Writer Identification Based on Letter Frequency Distribution

Polina Diurdeva, Elena Mikhailova, Dmitry Shalymov
Saint Petersburg State University
St. Petersburg, Russia
polina.durdeva@ya.ru, e.mikhaylova@spbu.ru, dmitry.shalymov@gmail.com

***Abstract*—Lately writer identification problem has become actual due to huge amount of documents in digital form. In the current work an approach based on frequency combination of letters is investigated for solving such a task as classification of documents by authorship. This research examines and compares four different distance measures between a text of unknown authorship and an authors' profile: L_1 measure, Kullback-Leibler divergence, base metric of Common *N*-gram method (*CNG*)[8] and certain variation of dissimilarity measure of *CNG* method which was proposed in [12]. Comparison outlines cases when some metric outperforms others with a specific parameter combination. Experiments are conducted on different Russian and English corpora.**

I. Introduction

Due to increased amount of available documents in digital form, usage of a method for digital documents processing has become crucial. The challenge of digital documents processing includes wide variety of problems and, in particular, a problem of author attribution is among them. The latter at the same time includes the following subtasks: author identification, verification, author profiling or characterization and others[2]. The problem of author attribution becomes extremely important because of wide distribution of anonymous, faked data in the network nowadays. Also this problem is significant in linguistic, criminalistic and historical research.

In this paper, we considered tasks of writer identification. The process of writer identification can be defined as determination of an author by general and particular texts' properties that form a writer's style. A special case of the author identification problem is the classification of documents by authorship. The latter can be stated given that an unknown author is one of the predefined set of candidates whose authorship can not be disputed [9].

The writer's style is formed by extracting style marker (stylometric features) from their text[8]. Extracting of the most style markers types require NLP preprocessing for their measurement which can be complicated and time-consuming. For example, calculation of word frequencies requires tokenizer, stemmer, lemmataizer, at the same time part-of-speech tagging needs tokenizer, sentences splitter, POS tagger, etc. However, there are simpler methods that do not need much preprocessing and are not less effective. For instance, today more and more researchers pay attention to *N*-gram approach. *N*-gram approach implies operating a text as a set of *N* character combinations. This approach is tolerant to spelling and grammatical errors[10] and does not demand sophisticated preprocessing of input text, sometimes only basic filtering is needed: removing spaces or/and punctuation marks. One of the main issues of the approach is choosing appropriate method for solving the problems of author attribution.

In this paper, we discuss profile-based methods and compare some variations with each other. These methods use a concept of profile - formal representative of author's style which is the character *N*-gram distribution. A mapping between authors' styles and profiles is established and the problem is solved by comparing profiles (distributions) using a special metric function. We investigated four metrics: L_1 measure, Kullback-Leibler divergence which were presented by Y. N. Orlov and K. P. Osminin in [7], base metric of Common *N*-gram method (*CNG*)[8] and certain variation of dissimilarity measure of *CNG* proposed by E. Stamatatos in [12].

II. Related works

Problems of authorship attribution have been studying over the 19-21 centuries and now have a wide background and a long history. The pioneers in this field are Mosteller and Wallace, who applied the approach based on Bayesian statistical analysis of common word frequencies to solve an author identification problem of The "Federalist Papers"[1] in 1964. Research in this field of science has been advancing ever since. As a result, today a comprehensive variety of tools for solution of aforementioned problems are provided.

In [2], [4], [3] different approaches, methods and wide va-riety of feature sets are reviewed and compared between each other. Furthermore, advantages and disadvantages of each are provided. There is more detailed theoretical background and description of particular problems of author attribution. Au-thors give various recommendations regarding circumstances under which their methods and approaches should be used or not.

