Journal of Process Management and New Technologies
Vol. 11, Issue 3-4, 2023, pp. 57-65

APPLICATION OF CHATBOT AI IN THE CREATION OF WEB MINING PROGRAMS AND THEIR ANALYSIS

Luka ILIĆ[1*], Aleksandar ŠIJAN[2], Bratislav PREDIĆ[3]

[1] *Faculty of Electronic Engineering, University of Niš, Niš, Serbia, luka.ilic@mef.edu.rs*
[2] *Faculty of Electronic Engineering, University of Niš, Niš, Serbia, aleksandar@mef.edu.rs*
[3] *Faculty of Electronic Engineering, University of Niš, Niš, Serbia, bratislav.predic@elfak.ni.ac.rs*

Abstract: *Artificial intelligence and ChatGPT became very popular in late November 2022. Their popularity has led to the development of other chat AI systems and AI tools. In order to monitor this progress and achieve key goals in the application of artificial intelligence, this paper has the task of presenting an analysis of the generated web mining program. One of the key aspects of this research is the presentation of advantages and disadvantages of this approach, including analysis of accuracy, speed and reliability. Also, potential challenges in the application of artificial intelligence (especially chat systems), such as the limitations of AI models, ethical and security aspects when processing data from the Internet, are explored. This research paper aims to contribute to a better understanding of the capabilities of AI in the development of web mining programs, identify potential areas of improvement and provide guidance for future development.*

Keywords: *artificial intelligence, chat gpt, web mining*

Original scientific paper
Received: 27.09.2023
Accepted: 14.10.2023
Available online: 19.11.2023
DOI: 10.5937/jpmnt11-46801

1. Introduction

Web mining is the process of extracting and analyzing data and information from the Internet in order to obtain useful and relevant data (Rahimi & Danesh, 2023; Gheisari *et al.*, 2023). Web mining tools enable the identification of meaningful information from large sets of texts found on the Internet (Karthigeiyan & Devi, 2023; Lalitha & Sreeja, 2023). It has a wide application, and the most important are market research, search personalization, analysis of user activity on the site, as well as optimization of websites for a better user experience (Samanta *et al.*, 2022; Mangat & Saini, 2022; Siddiqui & Aljahdali, 2013; Ting, 2008; Chen & Chau, 2004).

Artificial intelligence is a term that is often mentioned in the media and it is necessary to explain what it actually means. Artificial intelligence is a concept that includes a large number of concepts from computer science, generally speaking of mechanisms that use stochastic methods, that is, methods of randomness. A subset of artificial intelligence is machine learning, which has gained a lot of popularity among scientists and engineers with a large number of free

* Corresponding author

tools to work with (Helm *et al.,* 2020). The function of machine learning is to enable a computer to perform tasks without additional programming. Another branch of artificial intelligence is deep learning, the main characteristics of which are multi-layer neural networks. Neural networks represent a system consisting of a certain number of interconnected nodes where each node has its own local memory in which it remembers the data it processes (Goodfellow et al., 2016).

The focus of the project will be on the python program generated by AI. This program can offer good scalability ie. extension of the program depending on the functionality that the user will require from the AI. For example, an AI generated python program can show flexibility in different web mining requirements. Also, this program can be extended with other data processing programs (eg Excel), python libraries, etc.

On the other hand, this analysis can show the flaws of the python program generated by the AI, considering that there may be some limitations in the AI model. This can also be a representation of the current state of the AI model. The currently generated python program, created by AI, represents the current state of the web mining technology. As technology advances, there is a possibility that in a few years more advanced programs will be generated that will provide improved performance, more accurate data and more functionality will be available for this and similar projects.

2. Artificial intelligence

Artificial intelligence represents a revolutionary innovation in the history of human civilization. Its development was initially focused on machines and robots, especially space and similar technology (Branković, 2017). With comprehensive digitization, AI has experienced a significant boom, expanding its influence to all spheres of human life. The concept of "smart" entities has come to define various objects and fields, including smart machines, phones, cars, cities, toys, buildings, lighting, classrooms, and whiteboards. Today, AI is a key part of many technological and industrial sectors. Her ability to analyze, interpret and make decisions based on data enables improved performance and efficiency in various areas. This technology has the potential to change the way we live and work, improving processes and bringing innovative solutions.

