ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 5, No.17 , September 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Discrepancies Detection in Arabic and English Documents

Abdulwahed Almarimi[1] and Gabriela Andrejková[1]

[1]Institute of Computer Science, Faculty of Science, P. J. Šafárik University in Košice
04001 Košice, Slovakia
abdoalmarimi@gmail.com
gabriela.andrejkova@upjs.sk

Abstract

In the paper, there are analyzed and compared results of usable methods for discrepancies detection based on character n-gram profiles (the set of character *n*-gram normalized frequencies of a text) for English and Arabic documents. English and Arabic texts were analyzed from many statistical characteristics point of view. We covered some statistical differences between both languages and we applied some heuristics for measurements of text parts dissimilarities. The results for each text can call for an attention to the text (or not) if the text parts were written by the same author. We evaluate some Arabic and English documents and show its parts they contain discrepancies and they need some following analysis for plagiarism detection. The analysis depends on selected parameters prepared in experiments.

Keywords: *n-grams, stylistic measure, plagiarism, authorship*

1. Introduction

Written texts by the same author should have some similar basic features. For example, the author use the same own language corpus in his texts, the composition of sentences is the same (or very similar) in texts, and so on. Some of the features can't be influenced by author, some can be. The longer documents can be divided into two (or more) parts and it is possible to compare features of the parts. We suppose that stylometric properties are the same (or very similar) in all parts of the same document.

The motivation to the work we found in the papers oriented to intrinsic and external plagiarism, for example [1], [2], [3], [4]. Our main research goal is to compare the stylistic measurements applied in documents in both languages and to analyze if they work in the same way or it is necessary to do some changes of parameters in the analysis.

In this paper, we applied and propose a method for discrepancies detection in documents based on character *n*-gram profiles (the set of character *n*-gram normalized frequencies of a text) and an appropriate dissimilarity measure originally proposed for author identification [5]. Our method automatically creates document segments according to stylistic inconsistencies and decide whether or not a document is discrepancies-free. A set of heuristic rules is introduced that attempt to detect it on either the document level or the text passage level as well as to reduce the effect of irrelevant stylistic changes within a document [5]. The main result was covered for Arabic language where we found that it is better to use *4*-grams than *3*-grams from statistical point of view.

2. Document Analysis

The known methods for an intrinsic plagiarism detection [6], [7] use *3*-grams. The results given by the mentioned methods are quite interesting for English language but we are interesting in the number *n* for *n*-grams, for example why 3-grams are used. The time complexity is a very good argument for it because of the number of *3*-grams in the alphabet with 26 letters is 26^3. The number of *4*-grams is 26^4 and the analysis using *4*-grams should be more time consuming. It is 26^4 and the analysis using *4*-grams should be more time consuming.

We analyzed the documents from statistical point of view. Chosen studied English documents are from PAN 2011 documents corpus. The PAN corpus 2011 (PAN-PC-11) is a corpus for the evaluation comparison of automatic plagiarism detection algorithms. For research purposes the corpus can be used free of charge [8]. Arabic documents were chosen from [9]. We show results of 10 English documents and 10 Arabic documents. The number of letters and words of each document are in the first two rows in the Table 1 and Table 2. The next part of both tables contain the numbers of words with lengths 1-20. It was a full analysis according to the length of words. We present the number of words smaller than 20 letters, the number of longer words is a very small number. We can follow that the words with the length 2, 3, 4 have higher frequency in documents than the others words. There is computed the percentage of two highest frequencies of *n*-grams. The analysis of the *3*-grams and *4*-grams showed that some real words from the language have the highest frequency as *n*-grams.