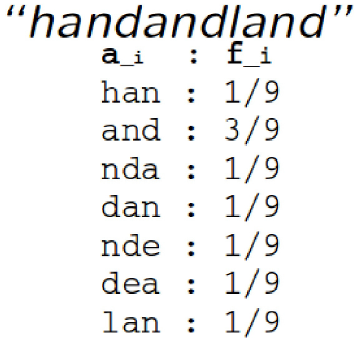
One of the most popular today approaches is one based on frequencies of character *N*-grams occurrences. Generally, a feature set is represented as *N*-gram probability distribution over *L* most frequent character *N*-grams. This approach is reported to be surprisingly efficient [2]. In [8] Common N-gram method (*CNG*) based on byte level *N*-gram is provided. *CNG* method is based on dissimilarity function between *N*-gram probability distributions. The best results for $1{,}000 \leq L \leq 5{,}000$ and $3 \leq N \leq 5$ were reported on Greek and English data sets. In [12] new dissimilarity measure is presented for the *CNG* method. Proposed modification was more stable in case of limited and imbalanced corpora. In [14] the profile-based method was explored for authorship verification problem. It is also some effective modification of *CNG* approach. Another alternative similarity measure for *CNG* method was proposed in [17] for task of Source Code Author Profiling(SCAP). This metric counts the amount of common *N*-grams of two profiles and is called simplified profile intersection (SPI). More complex *N*-grams-based method is investigated in [9], [11]. This approach uses a classifier ensemble to assign posterior probabilities of belonging to a possible class and rules of com-bining the probabilities to obtain combined decision. Results which are gained by this approach confirm potential of using *N*-gram approach for solving authorship problems.

N-grams-based approach is successfully applied along with machine learning method often enough. In [15], [16] new technique of *N*-gram constructed is provided. The technique enables to build *N*-grams basing on syntactic relations in syntactic trees and not on character sequence. Usage of SVM classifier along with this method leads to the most significant results.

III. Methodology

A. The description of the algorithm

During the identification of an author of input text it is assumed that the text displays the individual writing style of an author, which allows to distinguish it from the others. In order to compare different texts to each other, the texts must be provided with a certain numerical characteristic, which would be similar for texts of one author, and differ significantly for works from different authors. For this purpose, each text of the corpus associates is characterized with *N*-gram occurrence frequencies of letter combinations profile. The example of *N*-gram profile is presented in Fig. 1.



"handandland"

a_i	:	f_i
han	:	1/9
and	:	3/9
nda	:	1/9
dan	:	1/9
nde	:	1/9
dea	:	1/9
lan	:	1/9

Fig. 1. Three-grams example

The profile D of a text is defined as the set of pairs $\{(a_1, f_1), (a_2, f_2), (a_3, f_3), \dots\}$ where f_i is normalized occurrence frequency of *N*-gram a_i in the text. Moreover, texts, which authorship is determined clearly, all together form their author's profile. Thus, the next step is to determine the distance between a text's profile and the author's profile. The author with smallest distance to the profile will be considered a creator of given text. The algorithm of classification of texts by authorship can be formulated as presented in Listing 1.

B. Distance metrics

We examined and compared four distance metrics. The first metric was L_1 norm, that is defined by Eq. 1. The metric L_1 was considered in [7], [6], [5].

Algorithm 1 Generic *N*-gram classification algorithm

```
$T \leftarrow$ unseen set of texts
$t \leftarrow$ set of texts with known authorship
$A \leftarrow$ set of authors
$tr(a)$ returns all texts of the author $a$ from set $t$
for $a \in A$ do
   build author profile $D_a$, where
   $D_a = D$ of concatenation of all texts in $tr(a)$
end for
for $x \in T$ do
   build profile $D_x$
   $a^* \leftarrow \arg\min_{a}(distance(D_a, D_x))$
end for
```

$$L_1(D, D_a) = ||D - D_a|| = \sum_{i=1}^{\alpha(N,M)} |D(a_i) - D_a(a_i)| \qquad (1)$$

where M is a the alphabet power of the language of given data set and accordingly $alpha(N,M) = M^N$ is the number of *N*-grams.

The second metric taken for comparison was Kullback-Leibler divergence (KL), which is often used in information theory. This metric was also noted in [7]. KL is asymmetrical metric. The distance function is defined by Eq. 2.

$$KL(D, D_a) = \sum_{i=1}^{\alpha(N,M)} D_a(a_i) \log \frac{D_a(a_i)}{D(a_i)} \qquad (2)$$

The next metric was dissimilarity measure of Common *N*-grams method, which which was first proposed in [8] and determined by Eq.3.

$$CNG(D, D_a) = \sum_{i=1}^{\alpha(N,M)} \left(2\frac{(D(a_i) - D_a(a_i))}{D_a(a_i) + D(a_i)} \right)^2 \qquad (3)$$

The last taken function was proposed in article. The metric is not symmetric and takes into account only those *N*-grams a_i that are included in text profile D. This metric showed considerable results when only limited texts per author are available for training. It is a modification of basic *CNG* metric [12]. Formal definition of the last metric is presented in Eq. 4.

$$CNG_M(D, D_a) = \sum_{a_i \in D} \left(2\frac{(D(a_i) - D_a(a_i))}{D_a(a_i) + D(a_i)} \right)^2 \left(2\frac{(D(a_i) - D_C(a_i))}{D_C(a_i) + D(a_i)} \right)^2 \qquad (4)$$

where D_C is the set of normalized *N*-grams frequency of the given corpus, namely, corpus's profile [12].

