International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n1.14
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research on the Application of Big Data and Artificial Intelligence in Search Engine

Hanzhe Hu

Northeastern University, Shenyang, Liaoning, 110819, China
13760455897aw@gmail.com

ABSTRACT

With the development of the Internet, search engine as an important tool of information retrieval, more and more need to deal with large-scale, complex data. The combination of big data and artificial intelligence technology provides a new solution for search engine optimization. This paper mainly discusses the application of big data and artificial intelligence in search engine, including data collection and processing, data mining and pattern recognition, user behavior analysis, personalized recommendation algorithm, machine learning algorithm, deep learning algorithm, reinforcement learning algorithm and the intelligence of personalized recommendation system.

KEYWORDS

Big data; Artificial intelligence; Search engines; Applied.

1. INTRODUCTION

In the era of information explosion, search engine as an important tool for people to get information, its performance and user experience are of great importance. The traditional search engine mainly relies on keyword matching and web page sorting algorithm, however, this way in the face of massive data and complex information needs, it seems inadequate. Therefore, how to use big data and artificial intelligence technology to improve the performance of search engines and user experience is an important challenge we are currently facing.

The application of big data technology can help search engines deal with large-scale and complex data. Through distributed storage and computing technology, it can quickly collect and process massive data, including web page content, user behavior data, etc[1]. Through the in-depth mining and analysis of these data, it can be well understood by the needs of users, so as to provide more accurate information services. And the application of artificial intelligence technology in search engines can further improve the intelligence level of search engines. Technologies such as machine learning, deep learning and reinforcement learning can help us automatically learn the characteristics and rules of data from massive amounts of data, so as to predict and analyze unknown data. These technologies can be used not only for the classification, clustering and association rule analysis of web pages, but also for the mining and analysis of user behavior data to optimize the ranking of search results and personalized recommendations[2].

By combining big data and artificial intelligence technologies, more intelligent search engines can be built and more efficient and personalized information retrieval services can be provided. Such intelligent search engines can not only meet users' needs for information acquisition, but also improve the quality and relevance of information, thus enhancing the user experience.

2. APPLICATION OF BIG DATA IN SEARCH ENGINES

2.1. Data collection and processing

2.1.1. Sources and types of data

The data sources of search engines mainly include user data, web page data and log data of search engines themselves. User data includes the user's search history, click behavior, browsing behavior, etc. Web data includes the content of web pages, metadata, links, etc. The log data of search engines includes the keywords that users search for, the pages they click on, the time they search for, etc. This data can be of various types, including text, pictures, videos, etc.

2.1.2. Data cleaning and preprocessing methods

The purpose of data cleaning is to remove duplicate, incomplete, incorrect or abnormal data, thereby improving the quality and accuracy of the data. Methods of data cleaning include removing duplicate data, filling in missing values, deleting invalid data, etc. Preprocessing is one of the important links of data cleaning, which includes the normalization, standardization and normalization of data, so as to improve the readability and processability of data.

2.1.3. Data storage and management strategy

Since the amount of data that search engines need to process is very large, efficient storage and management strategies are needed to ensure the availability and scalability of data. Commonly used storage technologies include distributed file system, database system and so on. In the management strategy, it is necessary to consider data backup, recovery, permission control and other aspects to ensure the security and reliability of data. At the same time, it is also necessary to regularly maintain and update the data to ensure the accuracy and timeliness of the data.

2.2. Data mining and pattern recognition

2.2.1. Data mining technology and application

Data mining is a process of extracting useful information and knowledge from a large amount of data. Commonly used data mining technologies include cluster analysis, association rule mining, decision tree analysis, etc. In search engines, data mining technology can be used for user behavior analysis, webpage classification, keyword association and so on. For example, through cluster analysis, users' search history can be classified, so as to find users' interest preferences and behavior patterns[3]. Association rule mining can find the association relationship between users' search keywords, so as to optimize the search algorithm and recommendation strategy.

2.2.2. Pattern recognition algorithm and method

Pattern recognition is a method to identify specific patterns or regularities by analyzing data. In search engines, pattern recognition algorithms can be used to identify a user's search intent and relevancy in order to accurately rank search results. Commonly used pattern recognition algorithms include decision trees, naive Bayes, support vector machines, etc. These algorithms can be trained and predicted based on data such as the user's search history, clicking behavior, etc., so as to accurately classify and sort the user's search intent.