We will use the following symbols and definitions:

- $|A|$ – the length of the document A.
- ${}^{n}g$ - symbol for *n*-gram.
- $\#o_A({}^{n}g)$ - the number of occurrences ${}^{n}g$ in the document A.
- $f_A({}^{n}g)$ - the frequency of ${}^{n}g$ in the document A defined as

$$f_A({}^{n}g) = \frac{\#o_A({}^{n}g)}{|A|-n+1};$$

- for two documents A, B, $|A| \geq |B|$ let be defined the rate of lengths

$$k = \frac{|B|-n+1}{|A|-n+1} \leq 1$$

2.1 English Documents

Some results on *n*-grams for English documents can be found in [10] and [11]. The statistical analysis of 10 English documents is described in Table 1. We used documents from [8]. The longest/shortest analyzed document E8/E5 has 239881/93085 words and 1148960/417899 letters. In both documents, *3*-grams have highest frequency and *3*-gram "the" is the real word in English language. In some documents (E1, E5), *4*-gram the word "that" has the highest frequency. In the document E4/E8, the word "the" has 9.1% /10.02% occurrences in all document and 24.91%/20.67% occurrences among all word of the length 3.

Table 1: The number of words ordered by the length of 10 English documents (the words of the higher length than 20 have occurrences less than 10).

	Name of documents									
	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
# words	176598	132020	125487	106359	93085	101695	127921	239881	97960	101609
# letters	874761	607783	498696	471566	417899	448699	580411	1148960	399887	452265
# words by length										
1	5111	4642	4642	3474	4018	3034	3579	7297	4015	5167
2	27250* 15.43%	17048	20418* 16.27%	17285	14663	15991	18951	37844* 15.77%	**21826** 22.28%	17674* 17.39%
3	**38733** 21.93%	**27599** 20.90%	**23780** 18.95%	**26497** 24.91%	**19875** 21.35%	**26426** 25.98%	**32097** 25.09%	**49598** 20.67%	20220* 20.64%	**22805** 22.44%
4	26321	20909* 15.83%	18429	19224* 18.07%	18520* 19.89%	17440* 17.14%	19731* 15.42%	35486	17211	17423
5	16646	13328	15587	11459	10109	10874	12787	23311	10762	10024
6	14343	10125	9929	8372	6927	7717	9113	17281	7629	7869
7	12341	9130	9519	6514	6461	6727	8542	15894	4798	6369
8	10599	6686	6040	4020	3717	4071	5962	11690	3044	4581
9	7866	4534	5430	3039	2538	2850	4360	10085	2202	3418
10	4957	3335	3485	2039	1831	1879	3047	6876	1583	2133
11	3590	2081	2160	1384	1227	1253	1897	4781	1037	1319
12	2267	1452	1685	948	828	817	1392	3398	636	917
13	1703	945	1020	614	595	592	907	2337	467	644
14	1058	623	693	415	436	360	586	1564	268	401
15	748	420	543	289	262	227	361	1121	209	268
16	562	276	396	211	179	125	294	803	118	155
17	398	207	285	123	118	88	193	554	63	92
18	273	138	217	74	90	81	131	405	57	56
19	214	85	172	53	58	39	79	276	29	47
20	156	76	75	33	38	28	59	199	36	27
Max frq. *3*-grams	the 18790	the 12808	the 14829	the 9742	the 6313	the 10975	the 13408	the 24047	the 5917	the 10379
Max frq. *4*-grams	that 3026	nthe 1821	ofth 2582	ther 1774	that 1326	nthe 1892	ther 2009	nthe 3360	ther 1521	thes 1591

2.2 Arabic Documents

The statistical analysis of 10 Arabic documents is described in Table 2. We used documents from [9]. Some information on the intrinsic plagiarism we found in [12]. The longest/shortest analyzed document A7/A4 has 93668/31656 words and 375430/135573 letters. In the document A7, *4*-grams have highest frequency and *4*-gram "allah" is the real word in Arabic language. In the document A4, *3*-gram – the word "nal" has the highest frequency. In some documents (A2, A4, A9), "*4*-gram word "fiyal" has the highest frequency.

Table 2: The number of words ordered by the length of 10 Arabic documents (the words of the higher length than 20 have occurrences less than 10).

