Using a Decision Tree to Identify Non-uniform Fragments in a Text *

Aleksandr Rogov$^{1[0000-0002-8815-7920]}$, Kirill Kulakov$^{1[0000-0002-0305-419X]}$, Nikolai Moskin$^{1[0000-0001-5556-5349]}$, and Roman Abramov$^{2[0000-0001-9599-4906]}$

[1] Petrozavodsk State University, Petrozavodsk, Russia
rogov@petrsu.ru, kulakov@cs.karelia.ru, moskin@petrsu.ru
https://petrsu.ru/
[2] ITMO University, Saint Petersburg, Russia
monset008@gmail.com
https://itmo.ru/

Abstract. This article discusses the problem of searching for non-uniform text fragments. Each text fragment consists of several paragraphs or separate sentences which significantly differ from the rest of the text in terms of a set of characteristics. The problem of finding non-uniform fragments and their interpretation arises in the study of the pre-revolutionary magazines "Time" (1861-1863), "Epoch" (1864-1865) and the weekly "Citizen" (1873-1874). It's a known fact that F. M. Dostoevsky was their editor. It means that he could have made his own edits to the texts of articles written by other authors. In our research the texts were divided into separate parts. For every part the frequency of n-grams (encoded sequences of parts of speech) was determined. Further, the analysis was carried out using decision trees that classified texts by author. In particular, the texts of F. M. Dostoevsky and V. P. Meshchersky were subjected to this analysis.

Keywords: Text attribution · non-uniform fragment · n-gram · F. M. Dostoevsky · V. P. Meshchersky · decision tree · software complex "SMALT".

1 Introduction

A fragment of text that consists of several paragraphs or even separate sentences will be called non-uniform if it differs significantly from the rest of the text in terms of a set of characteristics. Each characteristic should be poorly controlled by the author of the work and be able to statistically separate two or more authors. For example, for this purpose (to distinguish the text of F. M. Dostoevsky, A. Grigoriev, V. Dahl), the frequency of different sequences of parts of speech found in fragments was used [11]. Based on the calculation of chi-square statistics for the selected fragments and the remaining parts of speech, the question of the uniformity of the fragments was solved.

* Supported by the Russian Foundation for Basic Research, project no. 18-012-90026.

At the same time, it is important to select the most informative features that allow you to identify non-uniform fragments [6]. For example, in [10], experiments were carried out using the FCBF (Fast Correlation-Based Filter), which was proposed by Lei Yu and Huan Liu [16]. The method does not target a specific machine learning model. It does not use classical correlations (for example, Pearson's), but it is based on information theory. As a result of its application, a subset of characteristics is formed by searching and sequential exclusion of uninformative features. Note that to solve the plagiarism search problem, the authors combine FCBF with such mathematical methods as the support vector machine (SVM) and the cumulative sum method (QSUM) [9, 10].

Based on the corpus of prosaic texts by Russian writers of the XVIII-XX centuries (215 texts by 50 authors) and scientific articles on philology, history, law, economics and other social and humanitarian sciences (500 texts written without co-authorship), the authors conclude that the method is quite accurate. Units of the symbolic level of the text, elements of grammar, idiosyncratic and special features of the text were used as characteristics to be compared, including [10]:

– features suggested by Morton: sentence length (in words) and a combination of words starting with a vowel letter and short words of two to four letters;
– sets of bigrams and trigrams of symbols, separated by frequency;
– sets of words and word combinations, divided by frequency;
– grammatical classes of words and combinations of grammatical classes;
– dictionaries of relevant scientific disciplines;
– dictionaries of male and female characters of the text, etc.

A close task is the identification of artificial texts [4]. It can be written by any author, group of authors or be the product of a software algorithm. To compare artificial and natural texts (written by a person) the following numerical characteristics were used: the number of sentences in the text, the number of service words, the average word length, the mention of certain words, the number of short words, the number of long words [12]. In [13], an invariant of artificial texts is proposed, which is a set of values of text characteristics, which allows us to classify texts according to the method of their creation. A method is also proposed for determining artificial texts based on the calculation of the measure of the input text belonging to the invariants (using the Mahalanobis distance), which makes it possible to reach a decision about the origin of the text. We also note the following works on this topic [1, 3].