C. Algorithm parameter specification

It should be emphasized that even though discussed algorithm 1 is relatively simple the most complex task is estimation of parameters with which the algorithm is parameterized. The parameters of the algorithm 1 is shown below.

- N - length of a character combination
- L - a profile size, amount of the most frequent *N*-grams. The number of *N*-grams in a profile was

limited to consideration of L the most frequently encountered in text N-grams.

- T/t - train/test ratio of texts per an author
- *size* - text length, characters. Fragments with a given length *size* were chosen from a randomly selected position in the text.

The first two parameters are basic parameters of investigated algorithm and they demand being tuned to obtain the best results on a corpus. The last two parameters are more general for task of authorship attribution and they are more likely to be called constraints of the task rather than algorithm parameters. In case of CNG and CNG_M metrics it was shown that $3 \leq N \leq 5$ and $1000 \leq L \leq 5000$ give the best results in most of the cases.

In this paper, we considered $N = 3$ and tried to estimate parameter L in case of using L_1 and KL metrics. The choice of N is related to results of other research [12] where it was shown that using 3-grams gives more accurate results. Our small personal research on this topic showed that generally using 3-grams performs better than using 4-grams. Also, it was important to investigate applicability of L_1 and KL metrics in generally. Different data sets and constraints, which were combinations of T/t and *size*, were considered in order to evaluate and compare the accuracy of the metrics on taken corpora. We tried to tune the parameters by comparing performances of algorithm with different combinations from predefined sets (except for certain cases):

- T/t : $\{2/4, 3/3, 1/5\}$
- L : $\{500, 700, 1000, 1500, 2000, 3000, 4000\}$
- *size* : $\{20000, 30000, 40000, 50000\}$

IV. Data set

In current research we conducted multiple experiments on different data sets: English data set, Russian data set, PAN12, which also contains English texts.

A. English data set

We made an English data set which contained texts of English literature. Texts of 19 authors formed the corpus, the following authors were among them: Charles Dickinson, Gilbert Keith Chesterton, Arthur Charles Clarke, Marie Corelli, Arthur Conan Doyle, Thomas Stearns Eliot, Edward Morgan Forster, John Galsworthy, Elizabeth Gaskell, George Gissing, Thomas Hardy, Joseph Conrad, Vernon Lee, Ford Madox Ford, Stephenie Meyer, Terry Pratchett, J. K. Rowling, J.R.R. Tolkien, Roger Zelazny. Lengths of texts from the data set were 60,000 - 1,500,000 symbols (without punctuation and gaps). Each author had 6 texts, which were divided into a training and a test sets according to the parameter T/t.

B. Russian data set

A Russian data set was build to test performances of different metrics on different language. Texts of 19 authors of Russian fiction were chosen from the corpus [20]: Paul Amnuel, Kir Bulychev, Ivan Efremov, Sergey Kazmenko, Leonad Kudryavtsev, Elena Khaetskaja, George Lockhart, Svyatoslav Loginov, Henry Lion Oldie, Victor Pelevin, Gennady Prashkevich, Andrey Schupov, Boris Shtern, the brothers Arkady and Boris Strugatsky, Alexander Tyurin, Garm Vidar, Yevgeny Lukin, Michael Veller, Lyubov and Yevgeny Lukiny. Lengths of texts from the data set were 50,000 - 1,000,000 symbols (without punctuation and gaps). Each author had 6 texts.This dataset is used in "Lingvoanalizator" sofware developement [21]. In this software system entropy and Markov chain approaches are applied. The reported accuracy of the system varies from 84 % to 89%.

C. PAN12 data set

The following series of experiments was conducted on data set which were presented on the PAN 2012 competition [18] to evaluate the relative accuracy. Competition results are presented in [18]. There were six problems of straightforward authorship attribution and we have considered three of them.

The problem A uses training and test sets which contain two samples each (six samples in total) for each of three authors. All samples were between 1800 and 6060 words.

The problem C contains texts with length of about 13000 words. There were 8 authors and two and one sample per author for training and test set respectively.