	Name of documents									
	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
# words	94197	48358	51938	31656	39340	36977	93668	40076	60503	82325
# letters	395065	198019	247448	135573	152905	155301	375430	163212	258346	325972
# words by length										
1	6	48	183	81	14	7	89	690	12	14
2	13217	7358	5365	4816	6188	6397	15396	7456	8079	17911
3	**23287** 24.72%	**12130** 25.08%	9353* 18.00%	**7795** 24.62%	9619* 24.45%	7575* 20.48%	23520* 25.10%	**8405** 20.97%	**12393** 21.48%	**21120** 25.65%
4	22426* 23.80%	11653* 24.09%	**9779** 18.82%	6324 * 19.97%	**11476** 29.17%	**8231** 22.25%	**25075** 26.77%	7850 * 19.58%	10592* 17.50%	19987* 24.27%
5	15887	8336	9175	5417	6134	7190	14198	7252	10540	8295
6	9459	4931	6605	3364	3139	4114	7889	4607	7747	5908
7	5559	2368	4187	2004	1253	2185	3608	2038	4559	2402
8	1978	891	2460	802	516	898	1420	861	1788	1541
9	921	334	973	349	258	175	532	321	678	1646
10	336	91	376	182	125	126	603	138	233	1326
11	205	56	271	161	61	34	388	62	126	755
12	118	24	286	120	41	13	268	41	67	316
13	92	16	304	86	26	19	121	20	60	175
14	64	15	336	51	18	8	61	7	20	76
15	71	14	243	24	11	1	19	8	13	14
16	54	4	175	13	8	0	7	0	3	0
17	42	3	96	15	1	0	2	0	4	2
18	22	2	48	14	0	0	0	0	2	2
19	21	0	18	14	0	0	0	1	1	0
20	7	2	6	0	0	0	0	0	1	0
Arabic	الم	واا	واا	نال	قال	الم	الل	الا	الم	بنع
Latin	alm	waa	waa	nal	qal	alm	all	ala	alm	bina
Max frq. *3*-grams	3027	1797	4242	789	2294	971	3468	1647	1983	3526
Arabic	الله	فيال	هوهو	فيال	الله	لبصر	الله	ثمال	فيال	بنال
Latin	allh	fiyal	huhu	fiyal	allh	lbsar	allah	thmal	fiyal	binal
Max frq. *4*-grams	1479	525	797	346	1986	2230	2958	1030	841	2077

3. Analysis of a Document Style

The main idea is the following: The profile of all document should be the same as a profile of some chosen parts from the document.

We use an approach in which we define a sliding window over the text length and compare the text in the window with the whole document. Thus, we get a function that quantifies the style changes within the document. Then we can use the anomalies (discrepancies) of that function to detect the plagiarized parts. In particular, the peaks of that function (corresponding to text parts of great dissimilarity with the whole document) show discrepancies in the parts. Therefore, what we need is a means to compare two texts knowing that one of the two (the text in the window) is shorter or much shorter than the other (the whole document).

Following the practice of recent methods each text is considered as a bag-of-character *n*-grams. That is, given a predefined *n* that denotes the length of strings, we build a vector of normalized frequencies (over text length) of all the character *n*-grams appearing at least once in the text. This vector is called the ***profile*** of the text. Note that the size of the profile depends on the text length (longer texts have bigger profiles). An important question is the value of *n*. A high *n* corresponds to long strings and better capture intra-word and inter-word information. On the other hand, a high *n* considerably increases the dimensionality of the profile.