AI is a field of computer science that deals with the development and implementation of systems that aim to create computer systems that can analyze, learn, and make decisions to solve problems in a similar way to how humans do. AI encompasses various techniques, the most popular being the following:

- Machine learning: A branch of AI that focuses on the development of algorithms and techniques that allow computers to learn from data and improve their performance through experience.
- Deep learning: A part of machine learning that uses neural networks with multiple layers to learn complex patterns from data.
- Natural language processing: It deals with understanding and generating natural language, enabling computers to read, understand and communicate with people in the language the user is using (the best example of this is ChatGPT)

We can see the application of artificial intelligence in all spheres of human life where computers are used: autonomous vehicles, medicine, banking, finance, e-commerce and others. Its development contributes to new innovations and improvements in various areas of human society.

When we talk about artificial intelligence, the first thing that usually comes to mind is machine learning. Machine learning is a key pillar of artificial intelligence that allows

computers to learn from data and improve their performance without the need for explicit programming (Kujundžić, 2015). This technique is very impressive because computers can identify patterns in data and thus apply them to new situations they encounter. This kind of work can lead to solving complex problems and making decisions without the need to re-promote. Some of the definitions of machine learning say that it represents the design of computer algorithms that use past experience when making future decisions, it is actually a model that learns and makes decisions based on certain data. Machine learning algorithms build a mathematical model based on data samples (training data) that aim to perform a task without additional programming..

In the last few years, machine learning occupies a very important place in the world of information technology, especially when it comes to analyzing large amounts of data and answering queries from huge databases. The task of machine learning algorithms is to "learn" the prediction function (that is, train the model) h(x) based on the given training, so that h(x) is an optimal approximation for the corresponding values (almost, but not completely accurate) (Milošević *et al*, 2014; Stojanović *et al*, 2017).

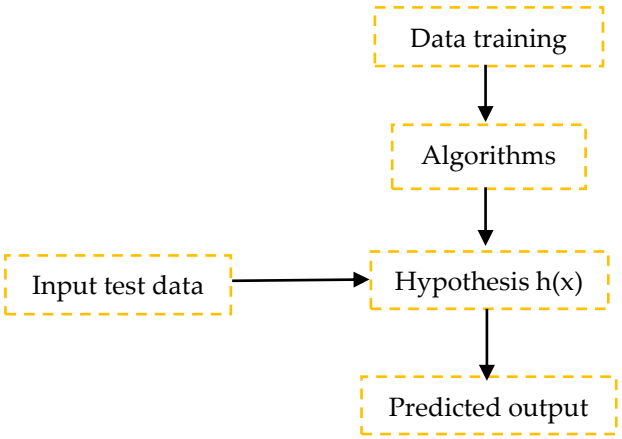

Diagram 1. Machine learning process

3. Chatbot AI

Chatbots are programs that use artificial intelligence (AI) to simulate a conversation with a user through text messages or speech. The main goal of a chatbot program is to provide a user experience similar to a conversation with a real person. Chatbots can be used in various fields, such as customer support, sales, education, entertainment, healthcare and other areas of human daily life..

There are two main types of chatbots:

- Rule-Based Chatbots: They work based on pre-defined rules and templates. They identify keywords in user questions and use pre-determined answers that best match the given words. Such chatbots are less flexible because their abilities depend on pre-programmed information. This type of chatbot is based on if/else logic.
- AI Machine learning chatbots: They use advanced techniques of artificial intelligence (natural language processing - NLP) and machine learning to better understand and answer the questions asked. They rely on a large amount of data to improve their responses over time of use. These AI chatbots have more flexibility in talking with the user and better understand the set requests.

The most popular chatbot is ChatGPT which is based on GPT-3.5 technology developed by OpenAI. It is actually an advanced language model that has a wide range of applications. Through interaction with users, this tool can answer various questions, provide support, generate textual content, solve programming tasks and other things that require natural language processing. ChatGPT has quickly become popular and has the potential to significantly impact various aspects of our lives, including education and research. Thanks to advanced natural language processing (NLP) capabilities, ChatGPT can understand and interpret human language in a way never before possible, allowing users to ask questions and get answers in a conversational and intuitive way (Adamopoulou & Moussiades, 2020; Zhu *et al*, 2023).

4. Web mining

Web mining is based on discovering content from the web. Web mining is like a graph and all pages are nodes and each connects with hyperlinks. This mode is good for getting information, images, texts, documents and multimedia (Dzitac & Moisil, 2017). By using web mining we can easily extract all information about multimedia. We can divide it into three categories (Kareem & Ibrahim, 2021; Sharma *et al*, 2018):

- Web Content Mining: It represents the process of collecting, extracting and analyzing data from Internet content. It is a type of data mining that focuses on extracting useful information from websites, blogs, forums, social networks and other sources on the Internet. The goal of this process is to discover useful patterns, trends, and information that can be used for business decision making, market research, competitor monitoring, service personalization, user behavior analysis, and more..
- Web Structure Mining: It represents the process of analyzing and extracting information from the structure of Internet content, such as links between web pages, hierarchical organization of web pages, and connection graphs. This type of data mining focuses on studying the relationships between different elements on the web in order to discover useful patterns, structures, and information..
- Web Usage Mining: It is the process of analyzing and extracting user behavior on the web to uncover useful patterns, trends and information. This type of data mining is aimed at studying user activities on web pages, such as clicks, navigation, search, interaction with web elements, time spent on pages and the like.