A significant disadvantage of the considered algorithms is the poor interpretability of the results, which is very important when solving the problem of attribution. In addition, modern neural classifiers require a large amount of training sample for their construction. Decision trees and forests differ better in this case.

The work consists of four sections and a conclusion. The introduction describes the problem of detecting non-uniform fragments in the text and the existing solutions. The second section shows how, based on the frequency of occurrence of certain n-grams, decision trees are constructed for the problem of attribution of texts from the magazines "Time" (1861-1863), "Epoch" (1864-1865) and the weekly "Citizen" (1873-1874). The third part provides a mathematical model used for attribution of a text fragment. The fourth section describes the developed tools for visualizing research results (highlighting text fragments, highlighting n-grams, coloring text), implemented in the SMALT information system.

2 Description of the Method Based on the Decision Tree

The problem of finding non-uniform fragments and their interpretation arises in the study of the pre-revolutionary magazines "Time" (1861-1863), "Epoch" (1864-1865) and the weekly "Citizen" (1873-1874). It's a known fact that the famous Russian writer F. M. Dostoevsky was their editor. It means that he could have made his own edits to the texts of articles of other authors. There is a number of works that have been published without the author's name (or under a pseudonym), which also allows specialists in the field of literary studies to formulate attributive hypotheses.

Since any hypothesis need a convincing proof, mathematical methods that complement philological research are becoming more and more popular. Previously, the authors have already used decision tree method and it has shown good results. For example, in [7], it was found that the relative frequency of the bigram "particle-adjective" greater than 6.5 is a distinctive feature of the journalistic style of Apollon Grigoriev, who published his articles in these journals. In [8], the analysis of the strong positions of texts (i.e. fragments located at the beginning or end of the text) using decision trees demonstrates the possibility of stylistic edits that F. M. Dostoevsky made to the texts of the original authors. Using such trees as an example, one can easily show how the set of texts that fall into a certain node is divided into two subsets, one of which contains objects that satisfy a certain rule and the other one does not.

Note that the first work with the use of decision trees appeared in the 60s of the 20th century and since then they are often used by data mining specialists. There are Classification trees and Regression trees [2]. In the first case, the method predicts if an object belongs to a particular class. In the second case, the predicted result is a real number. The problem of obtaining an optimal decision tree is NP-complete, so heuristic algorithms are needed. Currently, there are the following methods for training decision trees: ID3, C4.5 (improved version of ID3), C5.0, CART (and its modifications IndCART, DB-CART), NewId, ITrule, CHAID, CN2, etc.

In order to carry out such an analysis, it is necessary to split the source texts into fragments, while setting the left and right borders. Then grammatical markup is carried out, which takes into account 14 parts of speech (noun, adjective, numeral, pronoun, adverb, category of state, verb, participle, gerund, preposition, conjunction, particle, modal word, interjection) and also allows to mark quotes, foreign words, introductory words, abbreviated words and non-linguistic symbols. Sequences of parts of speech of different lengths can be represented as n-grams (n indicates how many elements should be taken and determines the size of the sequence). There are bigrams (2-grams), trigrams (3-grams), 4-grams, 5-grams, etc. When we talk about unigrams, or in other words n-grams consisting of one element, we mean the words themselves. N-grams are widely used in natural language processing: for example, for prompting the next word in a search string, searching for plagiarism, correcting errors, etc. Let's take a look at a well-known example of determining the authorship of the novel "The Quiet Don". Some critics argued that M. Sholokhov was too young at the time of writing the novel (23 years old), while the novel was very mature. There was a hypothesis that the author was Fyodor Kryukov, who also wrote about the life of cossacks. To test this hypothesis the frequency of the combination of different classes of words at the level of bigrams, trigrams and tetragrams was studied [5]. In another study of "The Quiet Don" informative signs were formed from the frequency of the symbolic 3-grams, formed from the space character and 33 letters of the Russian alphabet [15]. Other examples of using n-grams for text attribution can be found in [14].