The problem I uses novels of 14 authors the lengths of which are 40,000 – 170,000 words. Test/training data sets capacities were in the 1/2 ratio.

V. Experiments

Below we provided detailed descriptions with data sets mentioned above and tables containing configurations of experiments with the most considerable results.

A. Experiments on English data set

Results of experiments on English data set are presented in Table I, Table II, Table III. To evaluate average accuracy of each metric on English texts we conducted several experiments with different L on texts of length *size*. Table I demonstrates the average values of accuracy which were obtained for L from aforementioned sets. The accuracy was calculated as the percentage of correctly recognized authors.

The best obtained results in the experiment are listed in Table II. The parameters with which significant results were also achieved are reported in Table II.

Moreover, a series of experiments with no limit on the *size* was conducted and texts with their original lengths was taken into consideration. Thus, the training and test sets were imbalanced and depend on formed them texts. Average accuracy values were obtained for $L = \{300, 400, 500, 800, 900, 1000, 2000, 3000, 4000, 5000, 8000\}$.

B. Experiments on Russian data set

Results of analogous experiments on Russian Data Set are presented in Table IV, Table V and Table VI.

Likewise, the parameter values in case of maximum accuracy are shown in Table V.

TABLE I. The average accuracy for English data set

size	CNG_M,%	CNG,%	L_1,%	KL,%
$T/t = 1/5$				
20000	42.7	16.4	47.0	34.8
30000	53.5	24.8	53.3	36.0
40000	56.5	24.6	55.3	40.7
50000	58.4	29.0	56.8	48.0
$T/t = 2/4$				
20000	59.8	29.0	56.8	50.9
30000	63.8	36.8	62.7	48.6
40000	66.1	45.0	69.3	56.8
50000	70.9	48.2	71.0	55.6
$T/t = 3/3$				
20000	64.8	36.0	65.8	54.4
30000	62.5	40.1	66.1	56.4
40000	70.4	50.5	78.3	62.5
50000	73.2	51.3	76.0	64.5

TABLE II. The best results on English data set

metric	size	L	accuracy, %
$T/t = 1/5$			
CNG_M	50000	500	63.8
CNG	50000	700	60.6
L_1	50000	3000	66.0
KL	50000	2000	60.6
$T/t = 2/4$			
CNG_M	50000	700	80.0
CNG	40000	500,700	77.3
L_1	50000	1500	77.3
KL	40000	2000	74.6
$T/t = 3/3$			
CNG_M	50000	4000	84.0
CNG	40000	1000	85.7
L_1	40000	1000;3000	85.7
KL	40000	4000,3000	80.4

TABLE III. The average and maximum accuracy for English data set without limitation on the size of text (*size*)

	CNG_M	CNG	L_1	KL
$T/t = 1/5$				
max	79.8	78.7	83.0	63.8
avg	63.1	65.2	69.1	45.3
$T/t = 2/4$				
max	86.7	86.7	86.7	84.0
avg	75.4	79.9	81.3	58.1
$T/t = 3/3$				
max	89.3	94.6	94.6	91.1
avg	81.8	87.0	88.6	67.4

TABLE IV. The average accuracy for Russian data set

size	cng_st	cng	L_1	KL
$T/t = 1/5$				
20000	42.5	20.4	43.6	19.1
30000	52.1	25.7	52.1	21.1
40000	52.9	30.5	53.0	23.8
50000	62.3	34.3	54.3	30.9
$T/t = 2/4$				
20000	65.6	35.3	59.6	25.9
30000	65.0	41.7	66.1	26.3
40000	75.4	47.8	70.8	27.5
50000	78.1	55.4	74.8	31.5
$T/t = 3/3$				
20000	71.1	41.7	67.9	29.8
30000	76.8	51.8	77.4	30.4
40000	78.9	57.7	81.0	36.0
50000	79.8	64.3	80.7	37.8

TABLE V. The best results on Russian data set

metric	size	L	accuracy, %
metric	size	L	accuracy, %
$T/t = 1/5$			
CNG_M	50000	500	63.8
CNG	50000	500	60.0
L_1	40000	700	67.5
KL	50000	4000	45.0
$T/t = 2/4$			
CNG_M	40000	3000	86.0
CNG	50000	700,1000	81.3
L_1	50000	1000	82.8
KL	50000	3000	46.9
$T/t = 3/3$			
CNG_M	40000	1500	87.5
CNG	50000	700	90.0
L_1	40000;50000	4000;700	90.0
KL	50000	4000	60.4