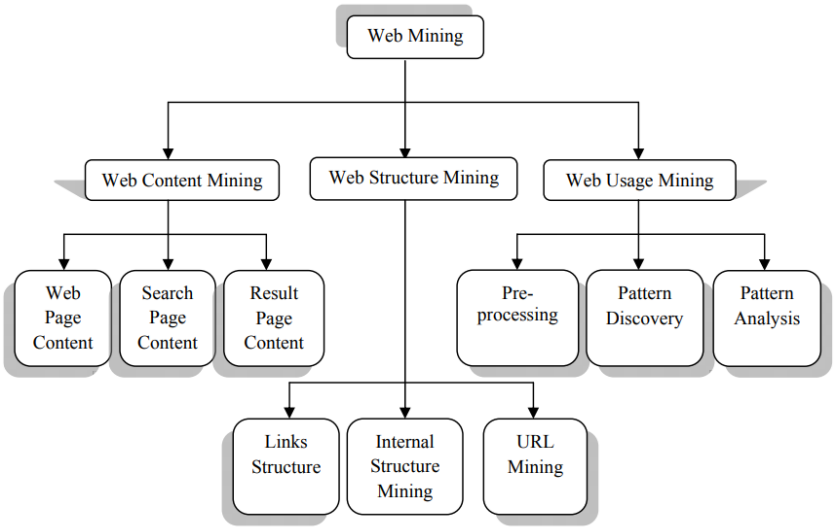

Figure 1. Web mining

4.1. Web mining tools

These are tools used to collect useful data and information from the Internet. They search the Internet, collect data from the sites and analyze them, they can also analyze the behavior of users on the site. The most popular types of tools are web crawlers that search the Internet, tools for extracting text and multimedia from the site, and web analytical tools that serve to analyze and monitor site visitors. When using these tools, it is very important to respect the terms of use of websites and in accordance with their guidelines and legal regulations.

There are many ready-made solutions for work, among them, one of the most popular is WebHarvy. WebHarvy is a commercial web scraping software used to extract data from websites without the need to write code. It allows users to set appropriate parameters for extracting data from sites, such as selectors, filtering and pagination. WebHarvy is a popular web mining tool and can be used for a variety of applications, including data extraction from e-commerce sites, web directories, real estate searches, gathering product information, reviews, and more.

With the recognition of the potential of the vast amounts of data available on the Internet, the field of web mining is becoming increasingly important in various industries. By analyzing this data, valuable insights about consumer habits, market trends and other relevant information can be obtained. Despite the benefits provided by web mining, we are often faced with the challenge of obtaining data from various web pages that are not structured in a way that allows direct analysis. This is exactly where web scraping comes into play.

A web scraping program should be carefully designed, taking into account a number of key elements that enable efficient data collection from web pages. These elements form the basis of successful web scraping, allowing the program to access web content, analyze HTML structures, accurately extract information, and finally process and save it in a useful form. Building such a program requires mastering techniques such as HTTP requests, HTML parsing, data processing and cleaning, iteration through pages, as well as careful processing of potential errors and respect for ethical principles of scraping. This versatile approach allows developers to create sophisticated tools for collecting information from web pages, market research, tracking changes and many other purposes, taking into account not only the technical aspects but also the legality and ethics of the process.

Building such a program requires mastering techniques such as (Mitchell, 2018):

1. HTTP requests: Enable the program to communicate with web servers and retrieve HTML content of pages. Through properly placed HTTP requests, the program gains access to the data on the target pages.
2. HTML parsing: Using HTML parsing libraries (the most popular beautifulsoup4), the program analyzes the structure of HTML, identifies specific elements and their attributes that contain the desired information.
3. Data extraction: Using selectors or methods of the HTML parsing library, the program extracts text, links, images and other resources from identified HTML elements, enabling the collection of targeted information.
4. Processing and cleaning: Collected data often require further processing to be useful. The program should clean the data, removing redundant characters, white space and any unwanted information.
5. Data storage: Extracted data should be stored in a suitable format, such as CSV, JSON or a database, allowing for further analysis and use.