2.3. Discover useful information and knowledge from the data

The purpose of data mining and pattern recognition is to discover useful information and knowledge from data. In search engines, this information and knowledge can be used to optimize search algorithms and recommendation strategies to improve user experience and search engine performance. For example, by analyzing a user's search history and clicking behavior, users' interest preferences and behavior patterns can be found, thus providing users with more relevant and personalized search

results[4]. At the same time, the information and knowledge can also be used for business analysis and decision support, providing valuable data support for enterprises

2.4. User behavior analysis

2.4.1. Feature extraction and analysis of user search behavior

User search behavior is one of the main interaction ways of search engine users, and its analysis can help search engine better understand the needs and intentions of users. Feature extraction is the basis of user search behavior analysis. Users' search features can be extracted by word segmentation, word frequency statistics, emotion analysis and other operations on search keywords and query statements. By analyzing these features, users' search preferences, demands and search habits can be found. The feature extraction of the user's search behavior mainly depends on the features you want to get from the search behavior. In general, features that you might want to extract might include the frequency of the user's search, the number of keywords searched per search, the keywords searched, the frequency of searches, and so on. For example, the search log is a CSV file that contains one search query per line[5].

More complex feature extraction is required in practical applications, such as calculating correlations between keywords, or using natural language processing (NLP) techniques to understand the user's search intent.

2.4.2. Modeling of user clicking behavior and browsing behavior

Users' clicking behavior and browsing behavior are important indicators that reflect their interests, preferences and needs. By modeling users' clicking behavior and browsing behavior, users' interest preferences and behavior patterns can be deeply understood. Common modeling methods include probabilistic model, collaborative filtering, deep learning, etc. By processing data such as users' click records and browsing records, users' interest models can be constructed and their future behavior trends can be predicted.

In Python, data analysis and machine learning libraries can be used to model user click behavior and browsing behavior.

For practical use, use a data set called user_clicks.csv, which contains data on the user's clicking behavior. The code first reads the data set and then extracts the features and labels. Next, the code divides the data set into a training set and a test set, and trains it using a random forest model. Finally, the code makes predictions on the test set and calculates the accuracy.

2.4.3. Recognition and prediction of user's interest preference and behavior pattern

Through the comprehensive analysis of the user's search behavior, click behavior,browsing behavior and other behaviors, the user's interest preference and behavior pattern can be identified. For example, some users may have a strong interest in electronic products, while others may pay more attention to entertainment news. By analyzing user behavior, users can be labeled with interests, which can better recommend content and personalize search. At the same time, by predicting user behavior, it can also provide users with content that may be of interest in advance to improve user experience.

The code uses the TF-IDF method to extract the features of the item, and uses cosine similarity to calculate the similarity between the user and the item. Then, the user behavior data is merged with the extracted features, and the logistic regression model is used to predict whether the user clicks on the item. Finally, a list of recommendations is generated based on the predicted results.

2.5. Personalized recommendation algorithm

2.5.1. Definition and objectives of personalized recommendation

Personalized recommendation is a technology that predicts the future needs and interests of users based on their historical behavior and interest preferences, and recommends relevant content for them. The goal of personalized recommendation is to improve the user experience, meet the individual needs of users, and improve the efficiency and accuracy of information acquisition[6]. In search engines, personalized recommendation can be used to provide users with personalized search results and related links, so as to meet users' personalized needs.

2.5.2. Content-based recommendation algorithm

Content-based recommendation algorithm is a method of making recommendations based on the content characteristics of an item. In search engines, content-based recommendation algorithms can be used to recommend items to users that relate to their search history or browsing history. The algorithm matches users' interests and preferences by analyzing the content features of the items, such as keywords, topics, categories, etc., so as to provide users with personalized search results and related links.

2.5.3. Collaborative filtering recommendation algorithm

Collaborative filtering recommendation algorithm is a method of making recommendations based on the user's behavior and the behavior of other users. In search engines, collaborative filtering recommendation algorithms can be used to recommend items to users who have similar interest preferences to other users. The algorithm finds the user's interest preferences and behavior patterns by analyzing the user's behavior and that of other users, so as to provide users with personalized search results and related links.

2.5.4. Evaluation of hybrid recommendation algorithm and personalized recommendation system

Hybrid recommendation algorithm is a method to combine a variety of recommendation algorithms so as to obtain better recommendation results. In search engines, hybrid recommendation algorithm can combine content-based recommendation algorithm and collaborative filtering recommendation algorithm, etc., and comprehensively consider the content characteristics of items and the behavioral characteristics of users, so as to provide users with more accurate and personalized search results and related links[7].

The evaluation of personalized recommendation system is one of the important steps to measure the performance of recommendation algorithm. The commonly used evaluation indexes include accuracy rate, recall rate, F1 score, AUC, etc. Through the evaluation of the recommendation results, the advantages and disadvantages of the recommendation algorithm can be found, so as to optimize and improve. At the same time, the parameters and strategies of the recommendation algorithm can be adjusted according to the evaluation results, so as to obtain better recommendation results.