TABLE VI. The average and maximum accuracy for Russian data set without limitation on the size of text (*size*)

	CNG_M	CNG	L_1	KL
$T/t = 1/5$				
max	67.5	66.3	66.3	37.5
avg	55.2	43.9	53.2	26.3
$T/t = 2/4$				
max	85.9	85.9	84.4	75.0
avg	78.6	74.7	80.1	36.4
$T/t = 3/3$				
max	97.9	93.8	95.7	77.1
avg	86.0	83.1	86.7	44.1

In the following experiment the texts with original length were similarly examined and average accuracy values were obtained for $L = \{300, 400, 500, 800, 900, 1000, 2000, 3000, 4000, 5000, 8000\}$. The results are presented in Table VI.

Also, in addition to the experiments described above, we conduct experiments on Russian and English data sets in which we considered neither *size* nor L in. These experiments showed that accuracy of distance metrics CNG and CNG_M on mentioned corpus is not more than 80%, KL is not more 68%, whereas $L1$ is 80 - 95%. Summarized, exploration on English end Russian data sets reveal the following:

- Applying mentioned approach leads to more accurate results on Russian data set. It can be related to specifics of taken $N = 3$ for this language and demand additional research to figure it out.
- Using L_1 and CNG_M metrics on balanced test and training sets (with limitation on *size*) gives more

precise results.

- Using L_1, *CNG*, CNG_M metrics on imbalanced data sets which contains sufficiently long texts enable to achieve 80%-98% accuracy. It can be seen that L_1 norm mostly outperforms other metrics in the case of larger volume of training and test set.
- A predefined profile length L for L_1 and *KL* distance metrics should be larger ($4000 \leq L \leq 8000$) than for *CNG* and CNG_M to achieve the best results. As well, high accuracy is achieved without restriction on profile size. That enables to narrow space of the unknown parameters, and thus simplifies the process of solving author identification problem.

C. Experiments on PAN12 data set

In Table VII the best results of experiments on PAN12 data set are presented. Accuracy is presented as r/w, where r – the number of correctly classified texts and w – the total number of texts. The parameter *size* was not considered whereas L took values $L = \{300, 400, 500, 800, 900, 1000, 2000, 3000, 4000, 5000, 8000\}$.

TABLE VII. THE BEST RESULTS FOR PAN12 DATA SET

problem	*CNG*	CNG_M	L_1	*KL*
A	6/6	5/6	6/6	5/6
C	6/8	6/8	7/8	5/8
I	13/14	10/14	12/14	12/14

The results obtained in this experiment demonstrate high competitiveness of the algorithm 1 with using the L_1 and *CNG* metrics compared to most other algorithms for solving problems A, C and I.

VI. Additional experiments

Our experiments showed that L_1 norm can be successfully used for solving authorship problem with the above approach with high accuracy along with *CNG* and CNG_M metrics which were thoroughly investigated in parallel works [8] [12]. However, we think that L_1 norm has not been explored enough considering its simplicity and high precision relative to its more complex analogues. Therefore we decided to conduct extra experiments aimed to more detailed exploration of L_1 norm applicability. Significant results on relatively long texts made attempts to use L_1 on shorter texts look promising.

We took a dataset which consisted of the English journalistic texts. The initial dataset was "Reuter_50_50 Data Set" of [13]. This data set contains news articles on the topic of "Corporate/Industrial". This fact allows to minimize the factor of thematic difference between texts.

This data set consists of 5000 articles of 50 authors: 50 training texts and 50 test texts for each author. The average size of articles is 2455 characters. It should be noticed that the variation of the texts' lengths is quite large: there are articles longer than 4 thousand characters and less than 300 characters. The last fact enables to consider the case of imbalanced corpus.

The result of classification experiment on all corpus achieved 67,8% accuracy. The factor that may affect the accuracy of the classification is the number of training data. In our case the volume of an author's profile is the direct measure of training data amount as it is built upon concatenation of all training texts. Dependency on author's profile size creates a problem of time-consuming search of author's texts of the above size, since the number of journalistic texts belonging to an author during solving real tasks may be much less.

The next experiments were conducted in order to investigate dependency of classification accuracy on author's profile size. For each author 30 test texts were taken, while training texts amount differed. The initial number of training texts for each author was equal to 6, then author's number of texts was consistently increased. Results for conducted experiments for 10, 30, 40, 50 authors are presented in Fig. 2.