The studies carried out by the authors have shown that bigrams are suitable for solving the problems of attribution of texts from the journals "Time" (1861-1863), "Epoch" (1864-1865) and the weekly "Citizen" (1873-1874) [7, 8]. The text is divided into fragments of 1000 words in increments of 100 words. Having chosen the reference texts of F. M. Dostoevsky and V. P. Meshchersky, a decision tree was constructed that separates the texts of these authors.

After that the text No. 160 "Dvoryanin, zhelayushchij byt' krest'yaninom" ("A nobleman who wants to be a peasant") was analyzed, which was published in the magazine "Time" (1861, volume VI, No. 12, pp. 117-123). However, it is difficult for specialists in philology to analyze the tree in this representation. In order to simplify this process, an algorithm for attributing text fragments was developed.

3 The Attribution Model of a Text Fragment

The method of attribution of a text fragment proposed in the report is based on the ensemble of classifiers. Let's select the fragment in the text that needs to be attributed, denote it by x. With each fragment x we will associate a binary variable z, which will take the value 0 if the text belongs to one author and value 1 if it belongs to the other. Let this fragment x is contained in j fragments y_j whose classification is known. Let's denote it by v_j. The general classification of all fragments is denoted by $f(v_1, ..., v_j)$.

As an example, you can use:

$$f(v_1, ..., v_j) = \frac{1}{j} \sum_{k=1}^{j} v_k \tag{1}$$

$$f(v_1, ..., v_j) = \begin{cases} 1, & \text{if } \sum_{k=1}^{j} v_k \geq \frac{j}{2} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

If $f(v_1, ..., v_j)$ is a binary function, then the obvious solution would be $z = f(v_1, ..., v_j)$. If we have not one tree, but a forest of decisions, then the function $f(v_1, ..., v_j)$ may have a different form.

When it is required to take into account a possible rejection of the classification (for example, when j is an even number), then the variable z_i must take three values. To the values 0 and 1, 1/2 must be added, as a rejection of the classification

Then the formula (2) will take the form:

$$f(v_1, ..., v_j) = \begin{cases} 1, & \text{if } \sum_{k=1}^{j} v_k > \frac{j}{2} \\ 1/2, & \text{if } \sum_{k=1}^{j} v_k = \frac{j}{2} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Note that this technique can be generalized to the case of evaluating the degree of belonging of a given fragment to the author (high, medium, low and excluded) or evaluating the probability of belonging.

Let's consider an example of the proposed algorithm based on the constructed decision tree. In our case, the size of the fragment x is equal to the split step. As a criterion we used the formula (3). The results of the algorithm are presented in Table 1. Here, the first author is F. M. Dostoevsky and the second author is V. P. Meshchersky. The text in question is marked with a number 160 "Dvoryanin, zhelayushchij byt' krest'yaninom" ("A nobleman who wants to be a peasant"). The unit of measurement is the word. Based on the source data, the SMALT system builds a text that is colored in accordance with the identity of the author (this is described in more detail in the next section).

The authors applied this approach when a transformer model was used as a classifier. The results are presented on the resource http://smalt.karelia.ru/. A significant limitation of the use of the classifier built on the transformer model is the need for a large volume of the training sample.

Due to the absence of marked texts with foreign fragments, it is not possible to assess the accuracy of the proposed method. However, philological experts rated it highly. Their expert assessment of some fragments coincided with the assessment obtained using the method described in this report.

4 Software Support for Text Analysis Methods and Algorithms

A large number of routine operations is required to identify non-uniform fragments: marking up texts, calculating statistics and analyzing the results. Usually such tasks are performed by a team of specialists, where the problem of interaction and exchange of the obtained results arises. The information system "Statistical methods of literary text analysis" (SMALT) allows you to speed up these operations at the expense of computer technologies. SMALT has a modular structure [7], which allows automatic text markup, markup correction by specialists and calculation of statistical characteristics of a separate text for research. The

SMALT system is available at http://smalt.karelia.ru/shower. The algorithms and methods presented in this article are focused on obtaining characteristics of groups of texts. Thus, the researcher performs the following process:

- obtaining marked up texts from SMALT;
- running the required algorithms on groups of texts;
- analysis of the obtained results.

Table 1. Marking the authorship of the text.

