Voronezh State University
Federal Research Center
“Computer Science and Control”
of the Russian Academy of Sciences
ACM SIGMOD Chapter

Data Analytics and Management in Data Intensive Domains

Extended Abstracts of the XXII International Conference DAMDID / RCDL'2020

October 13–16, 2020

Voronezh, Russia

Edited by Bernhard Thalheim, Sergey Makhortov,

Alexander Sychev

Voronezh
Publisher “Research publications”
2020

УДК 004.6+004.89
ББК 32.973+32.973.342
Д17

Д17 Data Analytics and Management in Data Intensive Domains: XXII International Conference DAMDID/RCDL' 2020 (October 13–16, 2020, Voronezh, Russia): Extended Abstracts of the Conference. Edited by Bernhard Thalheim, Sergey Makhortov, Alexander Sychev. – Voronezh : Voronezh State University, 2020. – 241 p.

ISBN 978-5-6045486-0-8

The “Data Analytics and Management in Data Intensive Domains” conference (DAMDID) is held as a multidisciplinary forum of researchers and practitioners from various domains of science and research, promoting cooperation and exchange of ideas in the area of data analysis and management in domains driven by data-intensive research. Approaches to data analysis and management being developed in specific data-intensive domains (DID) of X-informatics (such as X = astro, bio, chemo, geo, med, neuro, physics, chemistry, material science etc.), social sciences, as well as in various branches of informatics, industry, new technologies, finance and business contribute to the conference content. DAMDID conference was formed in 2015 as a result of transformation of the RCDL conference (“Digital libraries: advanced methods and technologies, digital collections”, http://rcdl.ru) so that the continuity with RCDL has been preserved after many years of its successful work.

ISBN 978-5-6045486-0-8
9 785604 548608

УДК 004.6+004.89
ББК 32.973+32.973.342

Contents

Data Management in Semantic Web

Advanced Data Analysis Methods

Text Attribution in Case of Sampling Imbalance by the Method of Constructing an Ensemble of Classifiers Based on Decision Trees *

Alexander Rogov[1], Roman Abramov[1], Alexander Lebedev[1][0000−0001−9939−9389], Kirill Kulakov[1][0000−0002−0305−419X], and Nikolai Moskin[1][0000−0001−5556−5349]

Petrozavodsk State University, Petrozavodsk, Russia
rogov@petrsu.ru, monset008@gmail.com, perevodchik88@yandex.ru, kulakov@cs.karelia.ru, moskin@petrsu.ru
https://petrsu.ru/

Abstract. When solving the attribution problem, the question of determining the author's style of a writer who created a smaller number of texts (both quantitatively and in terms of the total number of words) in comparison with other analyzed authors arises. In this paper we consider possible solutions to this problem by the example of determining the style of Apollon Grigoriev. As a method for constructing an ensemble of classifiers we use *Bagging* (*Bootstrap aggregating*). The SMALT information system ("Statistical methods for analyzing literary texts") was used to determine the frequency characteristics of the texts and Python 3.6 was used to build decision trees. As a result of calculations we can assume that the relative frequency of the "particle-adjective" bigram more than 6.5 is a distinctive feature of the journalistic style of Apollon Grigoriev. There also was a study of the article "Poems by A. S. Khomyakov", which confirms the previously conclusion that there is no reason to consider it as belonging to Apollon Grigoriev.

Keywords: Text attribution · F. M. Dostoevsky · Apollon Grigoriev · Poems by A. S. Khomyakov · sampling imbalance · decision tree · software complex "SMALT".

1 Introduction

Authorship identification of anonymous texts (attribution of texts) is one of most urgent problem for the philological community [3]. One of the issues, which is far from its final decision, is the affiliation of anonymous articles published in the magazines "Time" and "Epoch" (1861-1865) [2]. The solution to this problem is additionally hampered by the uneven amount of available textual material: there are many articles owned by F. M. Dostoevsky, while the remaining authors published in these journals (for example, A. Grigoriev, N. N. Strakhov, Ya. P. Polonsky, etc.), don't have so many texts that are uniquely attributed to them.