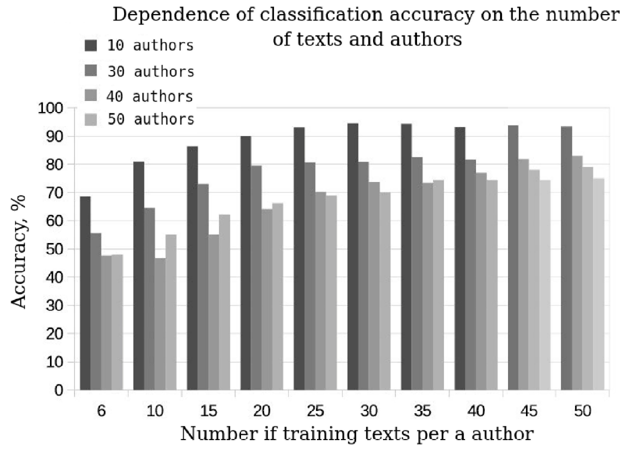


Fig. 2. Dependence of classification accuracy on the number of texts and authors

It can be seen that behavior of an accuracy plot for each author amount is the same : the plot increases first and stabilizes after some text amount (approximately 25 texts).

We decided to investigate more thoroughly the dependency of accuracy on L parameter. To implement this we took texts of 20 authors. For each author 30 of all their texts were considered as train along with 20 as test. L was taken from the set of values $\{500, 800, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 6000, 8000, 10000, 15000, max\}$ where max is all possible N-grams. Results for conducted experiments are presented in Fig. 3. The best accuracy equal to 91% was reached without L-constraint on profile size.

VII. Conclusion

In this paper we have investigated the method for writer identification based on letter frequency distribution. We examined profile-based Common N-Gram method (*CNG*) with different metrics to calculate the proximity between author' and texts' profiles:

- L_1 measure,
- Kullback-Leibler divergence (*KL*),
- base metric of *CNG*

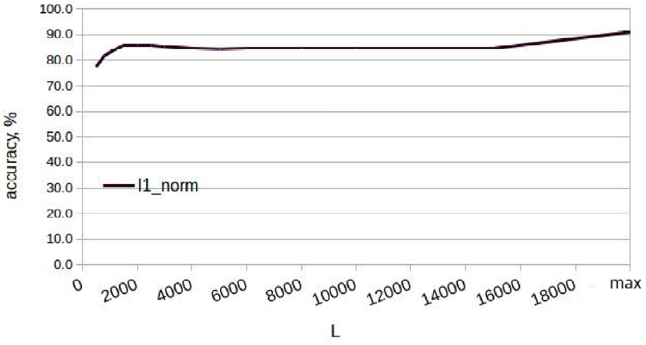


Fig. 3. Dependence of classification accuracy on L constraint

- certain variation(CNG_M) of dissimilarity measure of CNG, proposed by E. Stamatatos in [12].

The analysis was conducted on Russian and English literary texts, on PAN12 corpora and on a variety of nonfiction English texts [13].

Common N-gram method has a general problem of N (length of character combinations extracted from texts) and L (profile size - number of most frequently used N-grams) parameters tuning. We conducted experiments with using of three-grams for each metric with different L values. Experiments confirmed that base CNG metric and $CNGM$ metrics achieve high accuracy perform with bounded L (1000 - 5000), whereas KL and L_1 norm achieve maximum accuracy mostly on L more than 4000. However, it also can be observed that L_1 reaches its best results without L constraints. Using L_1, in contrast to other metrics, L tuning task can be avoided.

It can be undoubtedly concluded that CNG_M metric outperforms others on data sets with small training data capacities. However, it can be noticed that in more generic cases, like imbalanced training and test sets capacities corpora, L_1 norm shows slightly more accurate results than others.

Generally, KL metric showed worse results than other metrics, while L_1 showed comparably results with well-explored CNG and CNG_M metrics. Due to this observations experiments on Reuter data set only with L_1 norm were conducted.

More exact relationships between the classification accuracy and the number of training data, the number of recognized authors and L parameter have been determined.

A limitation of the method on small texts is a low separation ability in case of a large number of authors (highest accuracy equal to 94 % is achieved on 30 training texts for 10 authors). Limitations also include the requirements for number of training data which is large enough. However, experiments showed that limiting L to small values does not improve accuracy. Also the additional analysis of the metric has shown that it gives quite good results not only for long text, but also for quite short texts. These facts positively distinguishes this method from others.