Fragment number	Left border	Right border	1 author	2 author	$f(v_1, ..., v_j)$
0	0	100	1	0	1
1	100	200	2	0	1
2	200	300	3	0	1
3	300	400	4	0	1
4	400	500	5	0	1
5	500	600	5	1	1
6	600	700	5	2	1
7	700	800	5	3	1
8	800	900	5	4	1
9	900	1000	5	5	1/2
10	1000	1100	5	6	0
11	1100	1200	4	7	0
12	1200	1300	3	7	0
13	1300	1400	2	7	0
14	1400	1500	1	7	0
15	1500	1600	0	7	0
16	1600	1700	0	6	0
17	1700	1800	0	5	0
18	1800	1900	0	4	0
19	1900	2000	0	3	0
20	2000	2100	0	2	0
21	2100	2200	0	1	0

Information system SMALT allows you to get the marked-up text in the form of a table in the format xls (Microsoft Excel), csv or ods (Libroffice Calc). The table contains three columns:

- source word;
- initial form;
- part of speech code.

Sentences are separated by an empty line, paragraphs are separated by two empty lines. If there is loaded text with punctuation marks, a punctuation mark is displayed instead of an empty line in the first column. The results of text processing are submitted in JSON format as an array of objects containing the following fields:

- *id*: the sequence number of the fragment;
- *left_border*: the number of the first word (numbering from zero);
- *right_border*: the number of the last word;
- *pros*: assessment of the degree of belonging to the first author;
- *cons*: assessment of the degree of belonging to the second author.

To visualize the obtained results, SMALT provides the following tools:

- selection of a fragment of text;
- highlighting n-gram;
- coloring text.

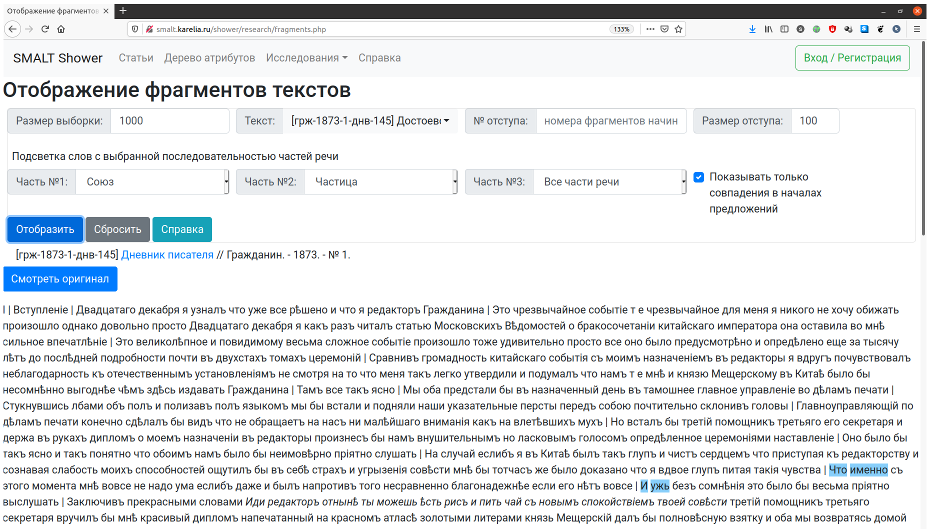
Displaying fragments of text allows you to select the required fragment in a text work. To view the text fragments go to the view form through the "Research" → "Text Fragments" menu. The form consists of fields for defining the selection parameters (fragment size, fragment number, indent size) and a block for displaying the selection results. To select a fragment its size (sample size) and also its offset relative to the beginning of the text are used. The offset is calculated based on the indent number and its size (for example 3 indent of size 10 gives an offset of 30 words). The indent number is the fragment number.

The system allows you to specify several fragments using the generally accepted notation of enumeration and range. For example, for a fragment size of 15 and an indent size of 10, specifying fragment numbers in the form "1-3,5,7" results in the selection of the following word ranges: from 10 to 45, from 50 to 65, and from 70 to 85. The highlighting of the n-gram is done in the same form. To do this in the fields "Part No. 1", "Part No. 2" and "Part No. 3" you need to select the required part of speech (see Fig. 1). You can also show only matches at the beginning of sentences by specifying the appropriate flag. You can search for unigram, bigram and trigram.

