АНОТАЦІЯ

Демидович І. М. Визначення авторства природньомовних текстів методами та засобами конструктивно-продукційного моделювання.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки» – Український державний університет науки і технологій, Дніпро, 2023.

Дисертація присвячена дослідженню та розробці різних методів й засобів встановлення авторства природньомовних текстів на основі різних показників, що відображають особливості авторського стилю мовлення.

У дисертаційній роботі отримані нові науково обґрунтовані теоретичні та експериментальні результати, що у сукупності дозволять застосовувати досліджені методи самостійно або у комплексі з іншими для встановлення авторства текстів та пошуку запозичень.

У першому розділі виконано огляд та аналіз існуючих наразі методів та підходів, що допомагають вловити авторський стиль для різних мов світу. Показано, що різні підходи зумовлені складністю задачі та особливостями різних мов. Встановлено, що досконалого 100% результату у питанні встановлення авторства текстів досі не набуто, незважаючи на широкий перелік використаних інструментів та підходів.

Виявлено, що дослідження підходів для роботи саме з україномовними текстам мають невеликий відсоток на відміну від робіт присвячених іншим мовам, що зумовлено складністю нормалізування та вільністю побудови речень.

З'ясовано, що через особливості побудови речень українською мовою, широкі можливості автора щодо надання тексту певної стилістики на вимогу ідеї твору чи призначенні роботи, поширені методи та підходи роботи з іншими мовами не зможуть в достатній мірі відобразити авторський стиль.

У другому розділі представлені досліджені методи та розроблені моделі статистичного аналізу, аналізу складності текстів, рекурентного аналізу конструктивно-продукційного моделювання.

Виконано адаптацію методів для роботи з природньомовними текстами українською мови. Запропоновано метод створення профілю автора та метод роботи з багатьма показниками для найкращого врахування особливостей авторського стилю.

Розроблена модель природньомовного тексту у вигляді множини правил стохастичних граматик та розроблені метод порівняння текстів на основі порівняння цих правил, що дозволяє враховувати синтаксичні та стилістичні особливості тексту автора

Розроблені конструктори для перетворення природньомовного тексту на множину стохастичних правил та подальше порівняння таких множин для встановлення ступеня їх співпадіння.

У третьому розділі приведені результати експериментальних досліджень. Перевірена та підтверджена ефективність кожного з методів та розроблених моделей. Виконано експерименти за допомогою репрезентативних вибірків як художніх творів різних авторів, так технічних текстів різного розміру та складу. Встановлено ступінь ефективності кожного з досліджених методів окремо.

В подальшому методи було об’єднано для отримання кращого результату та врахування різних особливостей авторського стилю. Було розвинуто та експериментально доведено ефективність методів роботи з великою кількістю різних показників для отримання кращого результату.

У четвертому розділі розроблено інструменти для автоматичного аналізу тексту, підрахунку відповідних показників та подальшого порівняння робіт за ними. Та інструменти що на основі розроблених конструкторів автоматично будують множини правил для різних текстів та порівнюють обрані на ступінь схожості.

Ключові слова: багатокритеріальна оптимізація, генетичний алгоритм, рекурентний аналіз, розпізнавання образів, конструктивне моделювання, авторство текстів, стохастичні граматики, формальні мови, природньомовні тексти, атрибуція текстів, українська мова, авторська атрибуція, критерій Стьюдента.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Праці у фахових виданнях затверджених МОН України:

1. Shynkarenko, V. I., and Demidovich, I. M. "Determination of the attributes of authorship of natural texts." *Artificial intelligence* 3 (2018): 27-35.
2. Shynkarenko, V. I., Demidovich, I. M., and Kuropiatnyk, O. S. "A Dual Approach to Establishing the Authority of Technical Natural Language Texts and Their Components." *Science and Transport Progress* 2 (102) (2023): 71-85. doi: 10.15802/stp2023/288958.
3. Shynkarenko, V. I., and Demydovych, I. M. "Methods and software for significant indicators determination of the natural language texts author profile." *Problems in programming* 3 (2023): 22-29. doi: 10.15407/pp2023.03.22
4. Shynkarenko, Viktor, and Demidovich, Inna. "Constructive-synthesizing modeling of natural language texts." *Computer systems and information technologies* 32023, p. 81-91. doi: 10.31891/csit-2023-3-10

Праці включені до міжнародних наукометричних баз (МНБД) Scopus та Web of Science:

5. Shynkarenko, Viktor, and Demidovich, Inna. "Natural Language Texts Authorship Establishing Based on the Sentences Structure." *COLINS*, 2022, p. 328-337.
6. Shynkarenko, Viktor I., et al. " Processing Words Effectiveness Analysis in Solving the Natural Language Texts Authorship Determination Task." *IEEE 16th International Conference on Computer Sciences and Information*

Technologies (CSIT), 2021, p. 48-51. doi: 10.1109/CSIT52700.2021.9648829.