3. APPLICATION OF ARTIFICIAL INTELLIGENCE IN SEARCH ENGINES

3.1. Machine learning algorithms

A machine learning algorithm is a technique that analyzes large amounts of data and automatically learns the characteristics and regularities of the data. In search engines, machine learning algorithms can be used to classify web pages, cluster them, analyze association rules, and mine and analyze data such as users' search history and click behavior. First of all, machine learning algorithms can be used to classify and cluster web pages. By classifying a large number of web pages, similar web pages can be grouped into the same category, making it easier for users to find the web pages they are interested

in more quickly. At the same time, by clustering web pages, related web pages can be gathered together to improve the relevance and accuracy of search results. Common machine learning algorithms include naive Bayes classifiers, support vector machines, decision trees, etc. Secondly, machine learning algorithms can also be used for association rule analysis. Association rule analysis can discover associations and dependencies between data, and thus discover association rules between web pages[8]. For example, if two web pages are clicked by many users at the same time, then there may be association rules between the two pages, and when a user searches for one of the pages, he can recommend another page that is associated with it. In addition, machine learning algorithms can also be used to mine and analyze data such as users' search history and click behavior. By analyzing a user's search history and click behavior, it is possible to understand the user's interest preferences and search habits, thereby optimizing the ranking of search results and personalized recommendations. For example, if a user frequently searches for topics related to science and technology, we can infer that the user is more interested in science and technology content, so we can prioritize related science and technology pages in their search results.

3.2. Deep Learning algorithms

Deep learning algorithms have a wide range of applications in search engines and can be used for semantic analysis and sentiment analysis of web pages, as well as mining and analysis of user behavior data.

Convolutional Neural Networks (CNNS): CNNS are a commonly used deep learning algorithm for processing data such as images and text. In search engines, CNNS can be used for image and text classification of web pages, sentiment analysis, keyword extraction and other tasks. For example, by classifying pictures and text in a web page, the topic and content of the web page can be judged, so that the web page can be accurately classified and indexed. Recurrent Neural Networks (RNN): RNN is a deep learning algorithm for processing sequential data. In search engines, RNNS can be used to mine and analyze data about a user's search history, clicking behavior, and more[9]. For example, by analyzing a user's search history and clicking behavior, it is possible to predict a user's interest preferences and search habits, so as to personalize the ranking and recommendation of search results.

Transformer: Transformer is a deep learning algorithm based on a self-attention mechanism that is suitable for processing long sequences of data. In search engines, Transformer can be used to model and analyze a user's search history and web content. For example, by analyzing a user's search history and web page content, the similarity between them can be calculated to accurately rank and recommend search results.

Generative Adversarial Networks (Gans): Gans are generative models that consist of two neural networks: a generator and a discriminator. In search engines, Gans can be used to generate new web content, such as generating articles, images, etc. that are relevant to a user's search. By training Gans, they can learn to generate content that meets the needs and interests of users, thereby improving the performance of search engines and user experience.

Deep learning algorithms require large amounts of data and computational resources to train and optimize. When building a search engine, it is necessary to select the right deep learning algorithm according to the specific application scenario, and adjust and optimize its parameters to improve the performance and accuracy of the search engine. At the same time, it is also necessary to deal with problems such as data imbalance and outliers to ensure the stability and reliability of the model.

3.3. Reinforcement learning algorithm

Reinforcement learning algorithms are also widely used in search engines and can be used to optimize search algorithms and improve search efficiency and accuracy.

Q-learning: Q-learning is a common reinforcement learning algorithm that guides agents in making decisions by building a Q table to record the Q value of each state and action. In search engines, Q-learning can be used to optimize search ranking algorithms to achieve more accurate ranking of search results. For example, through Q-learning algorithm, the search engine can learn to adjust the ranking of search results according to the user's search history and click behavior, so as to improve the user experience and the performance of the search engine.

Policy Gradient: Policy Gradient is a policy-based reinforcement learning algorithm that maximizes the expected return value by optimizing the strategy. In search engines, Policy Gradient can be used to optimize the recommendation strategy in the search algorithm to achieve a more accurate and personalized recommendation of search results. For example, through the Policy Gradient algorithm, search engines can learn to recommend relevant web pages and content based on users' interests, preferences and historical behavior[10].

Deep Q-Network (DQN) : DQN is an algorithm that combines deep learning and reinforcement learning to estimate Q values by building a deep neural network. In search engines, DQN can be used to optimize search ranking algorithms and web page ranking algorithms for a more accurate and efficient return of search results. For example, through the DQN algorithm, the search engine can learn to select the most relevant pages from a massive number of web pages, and make a personalized ranking according to the user's search history and click behavior.