3.1 Evaluation of a Documents Similarity

Let *P(A)* and *P(B)* be the profiles of two texts A and B, respectively. We studied the performance of various distance measures that quantify the similarity between two character *n*-gram profiles in the framework of author identification experiments. The following dissimilarity measure has been found to be both accurate and robust when the two texts significantly differ in length [5].

$$d(A,B) = \sum_{{}^{n}g \in p(A)} \left[\frac{2(f_A({}^{n}g) - f_B({}^{n}g))}{f_A({}^{n}g) + f_B({}^{n}g)} \right]^2$$

$$= \sum_{{}^{n}g \in p(A)} \left[\frac{2(k\#o_A({}^{n}g) - \#o_B({}^{n}g))}{k\#o_A({}^{n}g) + \#o_B({}^{n}g)} \right]^2, \quad (1)$$

where $f_A(g)$ is the frequency of occurrence (normalized over the text length) of the *n*-gram g in text A and text B, respectively. k is the rate defined by

$$k = \frac{\#o_B({}^{n}g)}{\#o_A({}^{n}g)},$$

the number of occurrences are in the same rate as the lengths of documents. The maximal contribution *d(A,B)* is 4 for $\#o_B({}^{n}g)=0$. The higher contribution means more differences in frequencies of *n*-grams. Note that d(A,B) is not a symmetric function (typically, this means it cannot be called distance function). That is, only the *n*-grams of the first text are taken into account in the sum. This function is designed to handle cases where text A is shorter than text B. showed that d is quite stable even when text A is much shorter than text B. This is exactly the case in the proposed method for intrinsic plagiarism detection where we want to compare a short text passage with the whole document that may be quite long. In this paper, we modified this measure as follows:

$$nd(A,B) = \frac{\sum_{g \in P(A)} \left[\frac{2(f_A(g) - f_B(g))}{f_A(g) + f_B(g)} \right]^2}{4|P(A)|} \quad (2)$$

where |P(A)| is the size of the profile of text A. The denominator ensures that the values of dissimilarity function lie between 0 (highest similarity) and 1. The value 4 in the denominator follows from the previous analysis. We call this measure normalized measure (*nd*).

3.2 Parameters for Practical Evaluation

We have applied (*d*, *nd*) using the Java program to find (the profile for English and Arabic documents). We want to compare the results based on what has already chosen in Arabic Corpus [9]. The complete set of parameter settings for the proposed method is given in Table 3.

Table 3: Parameter settings used in this study.

Description	Symbol	value
Character *n*-gram length	English	3
Character *n*-gram length	Arabic	4
Sliding window length	w	1000
Sliding window moving	s	100
Threshold of plagiarism free criterion	t_1	0.2
Real window length threshold	t_2	1500
Sensitivity of plagiarism detection	a	2

4. Discrepancies Detection

Let W be a sliding window of length *w* (in letters) and step *s* (in letters). The windows will be moved in each time to the right by s letters and the profile will be computed for each window W. If w > s the windows are overlapping. It is possible to define the style function of a document A as follows:

$$sf(i,A) = nd(W_i, A), i = 1..\lceil |A|/s \rceil,$$

where W_i is a window, $\lceil |A|/s \rceil$ is the total amount of windows (it depends on text length). It means each window in the document will be evaluated in a comparison to all document. We expect that the style function is relatively stable (it does not change value dramatically) if the document is written by the same author. If the style function has very different values (some peaks [5]) for different windows, it is necessary to analyze the covered parts. The existence of such peaks can be indicated by the standard deviation. Let S denote the standard deviation of the style function. If S is lower than a predefined threshold, then the document is considered plagiarism-free.

In the first step we worked according to recommendations found in Stamatos [5]. The problem is much harder since if the windows are too long the stylistic anomalies would correspond to the style of the alleged author rather than to the sections with discrepancies. Stamatos [5] supposed that at least half of the text is not plagiarized so that the average of style function would indicate the style of that author.