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

Different mathematical methods are used to establish authorship of works. Among them decision trees are distinguished by the fact that they are easy to understand and interpret and also do not require special preliminary data processing. When solving the problem of classification into two classes, the problem of sampling imbalance often arises, i.e. when the number of objects of one class significantly exceeds the number of objects of another class. In this case the first class is called the majority class and the second class is called the minority class. In such samplings classifiers are configured for objects of the majority class, i.e. high accuracy of the classifier can be obtained without selecting objects of the minority class. When solving the attribution problem, the question of determining the author's style of a writer who created a smaller number of texts (both quantitatively and in terms of the total number of words) in comparison with other analyzed authors arises.

2 Construction and analyzing decision trees

As a method for constructing an ensemble of classifiers we use *Bagging* (*Bootstrap aggregating*) [1]. The authors believe that it meets the meaning of the task better than *Boosting*. Based on previous research, the fragment size was chosen to be 1000 words and the step size for choosing the beginning of the next fragment to be 100 words. The SMALT information system was used to determine the frequency characteristics [3]. Specialists in philology carried out grammatical markup of texts, which took into account 14 parts of speech. A set of data for training was compiled (118 fragments – Apollon Grigoriev, 899 – the rest). In this case fragments of the texts of Apollon Grigoriev are objects of the minority class and all the others are from the majority class. The text size is quite small (from 2000 to 7000 words).

Python 3.6 was used to build decision trees (libraries: *scikit-learn* – for tree implementation, *pandas* – for data reading). The original data set was divided into 7 parts. All fragments of Apollon Grigoriev were taken as a class with a label "1", the same number of fragments of other authors were taken randomly as a class with a label "0". Repetitions of fragments of other authors were not allowed. A decision tree was trained on each part of data. The training continued until accuracy reached 100% (tree depth). All trees formed an ensemble. The decision was accepted by a majority vote. *Accuracy* was calculated on the entire data as $(TP+TN)/(TP+TN+FP+FN)$, where TP is true-positive, TN is true-negative, FP is false-positive and FN is false-negative predicted class. As a result of experiments depth 1 corresponds to the classifier accuracy of 0,8628 (respectively 2 – 0,9592, 3 – 0,9841, 4 – 0,9891, 5 – 0,992, 6 – 0,9901).

In total 7 decision trees were built. Note that on the third level there are two leaves that contain a small number of fragments (summary from 12 to 27, on average less than 8%). You should take into account the possible inaccuracy of the source data. The texts of Apollon Grigoriev could be edited by F. M. Dostoevsky. In addition there is a slight volatility in the parameters of the author's style depending on external factors (such as mood, health status, etc). There-

fore, when solving the problem of text attribution, you should limit yourself to the first level or at most the first two levels of decision trees. The accuracy of the ensemble at the second level already falls into the generally accepted 5% significance level. Analyzing the decision trees contained in the ensemble, it can be noted that in 4 of them the first attribute was the "particle-adjective" bigram less than or equal to 6.5. In two cases the same attribute is found, but with a different threshold (less than or equal to 7.5). Only one tree had a different first attribute ("adjective-particle") less than or equal to 2.5.

The influence of the universally accepted methods for processing unbalanced data "UpSampling", "UnderSampling", "SMOTE" on the accuracy of classification of works by Apollon Grigoriev was analyzed. The available data set was divided into test (42 - Apollon Grigoriev, 310 - Other) and training samples. The training sample was subjected to the techniques listed above to confront class imbalance. Then the accuracy ("Accuracy", "roc-auc" curve) was calculated on a test sample, which was the same for all three techniques. This analysis showed approximately the same accuracy of all three methods. UpSampling looks worse. The advantage of UnderSampling is that it is easier to explain. Therefore, the authors decided to focus on it.

One of the controversial and still unresolved issues is the article "Poems by A. S. Khomyakov". This work has long been attributed to Apollon Grigoriev. However, recently it has been considered the copyright text of F. M. Dostoevsky [4]. It was interesting to check where our classifier will take it. The text will be attributed to the author that most of the text fragments belong to. If we take the classification on the first node, then 6 of the 7 decision trees classify it as "Other", i.e. as not the text of Apollon Grigoriev. Only on one tree, there was an equality (5 fragments "for belonging" and 5 "against"). During the split on the second level 3 "for belonging", 3 "against" and in one rejection of the classification. Our study confirms the earlier conclusion [4] that there is no reason to consider the article "Poems by A. S. Khomyakov" as belonging to Apollon Grigoriev.