Proximal Policy Optimization (PPO) : PPO is an efficient Policy Gradient algorithm that improves Policy Gradient stability by limiting differences between new and old strategies. In search engines, PPO can be used to optimize recommendation strategies and web page ranking strategies in search algorithms to achieve more accurate and efficient recommendation and return of search results.

Reinforcement learning algorithms require a lot of interaction and trial and error to learn optimal strategies. When building a search engine, a suitable reward function and environment need to be designed to guide the Agent to learn and make decisions. At the same time, it is also necessary to deal with the problem of data imbalance, outliers and other problems to ensure the stability and reliability of the model. In addition, the computational complexity of reinforcement learning algorithm is high, which requires efficient computing resources and optimized algorithm design to achieve efficient search engine performance.

3.4. Intelligence of personalized recommendation system

The personalized recommendation system based on artificial intelligence can realize intelligent recommendation through the analysis of user behavior data. Such a system can realize accurate prediction of users' interests and preferences and personalized recommendation with the help of various techniques such as machine learning, deep learning and reinforcement learning. In a personalized recommendation system, a common practice is to analyze users' historical behavior and interest preferences in order to provide personalized search results and recommendation services to users more accurately. For example, users' interest preferences and behavior habits can be learned by analyzing data such as the keywords they search in the search engine, the web pages they click on, and the time they browse. Then, based on this information, pages or content that are more in line with the user's interests and needs can be recommended.

Deep learning algorithms can play an important role in the process of realizing such intelligent recommendations. For example, convolutional neural networks (CNNS) can be used for feature extraction and classification of users' historical behavior data, thereby identifying users' interest preferences and behavior patterns. Recurrent neural networks (RNNS) can be used to model a user's search history and clicking behavior sequentially, so as to predict the user's next behavior. Generative adversarial networks (Gans) can be used to generate new content that matches a user's interests and needs, thereby enriching a user's search results and recommendation list.

In addition to deep learning algorithms, reinforcement learning algorithms can also be used to optimize personalized recommendation systems. For example, reinforcement learning algorithms can be used to optimize search algorithms to achieve more efficient and accurate return of search results. At the same time, reinforcement learning algorithms can also be used to optimize the recommendation algorithm to achieve more accurate and personalized recommendation of search results.

4. SUMMARY

With the development of the Internet, the application research of big data and artificial intelligence in search engines is becoming more and more important. The application of big data can realize large-scale data processing, improve the performance of search engine and user experience; And the application of artificial intelligence can improve the intelligence level of search engines through machine learning, deep learning and reinforcement learning and other technologies to achieve more accurate and efficient search results return and personalized recommendation services. In the future, with the continuous development of technology, the application of big data and artificial intelligence in search engines will be more and more extensive, providing people with more intelligent and efficient information retrieval services.

REFERENCES

[1] Zhang, X.Y. (2023) A study on the application of Big Data and Artificial Intelligence technology in English Teaching in Higher vocational Colleges. English Abroad, (5):238-240.

[2] Wu, S.Y. (2023) Application and research of big data and artificial intelligence technology in smart agriculture. Science and Information Technology, (14):160-162.

[3] Wang, H. (2023) Application of artificial intelligence and big data technology in digital marketing. Information and Computer.

[4] Zhang, W.Z., Ren, S. (2023) Railway data security and privacy protection technology system research. Journal of railway computer applications, 32 (11): 45-50.

[5] Yao, H.R. (2023) Big data analysis and mining technology in the application of marketing. Journal of computer science and artificial intelligence, 1 (4): 24-27.

[6] Wang, L. (2023) Research on Application of Artificial Intelligence in editing of Sci-tech journals. China media technology, 2023 (6): 95-98.

[7] Wang, M., Jia, F.Q. (2023) Artificial intelligence in the application of psychological consultation and treatment and development. Journal of psychologies, 18 (11): 227-230.

[8] Ding, S.Z., Jiang, X.Q., Meng,C. et al. (2023) Application of AI and big data technology to inverse synthesis route design. Science in China: Chemistry, 2023, 53(1):13.

[9] Xia, B.Q. (2023) Application research of big data, Internet of Things and Artificial Intelligence technology in offshore intelligent oilfield production. Digital Communication World, (2):121-123.

[10] Liu, H.Z., Liu, J.F. (2023) Big data and artificial intelligence in the diagnosis of primary liver cancer screening and application. Journal of the Chinese liver surgical operative surgery electronic journal, 12 (1): 5.