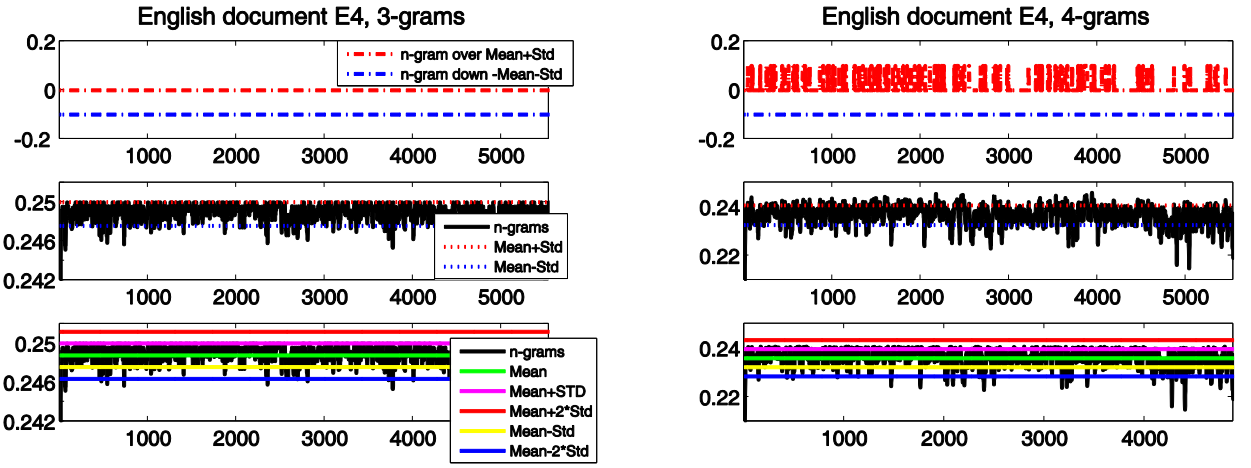

Fig. 1 The style function of English document E4, the window moving by 100 positions using *3*-grams (left) and *4*-grams (right). Down panels represent style function *fs*, The panels in the middle line represent modified style function, and the highest panels represent peaks of values (binary values). The binary function above indicates probably plagiarized passages (high values).

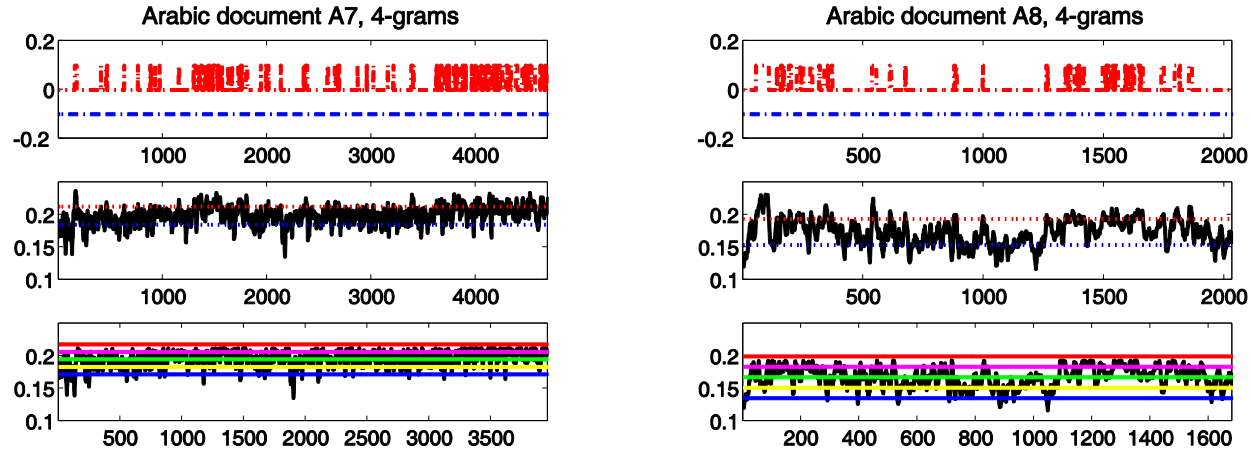

Fig. 2 The style function of Arabic documents (A7, A8), the window moving by 100 positions using *4*-grams. The binary function above indicates passages probably with discrepancies (high values). In both documents were found discrepancies. The legend used in the Fig. 1 is applied in this figure too.

However, the calculation of the average *sf* value would inevitably involve the plagiarized passages as well.