In further research it is possible to compare applicability of machine learning methods (e.g. SVM)[23] to the authorship attribution problem with current approach.

VIII. Acknowledgments

This research is supported by Saint-Petersburg State Unversity grant 6.37.181.2014.

References

[1] F. Mosteller, D. Wallace, *Inference in an authorship problem - a comparative-study of discrimination methods applied to authorship of disputed Federalist Papers*, Journal of the American Statistical Association, 58 (302), pp. 275-309, 1963

[2] E. Stamatatos, *A survey of modern authorship attribution methods* Journal of the American Society for Information Science and Technology, 60 (3), pp. 538-556, 2009

[3] P. Juola, *Authorship attribution. Foundations and Trends in Information Retrieval*, 1 (3), pp. 233-334, 2006

[4] M. Koppel, J. Schler, S. Argamon, *Computational methods in authorship attribution*, Journal of the American Society for Information Science and Technology, 60 (1), pp. 9-26, 2009

[5] L. A. Borisov, Y. N. Orlov, K. P. Osminin *Identification of an author of a text based on a distribution of frequencies of letter combinations*, Applied Informatics, T. 26. No 2, pp. 95-108, 2013

[6] Y. N. Orlov, K. P. Osminin, *The definition of the genre and the author's literary works based on statistical methods*, The Preprint IPM by M. V. Keldysh, No27, 26C, Web: http://library.keldysh.ru/preprint.asp?id=2013-27.

[7] Y. N. Orlov, K. P. Osminin *Methods of statistical analysis of literary texts* URSS/Knizhnyy Dom "LIBROKOM": 2012

[8] V. Keselj, F. Peng, N. Cercone, C. Thomas *N-gram-based author profiles for authorship attribution*, In Proceedings of the Pacific Association for Computational Linguistics, pp. 255-264, 2003

[9] E. Stamatatos, *Authorship attribution based on feature set subspacing ensembles*, International Journal on Artificial Intelligence Tools, 15(5), pp. 823-838, 2006

[10] William B. Cavnar, John M. Trenkle *N-gram-Based Text Categorization*, In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval

[11] 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, 2006

[12] E. Stamatatos *Author identification using imbalanced and limited training texts*, In Proceedings of the 4th International Workshop on Text-based Information Retrieval, pp. 237-241, 2007

[13] *The UC Irvine Machine Learning Repository. Reuter_50_50 Data Set*, Web: https://archive.ics.uci.edu/ml/datasets/Reuter_50_50.

[14] N. Potha, E. Stamatatos *A profile-based method for authorship verification*, In Proceedings of the 8th hellenic conference on artificial intelligence (SETN), LNCS, pp. 313-326, 2014

[15] G. Sidorov, F. Velasquez, E. Stamatatos, A. F. Gelbukh and L.Chanona-Hernndez *Syntactic N-grams as Machine Learning Features for Natural Language Processing*, Expert Systems with Applications, 41(3), pp. 853-860, 2014

[16] G. Sidorov, F. Velasquez, E. Stamatatos, A. F. Gelbukh and L.Chanona-Hernndez *Syntactic Dependency-Based n-grams: More Evidence of Usefulness in Classification*, In Proc. of the 14th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2013), Springer LNCS, 7817, pp. 13-24, 2013

[17] G. Frantzeskou, E. Stamatatos, S. Gritzalis, S. Katsikas *Effective identification of source code authors using byte-level information*, In Proceedings of the 28th International Conference on Software Engineering, pp. 893-896, 2006

[18] *PAN Competitiona* Web: http://pan.webis.de/clef12/pan12-web/author-identification.html

[19] P. Juola *An Overview of the Traditional Authorship Attribution Subtask*, In Proc. of CLEF12, 2012

[20] Web: http://www.rusf.ru/books/analysis/list.htm

[21] Web: http://www.rusf.ru/books/analysis/

[22] D. Shalymov, O. Granichin, L. Klebanov, Z. Volkovich *Literary writing style recognition via a minimal spanning tree-based approach*, Expert Systems With Applications, 61, pp. 145153, 2016

[23] J. Diederich, J. Kindermann, E. Leopold *Authorship Attribution with Support Vector Machines*, Applied Intelligence, 19, pp. 109123, 2003