The mentioned technical aspects are the basis of any web scraping program, but it is also important to pay attention to the following:

1. Iterate through pages: If the target information spans multiple pages, the program should support iteration through pagination, ensuring that data is collected from all relevant parts.
2. Error handling: The program should be resistant to errors such as site unavailability or changes in the HTML structure, in order to ensure stable and reliable data collection.
3. Ethics and legality: Before starting to collect information from HTML pages, it is mandatory to check the website's terms of use and compliance with the guidelines from the "robots.txt" file in order to ensure ethical and legal information collection.

Building a web scraping program requires a comprehensive approach, taking into account technical, legal and ethical aspects. This versatile approach enables developers to create tools that efficiently collect, analyze and store web content according to specific needs and guidelines.

5. AI generated program, improvement and optimization

AI generated software represents a revolutionary approach to the software development process, enabling automation and acceleration of program creation. As part of the research, an AI generated program was developed that focuses on web mining, applying advanced techniques for processing and analyzing web content. Compared to the usual ways of creating software, programs generated by artificial intelligence allow us to adapt more quickly and deal more easily with large amounts of data, as often occurs in the process of collecting information from web pages (web mining). When we compared the Bard model with ChatGPT, the research showed that ChatGPT delivers better results in code generation. ChatGPT has a better understanding of the task and the situation, which is why it can generate more accurate and efficient code that solves specific challenges in web content analysis. For the same questions asked, ChatGPT gave more specific answers and it was not necessary to ask additional questions to develop the code as was the case with Bard. Also, ChatGPT demonstrates greater flexibility in adaptive code generation, responding to different input scenarios for HTML parsing and data extraction. Although Bard can be used similarly to ChatGPT, its efficiency and accuracy in questions (requests) is slightly lower.

During the program generation process, the challenge was for the chatbot AI to write a program that would automatically identify a product category and save that category to JSON

and CSV files. When a product had only one category, both models were able to successfully store that category in the database. However, the problem arose when a product had multiple categories. Based on the generated code, the chatbot was asked to create a program that would collect all categories from that product. ChatGPT successfully performed this task, managing to identify and collect all the relevant categories. On the other hand, the Bard model was not able to do this, even after setting up more tasks and explaining the necessary steps in more detail.

During the program generation phase, we encountered challenges due to the limitations of the current model in creating a graphic interface that would allow selecting elements by clicking, and then further processing them through the program. ChatGPT, although successfully generating code for various functionalities, could not intuitively solve this problem. Creating an interactive user experience required consideration of complex interaction steps and appropriate interface design to enable efficient selection of elements and their transfer through the program for further processing. This part of the research points out how important it is that the technical capabilities of the model correspond to the needs of the user experience in order for the web content analysis program to be efficient and functional. We expect that in the future the models will be further improved in order to solve such challenges, which will enable the development of even more complete programs.

People with an understanding of programming will have the opportunity to further improve this program by adding additional modules, such as functionality for user registration, data search, custom interface elements and optimization of the result display format. This opens the door to expand the functionality of the program according to specific needs and requirements, providing greater adaptability and user experience. This ability to add modules allows the program to be adapted to different situations and requirements, thereby further enriching its use and usefulness.

This program will be constantly improved to make it better and more useful. Over time, new features and improvements will be added to facilitate interaction and achieve better results. For example, it is possible to add options for managing user accounts, advanced search and selecting elements for extraction through a graphical interface. Also, the program will become faster and better at analyzing web content and extracting information. Given the rapid development of artificial intelligence, the program will adapt to new techniques in order to remain efficient and useful for analyzing data from web pages. This continuous improvement will contribute to making the program an even stronger and more useful web content analysis and web mining tool.

The very essence of this program lies in the fact that its complete code is generated by an AI chatbot. When problems or difficulties arose during development, the chatbot AI solved them by the user asking those problems. This highlights the role of AI technology in the entire process, where the chatbot played an active role to ensure the functionality and quality of this generated program. Also, this approach has the potential to find application in various fields, not limited to web mining. In addition to the development of web content analysis programs, such models can be used to automate and improve processes in many other areas, such as database management, natural language processing, algorithm optimization, and more. The flexibility and adaptability of these models make them valuable tools for solving various challenges and opening up new possibilities for the application of artificial intelligence.

6. Conclusion

In this paper, we investigated the application of chatbot AI technology in the process of developing a web content analysis program. Our goal was to explore how chatbot AI models, with a special emphasis on ChatGPT, can contribute to code generation and solving web mining challenges.