Let *M* and *S* denote the mean and standard deviation of *sf*, respectively. We first remove from *sf* all the text windows with value greater than $M + S$. These text sections are highly likely to correspond to plagiarized sections.
Let *sf* (*i'*, A) denote the style function after the removal of these sections. Let *M'* and *S'* be the mean and standard deviation of *sf (i', A)*. Then, the following criterion was used to detect plagiarism, Stamatos [8]:

$$sf(i', A) > M' + a * S' \quad (3)$$

where parameter *a* determines the sensitivity of the plagiarism detection method. The higher the value of *a*, the less (and more likely plagiarized) sections are detected. The value of the parameter *a* was determined empirically at 2.0 [5] to attain a good combination of precision and recall. We used recommended value for the parameter *a*.

5. Evaluation

After preprocessing of the analyzed documents we applied the described algorithm to them. We illustrate the results of prepared algorithms on chosen documents in Fig. 1 and Fig 2. In the Fig. 1, it is analyzed the English document E4 using *3*-grams (left 3 panels) and *4*-grams (right 3 panels). The applied criterion in the formula (3) to the same document E4, using *3*-grams and *4*-grams gives the different results: using *3*-grams the document has not discrepancies, using 4-grams it has them. The legend used in the Fig. 1 is used in all the following figures. In the Fig 2, there are analyzed two Arabic documents A7 (left 3 panels) and A8 (right 3 panels). In the middle line panels, we can see results of style functions *sf*, the blue lines represents the values $M-S$, the red lines represent the values $M+S$. Down panels contain the style functions *sf'* for the reduced number of windows. The green lines represent the means, the yellow respectively magenta lines represent the values $M' + S'$, $M' - S'$ respectively and red and blue lines show values $M' + 2S'$, $M' - 2S'$. The top panels with binary values indicate the values higher than $M + S$ and lower than $-M - S$.

In Fig.1, we illustrate that the analysis using *3*-grams has not problems with the discrepancies according to prepared parameters but the analysis using *4*-grams gives the same results after the reduction of *sf* to *sf'*. In Fig. 2, both Arabic documents are analyzed using *4*-grams and according to prepare parameters they have any discrepancies.

In Fig. 3 we illustrate *sf* function of combined documents to show that it has a different course. The first part was written by one author and the second part by another author. The plagiarized part is the part with higher values of *sf* function (the part has more differences to the all document).

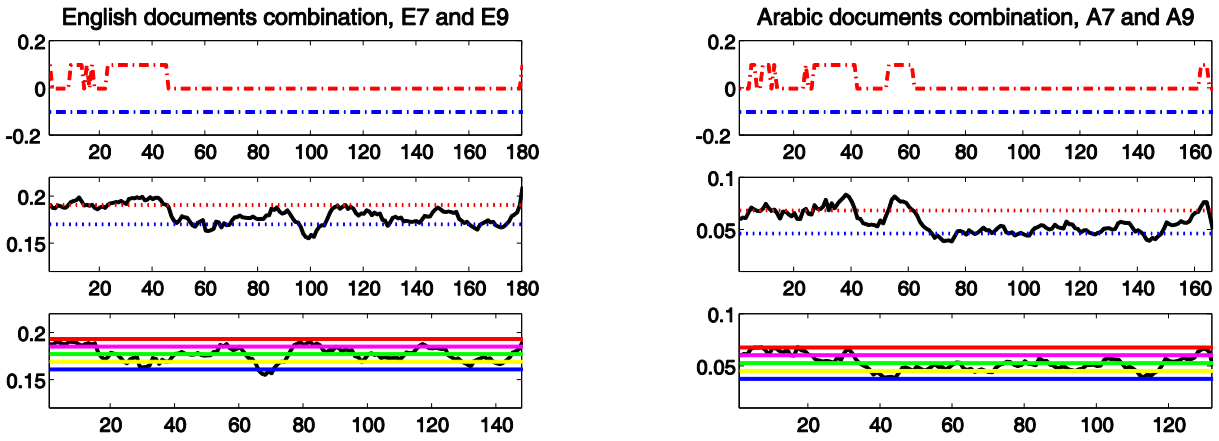

Fig. 3 The style function (the left panel) of combined two English documents E7 (7113 first characters) and E9 (8155 first characters) and (the right panel) two Arabic documents A7 (5347 characters) and A9 (8155 first characters), the window moving by 100 positions using *3*-grams for English documents and *4*-grams for Arabic documents. The binary function above indicates probably discrepancies in passages (high values). The legend used in the Fig. 1 is applied in this figure too.