3 Information system SMALT

The SMALT information system developed at Petrozavodsk State University is designed for the collective work of various specialists with texts [3]. The information system can be divided into three sections: import of new texts, verification of texts by philologists and the use of various analysis methods both on a single text and for a group of texts. As part of the text import process, the text is divided into sections, paragraphs, sentences and words, as well as matching each word with its morphological analysis. If the task of text separation is typical, then the task of comparing the morphological analysis is rather complicated. The problem is both in the wide variety of spelling of the word (using pre-revolutionary graphics, a more flexible dictionary allowing different spelling of the word), and in the need to take into account the context of the use of the word. At different times, algorithms for finding the first possible variant, a frequently used variant

and an algorithm based on n-grams were used to select the semantic analysis of the word. The latter has a great prospect due to the small number of subsequent corrections.

As part of the text verification process, philologists perform correction of text analysis (for example, combining or separating words), correction of morphological analysis of a word, or creation of a new analysis. Using the web interface allows several specialists to work on the text at the same time. During the analysis process, SMALT provides researchers with access to the accumulated database in various sections. For example, one of the popular statistical characteristics is Kjetsaa metrics [2]. SMALT calculates the characteristics of both a single work and a group of texts. Another objective of the analysis is to identify the causes of the results. For example, to identify the reasons for the separation of text fragments between different nodes of the decision tree. The SMALT information system allows you to access the source data of the required fragment for subsequent linguistic analysis.

4 Conclusion

When solving the problem of determining the author's style of Apollon Grigoriev, the problem of sampling imbalance often arises. Analyzing decision trees, we can assume that the relative frequency of the "particle-adjective" bigram more than 6.5 is a distinctive feature of the journalistic style of Apollon Grigoriev. The obtained knowledge was used to study the authorship of the article "Poems by A. S. Khomyakov", a discussion about whose authorship in the literary criticism continues over the past twenty years. If we take the classification on the first node, then 6 of the 7 decision trees classify it as "Other", i.e. as not the text of Apollon Grigoriev.

5 Acknowledgements

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

References

1. Bühlmann, P.: Bagging, Boosting and Ensemble Methods. In: Gentle J., Härdle W., Mori Y. (eds) Handbook of Computational Statistics. Springer Handbooks of Computational Statistics. Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-21551-3_33
2. Kjetsaa, G.: Attributed to Dostoevsky: The Problem of attributing to Dostoevsky anonymous articles in Time and Epoch. Oslo: Solum Forlag A. S. (1986)
3. Rogov, A., Kulakov, K., Moskin, N.: Software support in solving the problem of text attribution. Software engineering **10**(5), 234–240 (2019) https://doi.org/10.17587/prin.10.234-240
4. Zakharov, V.: Question about Khomyakov. In: Zakharov, V. The name of the author is Dostoevsky. Essay on creativity. Moscow, Indrik, 231–247 (2013)

Научное издание

Data Analytics and Management in Data Intensive Domains Extended Abstracts of the XXII International Conference DAMDID / RCDL'2020

October 13–16, 2020
Voronezh, Russia

Минимальные системные требования:
PC не ниже класса Pentium I, 32 Mb RAM,
свободное место на HDD 16 Mb,
Windows 95/98, Adobe Acrobat Reader,
дисковод CD-ROM 2-х, мышь

Подписано к использованию 12.10.2020
Объем данных 9 Мб. 1 электрон. опт диск (CD-ROM).
Тираж 500 экз. Заказ 200

ООО «ВЭЛБОРН»
Издательство «Научно-исследовательские публикации»
394068, г. Воронеж, Московский пр-т, 98
Тел. +7 (930) 403-54-18
http://www.scirep.ru E-mail: publish@scirep.ru

Изготовлено фирмой «Большой формат»
(ООО «Твой выбор»)
394018, г. Воронеж, ул. Кости Стрелюка, д. 11/13, офис 6
Тел. +7 (473) 238-26-38
http://big-format.ru, E-mail: 382638@mail.ru