At the moment we conducted the experiment, it is noticeable that ChatGPT already provides better results and better explanations compared to other models, hinting at greater potential for the future. We are witnessing the rapid development of artificial intelligence, the question arises what awaits us in a few years when the models are improved even more. The aim of this research paper was to show the professional public the advantages and disadvantages of chatbot AI services, especially in the context of the development of web content analysis programs. Analyzing the differences between models such as ChatGPT and others, the paper pointed to significant steps forward in the field of artificial intelligence and provided insight into the potential that this technology carries. This experiment lays the groundwork for further research and development, opening the door for more comprehensive applications of chatbot AI in various fields, including web mining, while future models are expected to achieve even greater innovation and improvement.

This research has shown how chatbot AI technology, especially ChatGPT, can contribute to the efficient development of web content analysis programs. Its current advantages point to a bright future, pointing to the possibilities of advancement and application of artificial intelligence.

References

Adamopoulou E., Moussiades L., (2020). An Overview of Chatbot Technology, Artificial Intelligence Applications and Innovations, 373-383

Branković S., (2017). Veštačka inteligencija i društvo, Srpska politička misao, 5-12

Chen, H., & Chau, M. (2004). Web mining: Machine learning for web applications. Annual review of information science and technology, 38(1), 289-329.

Dzitac I., Moisil I., (2017) Advanced AI techniques for web mining, MAMECTIS'08

Gheisari, M., Hamidpour, H., Liu, Y., Saedi, P., Raza, A., Jalili, A., ... & Amin, R. (2023). Data Mining Techniques for Web Mining: A Survey. In Artificial Intelligence and Applications (Vol. 1, No. 1, pp. 3-10).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., ... & Ramkumar, P. N. (2020). Machine learning and artificial intelligence: definitions, applications, and future directions. Current reviews in musculoskeletal medicine, 13, 69-76.

https://openai.com/ (25.08.2023)

https://www.academia.edu/80557057/AI_in_Web_Mining (25.08.2023)

https://www.geeksforgeeks.org/web-mining/ (25.08.2023)

https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html (25.08.2023)

https://www.sciencedirect.com/topics/computer-science/web-mining (21.08.2023)

https://www.techtarget.com/whatis/definition/ChatGPT (25.08.2023)

Kareem K., Ibrahim J., Ahmed J, (2021). Web Mining Techniques and Technologies: A Landscape View, Journal of Physics: Conference Series, 1-13

Karthigeiyan, K., & Devi, M. D. (2023). Farm Track Application Development Using Web Mining and Web Scraping. Advances in Science and Technology, 124, 566-573.

Kujundžić M., (2015). Analiza modela predviđanja trendova kretanja cijena vrijednosnih papira, Sveučilište u Rijeci, Tehnički fakultet, 2-30

Lalitha, T. B., & Sreeja, P. S. (2023). Potential Web Content Identification and Classification System using NLP and Machine Learning Techniques. International Journal of Engineering Trends and Technology, 71(4), 403-415.

Mangat, P. K., & Saini, K. S. (2022). Relevance of data mining techniques in real life. In System Assurances (pp. 477-502). Academic Press.

Milošević B., Milošević D., Obradović S., (2014). Mašinsko učenje u obrazovanju, Infoteh Jahorina, 1-5

Mitchell R., (2018). Web Scraping with Python, Allite books, 3-103, 215-218

Rahimi, F., & Danesh, F. (2023). Analyzing Persian Wikipedia's citations to discover the effectiveness of Persian scientific papers: applied web mining techniques. Performance Measurement and Metrics, 24(2), 85-100.

Samanta, D., Dutta, S., Galety, M. G., & Pramanik, S. (2022). A novel approach for web mining taxonomy for high-performance computing. In Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021 (pp. 425-432). Springer Singapore.

Shan T., Tay F. R., Gu L., (2020). Application of Artificial Intelligence in Dentistry

Sharma K., Sharivastava G., Kumar V., (2018). Web Mining: Today and Tomorrow, TarjomeFa, 1-5

Siddiqui, A. T., & Aljahdali, S. (2013). Web mining techniques in e-commerce applications. arXiv preprint arXiv:1311.7388.

Stojanović M., Cvetković S., Stančić G., (2017). Poređenje metoda mašinskog učenja za predikciju trenda promene finansijskih vremenskih serija, Infoteh Jahorina, 1-4

Ting, I. H. (2008). Web mining techniques for on-line social networks analysis. In 2008 International Conference on Service Systems and Service Management (pp. 1-5). IEEE.

Zhao B., (2017). Web Scraping, Encyclopedia of Big Data, 1-4

Zhu J. J., Jiang J., Yang M., Ren Z. J., (2023). ChatGPT and Environmental Research, American Chemical Society, 1-4