УДК 004.89

FEATURES OF DESIGNING A SOFTWARE SYSTEM FOR DETERMINING THE FALSEING INFORMATION BASED ON FREQUENCY PATTERN SEARCH ALGORITHMS

Khovrat A.V.
Scientific Supervisor – Cand. Sc. (Techn.), assoc. prof. Kobziev V. G.
Kharkiv National University of Radio Electronics, dpt. of Software Engineering
Kharkiv, Ukraine
e-mail: artem.khovrat@nure.ua

Робота присвячена процесу побудови алгоритму обробки навчальних робіт з метою визначення їх фальсифікації шляхом використання генераційних нейронних мереж, зокрема Chat GPT. У якості основного алгоритму обрано модель пошуку частотних патернів Apriori. Задля подолання проблеми швидкодії запропоновано використання технології MapReduce з прискоренням близько 3.1. Отриманий результат точності класифікації робіт та швидкість алгоритму дозволяють стверджувати про доцільність імплементації запропонованого підходу.

The rapid development of artificial intelligence and related technologies leads to the spread of the problem of their unscrupulous use, in particular in the educational domain. According to international studies [1], with the advent of Chat GPT, the volume of falsification of homework and exam answers has increased for both schools and universities. An important feature of these works is the use of phrases that are not typical for the target language, a large number of generalizing words and the absence of a specified vocabulary [2]. Such markers allow verifiers to concluder about the using of generation technologies. However, in this case, the subjectivity of the inspector's perception may be imposed on the evaluation process. In order to avoid possible conflict situations, it is proposed to develop a system that receives a text and returns the probability of its falsification by intelligent data analysis software. To build this system, it is suggested to consider a family of frequency pattern search algorithms and check their accuracy and processing speed for text data in the Ukrainian language.

As an example, it was decided to consider the Apriori algorithm [3]. Selected approach is not as comprehensive as transformers or complex recurrent neural networks, it works much faster and does not require large amounts of data for training [4]. The Apriori model has only four steps. First: finding support for each element. Support is the frequency of occurrence of an element in a data set.

After finding this parameter, it is enough to choose those elements that satisfy a certain pre-set restriction. As a result, all the most frequent patterns will be found, on the basis of which a set of associative rules can be built. The final step is to sort the received values in descending order of elevator (ascent).

In order for the result of the check to be correct, it was decided to add a simple algorithm of refinement that uses operations of stemming, lemmatization

and cleaning from words without a significant lexical load. And although these actions are not typical for polymorphic languages, in particular Ukrainian, when using Apriori, this is not essential. At the same time, one of the significant disadvantages of the proposed approach is the execution time, which will negatively affect the possibility of its implementation in a real time system. To overcome it, it was decided to use MapReduce technology based on Hadoop. Its essence lies in the distribution of processing of a common set of data to individual nodes. This procedure is performed using the mapping function, with subsequent application of the necessary algorithms, and a reducer that collects data from all nodes and unifies them [5].

In the current case, the information read from the database in different blocks. Each of these blocks executes a sequential form of the Apriori algorithm. Next, the reducer combines the values on each block and passes them to the reducer, which in turn filters the data. It is worth noting that the available mapper performs the ordering process necessary for the correct operation of the algorithm automatically. The conducted set of experiments showed that the acceleration could reach 3.1 and meet the generally accepted norms for implementing the algorithm in the proposed system. The relatively small acceleration is explained by the general simplicity of the basic algorithm.

When checking the accuracy of the proposed approach on a self-generated set of test data, a value of about 91% was obtained. This in general indicates the high efficiency of the use of frequency pattern searching algorithms to detect falsification of educational works, however, the question of optimizing the proposed solution and the feasibility of considering other models with in order to obtain a better approach and its implementation in the future system.

References:

1. Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 1–12.

2. Sandler, M., Choung, H., Ross, A., & David, P. (2024). A Linguistic Comparison between Human and ChatGPT-Generated Conversations. Cornell University Archive. https://arxiv.org/abs/2401.16587.

3. Subhani, S., Devarakonda, N., & Nagamani, C. (2018). Parallel Computing Algorithms for Big data frequent pattern mining. In Int. Conference on Intelligence & Data Engineering ICCIDE-2017. Springer Nature Singapore.

4. Generalized Semantic Analysis Algorithm of Natural Language Texts for Various Functional Style Types / N. Sharonova et al. In COLINS-2022. CEUR Workshop Proceedings.

5. Yakovlev, S., Khovrat, A., & Kobziev, V. (2024). Using Parallelized Neural Networks to Detect Falsified Audio Information in Socially Oriented Systems. In IT&I-2023. CEUR Workshop Proceedings.