Coloring text allows you to visually highlight text fragments. The selection is performed in accordance with the formula (3). There are three options for coloring:

- yellow: $f(v_1, ..., v_j) = 1$;
- green: $f(v_1, ..., v_j) = 1/2$;
- white: $f(v_1, ..., v_j) = 0$.

Insertions of quotations in the text do not participate in the coloring of the text, they are marked in italics. Text coloring pages are stored in the information system database. To download the coloring book you must specify the text, the description of the experiment and the json-file with the distribution of votes by fragments in the appropriate form.



Отображение фрагментов текстов

Размер выборки: 1000
Текст: [грж-1873-1-днв-145] Достоев
№ отступа: номера фрагментов начин
Размер отступа: 100

Подсветка слов с выбранной последовательностью частей речи

Часть №1: Союз
Часть №2: Частица
Часть №3: Все части речи
☑ Показывать только совпадения в началах предложений

[грж-1873-1-днв-145] Дневник писателя // Гражданин. - 1873. - № 1.

I | Вступленіе | Двадцатаго декабря я узналъ что уже все рѢшено и что я редакторъ Гражданина | Это чрезвычайное событіе т е чрезвычайное для меня я никого не хочу обижать произошло однако довольно просто Двадцатаго декабря я какъ разъ читалъ статью Московскихъ ВѢдомостей о бракосочетаніи китайскаго императора она оставила во мнѢ сильное впечатлѢніе | Это великолѢпное и повидимому весьма сложное событіе произошло тоже удивительно просто все оно было предусмотрѢно и опредѢлено еще за тысячу лѢтъ до послѢдней подробности почти въ двухстахъ томахъ церемоній | Сравнивъ громадность китайскаго событія съ моимъ назначеніемъ въ редакторы я вдругъ почувствовалъ неблагодарность къ отечественнымъ установленіямъ не смотря на то что меня такъ легко утвердили и подумалъ что намъ т е мнѢ и князю Мещерскому въ КитаѢ было бы несомнѢнно выгоднѢе чѢмъ здѢсь издавать Гражданина | Тамъ все такъ ясно | Мы оба предстали бы въ назначенный день въ тамошнее главное управленіе во дѢламъ печати | Стукнувшись лбами объ полъ и полизавъ полъ языкомъ мы бы встали и подняли наши указательные персты передъ собою почтительно склонивъ головы | Главноуправляющій по дѢламъ печати конечно сдѢлалъ бы видъ что не обращаетъ на насъ ни малѢйшаго вниманія какъ на влетѢвшихъ мухъ | Но всталъ бы третій помощникъ третьяго его секретаря и держа въ рукахъ дипломъ о моемъ назначеніи въ редакторы произнесъ бы намъ внушительнымъ но ласковымъ голосомъ опредѢленное церемоніями наставленіе | Оно было бы такъ ясно и такъ понятно что обоимъ намъ было бы неимовѢрно пріятно слушать | На случай еслибъ я въ КитаѢ былъ такъ глупъ и чистъ сердцемъ что приступая къ редакторству и сознавая слабость моихъ способностей ощутилъ бы въ себѢ страхъ и угрызенія совѢсти мнѢ бы тотчасъ же было доказано что я вдвое глупъ питая такія чувства | Что именно съ этого момента мнѢ вовсе не надо ума еслибъ даже и былъ напротивъ того несравненно благонадежнѢе если его нѢтъ вовсе | И ужь безъ сомнѢнія это было бы весьма пріятно выслушать | Заключивъ прекрасными словами *Иди редакторъ отнынѢ ты можешь Ѣсть рисъ и пить чай съ новымъ спокойствіемъ твоей совѢсти* третій помощникъ третьяго секретаря вручилъ бы мнѢ красивый дипломъ напечатанный на красномъ атласѢ золотыми литерами князь Мещерскій далъ бы полновѢсную взятку и оба мы возвратясь домой

Fig. 1. Displaying bigrams in the SMALT system.