6. Conclusion

One of the known methods for the intrinsic plagiarism detection is the method using *n*-grams. But usually *3*-grams are used. We do not write about a plagiarism, we write about discrepancies in documents that should be the base for plagiarism. We suppose some semantic analysis to write that document obtains some plagiarized parts. Our expectation was that different languages have to be analyzed by different *n*-grams, especially languages with different word compositions and different lengths of words. In the paper we showed the statistical analysis of Arabic language and cover that using *4*-grams are better for it. The prepared analysis is very formal and in the next work we will try to apply some new accesses to the problem.

Acknowledgements
The research is supported by the Slovak Scientific Grant Agency VEGA, Grant No. 1/0142/15.

References

[1] S. Meyer zu Eissen, B. Stein, "Intrinsic Plagiarism detection", Lalmas et al. (Eds.): Advances in Information Retrieval Proceedings of the 28th European Conference on IR Research, ECIR 2006. London, ISBN 3-540-33347-9, c Springer 2006, pp. 565-569.

[2] S. Meyer zu Eissen, B. Stein, M. Kulig, "Plagiarism detection without reference collections", In: Decker, R., Lenz, H.J. (eds.) GfKl. pp. 359 - 366. Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin / Heidelberg (2006), http://dblp.unitrier.de/db/conf/gfkl/gfkl2006.html

[3] G. Oberreuter, G. LHuillier, S. A. Ríos, and J. D. Velasquez, "Approaches for Intrinsic and External Plagiarism Detection", Notebook for PAN at CLEF 2011

[4] M. Potthast, A. Barrón-Cedeño, A. Eiselt, B. Stein, P. Rosso, "Overview of the 2nd international competition on plagiarism detection", In: Braschler, M., Harman, D. (eds.) Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy (2010) .

[5] E. Stamatatos, "Intrinsic Plagiarism Detection Using Character n-gram Profiles", SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, PAN'09, pp. 38-46, 2009.

[6] E. Stamatatos, "Authorship attribution based on feature set subspacing ensembles", International Journal on Artificial Intelligence Tools, 15.5, pp. 823-838.

[7] E. Stamatatos, "Ensemble-based author identification using character *n*-grams", In Proceedings of the 3rd International Workshop on Text-based Information Retrieval, pp. 41-46, Riva del Garda, Italy

[8] pan-plagiarism-corpus-2011.part1.rar, http://www.uniweimar.de/en/media/chairs/webis/corpora/pan-pc-11/

[9] King Saud University Corpus of Classical Arabic. http://ksucorpus.ksu.edu.sa.

[10] F. I. Haj Hassan, M. A. Chaurasia, "*N*-Gram Based Text Author Verification", IPCSI vol. 36 (2012), IACSIT press, Singapore.

[11] M. Kuta, J. Kitowki, "Optimisation of Character n-Gram Profiles Method for Intrinsic Plagiarism Detection", ICAISC 2014, Part II, LNAI 8468, pp. 500-511, 2014.

[12] I. Bensalem, P. Rosso, S. Chikhi, " A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection", CLEF 2013, LNCS 8138, pp. 53-58, 2013.

First Author: Master of Science from P. J. Šafárik University in Košice, Institute of Computer Science, Faculty of Science, now PhD. Student.

Second Author: Associate Professor, Deputy director for academic affairs in of Computer Science, Faculty of Science, P. J. Šafárik University in Košice. She is interested in artificial neural networks and in stringology.

.