rosso, and Salim Chikhi. "A new corpus for the evaluation of arabic intrinsic plagiarism detection." *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer Berlin Heidelberg, 2013.

[10]. Sellam, Rahma, et al. "Improved Statistical Machine Translation by Cross-Lingustic Projection of Named Entities Recognition and Translation." *Computación y Sistemas* 19.4 (2015): 701-711.

[11]. Chen, Hsin-Hsi, Changhua Yang, and Ying Lin. "Learning formulation and transformation rules for multilingual named entities." *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15*. Association for Computational Linguistics, 2003.

[12]. Aljohani, Adel, and Masnizah Mohd. "Arabic-English Cross-language Plagiarism Detection using Winnowing Algorithm." *Information Technology Journal* 13.14 (2014): 2349.

[13]. Alzahrani Salha. (2016, November), "Cross-language semantic similarity of Arabic-English short phrases and sentences". *Journal of computer sciences.* DOI: 10.3844/jcssp. (2016, November).
[14]. Atescelik, Esra. "CLUSTER ANALYSIS APPLIED TO EUROPEANA DATA." (2014).
[15]. Mausam. Course/cse573.''Document similarity in information retrieval". (2012, May)<http://Courses.cs.washington.edu/courses/cse573/.....17-ir.pdf>.
[16]. Tf-df::Single –page tutorial- information retrieval and text mining. <ttp://www.tf-idf.com/>. April (2017).
[17]. Bensalem I., Rosso P. , Chikhi S.: A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection. CLEF 2013,Valencia, Spain,September 23-26, Springer.
[18]. Kraus, Christina. "Plagiarism detection state-of-art systems (2016) and evaluation methods". <http://arxiv.org/pdf/1603.03014.pdf> May (2016).