7. Shynkarenko, Viktor I., and Demidovich, Inna. "Authorship Determination of Natural Language Texts by Several Classes of Indicators with Customizable Weights." *5th International Conference on Computational Linguistics and Intelligent Systems (COLINS),* 2021, p. 832-844.

Матеріали міжнародних наукових конференцій:

8. Шинкаренко, В.І, та Демидович, І.М. . “Статистичний та рекурентний аналіз природньомовних. текстів”. *Збірка тез XII Міжнародної науково-практичної конференції “Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті”*, Дніпропетровський нац. ун-т залізничного транспорту ім. акад. В. Лазаряна, 2018, с. 120.
9. Шинкаренко, В.И., та Демидович, И.Н. “Рекуррентный анализ естественно-языковых текстов”. *Збірка тез Всеукраїнської науково-методичної конференції «Проблеми математичного моделювання»*. Дніпропетровський державний технічний ун-т, 2018, с. 40-43.
10. Шинкаренко, В.І, та Демидович, І.М. “Використання генетичного алгоритму для покращення визначення авторства природньомовних текстів”. *Збірка тез XIV Міжнародної науково-практичної конференції “Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті”,* Дніпропетровський нац. ун-т залізничного транспорту ім. акад. В. Лазаряна, 2020, с. 127.
11. Шинкаренко, В.І, та Демидович, І.М. “Показатель структурного сходства естественно языкового литературного текста”. *Збірка тез XV Міжнародної науково-практичної конференції “Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті”,* Дніпропетровський нац. ун-т залізничного транспорту ім. акад. В. Лазаряна, 2021, с. 68.

12. Шинкаренко, В.І, та Демидович І.М. . “Застосування формальних стохастичних граматик при визначенні авторству текстів” *Матеріали Міжнародної науково-технічної конференції Інформаційні технології в металургії та машинобудуванні*. Український державний університет науки і технологій, 2022, с. 293
13. Шинкаренко, В.І, та Демидович, І.М. «Застосування конструктивного моделювання при визначенні авторства текстів» *Матеріали Міжнародної науково-технічної конференції Інформаційні технології в металургії та машинобудуванні*. Український державний університет науки і технологій, 2023, с.394.

ABSTRACT

Demidovych I. M. Methods and tools development for Ukrainian-language texts authorship determining based on constructive-synthesizing modeling.

Thesis submitted for obtaining the Doctor of Philosophy degree in the specialty 122 "Computer Sciences" – Ukrainian State University of Science and Technology, Dnipro, 2023.

The dissertation is devoted to the research and various methods and means development for establishing the natural language texts authorship based on various indicators that reflect the peculiarities of the author's speech style.

New theoretical and experimental scientifically based results were obtained, which together will allow applying the researched methods independently or in combination with others to establish the authorship of texts and search for borrowings.

In the first chapter, a review and analysis of currently existing methods and approaches that help to capture the author's style for different languages of the world is performed. It is shown the variety of different existing approaches due to the complexity of the task and the structure distinction in different languages. It has been established that a perfect 100% result in establishing the texts authorship has not yet been achieved, despite the wide range of tools and approaches used.

It was found that the research of approaches for working specifically with Ukrainian-language texts has a small percentage, in contrast to works devoted to other languages, which is due to the complexity of its formalization and the variety of sentence constructions.

It has been found that due to the complexity of sentence structure in the Ukrainian language, and the wide possibilities for the author to provide the text with a certain style at the request of the main idea or the purpose of the work, commonly used methods and approaches will not be able to sufficiently reflect the author's style.

The second section presents the researched methods and developed models statistical analysis, analysis of text complexity, recurrent analysis, structural and production modeling. Methods adaptation for working with natural language texts in the Ukrainian language has been developed. An author's profile creating and working with the range of indicators, finding the best among them to reflect the author's style crucial features methods are proposed.

A natural language text model in the form of stochastic grammars rules set was developed and the texts comparing method based on the comparison of these rules was developed, which allows working with the syntactic and stylistic features of the author's text.

Constructors have been developed for converting natural language text into a set of stochastic rules and further comparing such sets to establish the degree of their similarity.

The third section presents the results of experimental research. The effectiveness of each method and developed model has been tested and confirmed. Experiments were carried out with the help of representative samples: different authors fictional works and technical texts in different sizes and formats. The effectiveness degree of each investigated method was determined separately.

The methods were combined to obtain a better result and take into account various features of the author's style. The effectiveness of methods working with a large number of different indicators to obtain a better result was developed and experimentally proven.

In the fourth chapter, tools are developed for automatic text analysis, calculation of relevant indicators and further comparison of works based on them. And tools based on developed constructors that automatically build sets of rules for different texts and compare the selected ones for the degree of similarity.

Keywords: multicriteria optimization, genetic algorithm, recurrent analysis, pattern recognition, constructive-synthesizing modeling, authorship of texts, stochastic grammars, formal languages, natural language texts, attribution of texts, Ukrainian language, authorship attribution, Student's criterion.