5 Conclusion

This article depicts the problem of searching and interpreting non-uniform text fragments based on the material of the pre-revolutionary journals ”Time” (1861-1863), ”Epoch” (1864-1865) and the weekly ”Citizen” (1873-1874). The model of attribution of text fragments based on a heuristic algorithm using decision trees has been developed. The used sings were the frequency of the occurrence of certain n-grams (encoded sequences of parts of the speech). To analyze the obtained results in the SMALT information system (http://smalt.karelia.ru/), tools for highlighting text fragments, highlighting n-grams and coloring text have been implemented.

Acknowledgements. This work was supported by the Russian Foundation for Basic Research, project no. 18-012-90026.

References

1. Bakhteev, O. Yu., Kuznetsova, M. V., Romanov, A.V., Chekhov, Yu. V. About one method of detecting artificial and non-scientific texts in an extensive collection of documents. Electronic libraries **20**(5), 298–304 (2017)
2. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J.: Classification and regression trees. Wadsworth, Belmont, Ca (1984)
3. Grechnikov, E. A., Gusev, G. G., Kustarev, A. A., Raigorodsky, A. M.: Search for unnatural texts. Russian Conference on Digital Libraries, Petrozavodsk, 306–308 (2009)

4. Iskhakova, A. O. Method and software tool for determining artificially created texts. Tomsk (2016).
5. Kjetsaa, G., Gustavsson, S., Beckman, B., Gil, S.: Who wrote "The Quiet Don"? Moscow (1989)
6. Kolesnikova, S. I.: Methods of analyzing the informativeness of different types of attributes Tomsk State University Journal of Control and Computer Science **1**(6), 69–80 (2009)
7. Rogov, A. A., Abramov, R. V., Lebedev, A. A., Kulakov, K. A., Moskin, N. D.: Text Attribution in Case of Sampling Imbalance by the Method of Constructing an Ensemble of Classifiers Based on Decision Trees. Data Analytics and Management in Data Intensive Domains. Supplementary Proceedings of the 22th International Conference DAMDID/RCDL'2020 (October 13-16, 2020, Voronezh, Russia). CEUR Workshop Proceedings, 319–328 (2020)
8. Rogov, A. A., Lebedev, A. A., Abramov, R. V., Moskin, N. D., Kulakov, K. A.: Application of decision trees for analyzing the strong positions of the text in the problem of attribution of works by F. M. Dostoevsky. Computer Linguistics and Computing Ontologies. Vol. 4 (Proceedings of the XXIII International Joint Scientific Conference "Internet and Modern Society", IMS-2020, St. Petersburg, June 17-20, 2020). St. Petersburg: ITMO University, 118–127 (2020) https://doi.org/10.17586/0000-0000-2020-4-118-127
9. Romanov, A. S.: Modification of the method of accumulative sums for checking the uniformity of the text and detecting plagiarism // Materials of reports of the International scientific-practical conference "Electronic means and control systems" **2**, 30–38 (2013)
10. Romanov, A. S., Meshcheryakov, R. V., Rezanova, Z. I.: Plagiarism detection and text homogeneity checking technique based on one-class support machine and fast correlation-based filter. Proceedings of TUSUR University **2**(32), 264–269 (2014)
11. Sedov, A. V., Rogov, A. A.: Program system to detect heterogeneity in texts. Proceedings of the VI International Scientific and Practical Conference "Information environment of the university of the XXI century". Petrozavodsk, 135–139 (2012)
12. Shumskaya, A. O.: Choice of parameters for identification of artificial texts. Proceedings of TUSUR University **2**(28), 126–128 (2013)
13. Shumskaya, A. O.: Method of the artificial texts identification based on the calculation of the belonging measure to the invariants. SPIIRAS Proceedings **6**(49), 104–121 (2016) https://doi.org/10.15622/sp.49.6
14. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology **60**(3), 538–556 (2009) https://doi.org/10.1002/asi.21001
15. Usmanov, Z. D., Kosimov, A. A.: About metrization of works of fiction. New information technologies in automated systems **21**, 183–186 (2018)
16. Yu, L., Liu, H.: Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03). Washington DC, 856–863 (2003)