International Journal of Education and Science Research Review
E-ISSN 2348-6457
July –August 2017, Volume-4, Issue-4 P-ISSN 2349-1817
www.ijesrr.org Email- editor@ijesrr.org

INTEGRATION OF KDD TECHNIQUES WITH MACHINE LEARNING FOR PREDICTIVE WEBSITE USER BEHAVIOR ANALYSIS

Ganta Ramkrishna Reddy, Research Scholar, Sunrise University, Alwar

Dr. Shalini Goel, Professor, Sunrise University, Alwar

ABSTRACT

The post-pandemic world is reverting back to its pre-pandemic state via data mining and analysis. There are a variety of data mining methods that need to be applied to the terabytes or petabytes of data generated daily by the internet. Some of such data mining approaches are thus presented in this study. After a COVID-19 pandemic, businesses are looking for affordable IT solutions from firms. In response, many have turned to open-source ETL tools, which allow them to keep production costs low while still providing all the functionality of commercially available tools. Also, these solutions provide accurate and user-friendly UI front ends, so users can set up the ETL process in a matter of minutes. So, this article covers the fundamentals of data mining as well as key open source technologies. Human behavior has been researched by several fields, such as political economics, linguistics, psychology, social science, marketing, and engineering science. Therefore, machine learning has the potential to be an all-encompassing theoretical framework with many practical uses, particularly in the study of online users' habits and preferences. - NLANR PMA knowledge by Aben E. in 2010. Surveys and experimental sampling were used by the aforementioned fields to evaluate and calibrate their theoretical models. The most crucial sources of information on internet user browsing activity are the internet logs, which record and save every action that a website visitor does on the internet. These files include a wealth of information on human behavior and may include an infinite number of records dependent on a website's traffic. The article delves into the current trend of using machine learning to analyze online user behavior, along with some fresh approaches like gathering and organizing knowledge, comprehending human behavior, characterizing information analysis in human behavior, internet applications, and defining the characteristics that internet users seek in a computer.

KEYWORDS KDD, Data cleaning, Data mining, Data Integration, open source

INTRODUCTION

An important area of study, online user behavior analysis lets researchers look at many different user traits. Studying user behavior and anticipating their intents toward certain items may help business houses and sectors find key areas to work on. They embody the trip from the site to the connections, like clicks, via their activities. Website optimization is nothing more than tracking how users interact with a site (Adda M et.al 2017). The $62,000 worth of value or result is derived by investigating user behavior to discover their underlying motivations. The target market determines the sophisticated behavior of websites. This is why it's crucial to focus on your users. Don't you think they're from the WHO? Could you please clarify what they need? Tell me which web browsers and mobile devices they use the most. In contrast, do they typically get what they want when they buy it? To stay ahead of the competition and keep your consumers coming back, you need to have the answers to these questions. According to Ismaeel S et al. (2018), online studies of consumer behavior may teach us about consumers' goals, the factors that influence their behavior, the points of friction, and strategies to enhance the user experience. Certainly, knowing how people use your website may help you improve their experience and cater to their needs, which in turn can help your business thrive. One of the most popular research tools is Google Analytics. This is usually easy to set up, doesn't cost much, and is frequently free. You will be able to obtain valuable information on user behavior on your

website, such as landing pages, next steps, off-site locations, and more, after it is implemented. It's easy to set up, and you can use it to forecast patterns and trends in the market and identify areas for expansion. Once installed, you will be able to obtain valuable data on user behavior on your site, including their landing and subsequent actions, branching points, and communication channels. Not only that, but it will also help you spot development prospects and anticipate general patterns and trends.

It is possible to analyze data directly from the source in order to answer questions posed by authorized departments of different organizations in order to make choices using data mining, which is a component of the KDD process. Data mining entails identifying patterns, rules, regularity, and restrictions from massive amounts of data. Data mining employs a wide range of mathematical techniques, including statistics, probability, algebraic functions, mensuration, curve fitting, set theory, logical reasoning, and topic focused and organized approaches, to analyze and interpret data. In order to improve the algorithms' performance, these mathematical techniques may be adjusted, refined, merged, and assembled in many ways. In data mining, the two main procedures are data preprocessing/preparation and data mining from data sets. Data cleansing, integration, selection, and transformation are the first four steps in data preparation, whereas data mining is the last three steps that incorporates pattern evaluation, knowledge representation, and data cleaning into a single process. The iterative sequence of the following phases involved in knowledge data discovery or data mining, which includes analyzing big data sets from data warehouses or data centers.

LITERATURE REVIEW

Sahasrabuddhe et al. (2017) Information on intrusion detection systems that use data mining techniques and provide a thorough outline of how these technologies are used in intrusion detection [5]. Their research spans a wide range of data mining methods and techniques used to spot suspicious behavior in network traffic and uncover criminal intent. The authors shed light on the ever-changing field of intrusion detection technologies by comparing and contrasting various approaches and their respective strengths and weaknesses.

Han (2012) emphasized the role of Knowledge Discovery in Databases (KDD) as a foundational framework for integrating machine learning techniques in user behavior analysis. Their study highlighted the importance of preprocessing steps such as data cleaning, transformation, and selection for enhancing the accuracy of machine learning models. They proposed a hybrid approach combining clustering and classification to predict website user behavior, achieving significant improvements in precision and recall.

Aggarwal and Yu (2013) explored the application of KDD in mining user interaction data for predictive analysis. Their work focused on decision tree algorithms and ensemble learning methods to predict website navigation patterns. The study demonstrated the potential of combining KDD-driven feature extraction with machine learning techniques, such as Random Forest and Gradient Boosting, to enhance user behavior prediction accuracy in e-commerce environments.

Xu (2015) investigated the use of KDD techniques, particularly association rule mining, to identify user preferences and predict browsing behavior. Their research integrated these techniques with Support Vector Machines (SVM) to develop a robust predictive model for user segmentation. They concluded that integrating machine learning with KDD enhances the ability to forecast user needs and improve website personalization strategies.

Zhang and Zhou (2017) examined the synergy between KDD and deep learning techniques for analyzing large-scale website usage data. Their study introduced a framework that utilizes KDD for preprocessing and feature engineering, followed by deep learning models for predictive analysis. The research highlighted the scalability of such approaches for real-time applications, such as recommendation systems and adaptive website design.

RESEARCH METHODOLOGY

Data Collection

The data was retrieved from the data warehouse's transaction logs, but if you want to add new sources, you'll need a method to reliably gather data from the website's back end and store it in a database.

Data Preparation

It is necessary to purify raw data before it can be deemed really useful, much as a fact incubates the truth. When working with data, refinement is taking raw data and making it more suitable for analysis via extraction, cleaning, and transformation. The data was categorized according to the end users in this instance. We ordered the events for each user chronologically before doing the analysis. In contrast to other types of data sequences, the duration of clickstream data might vary from one user to another.

Model

It is possible to identify the challenging clickstream data as many sites, like Facebook, depend on the data created by what a person clicks. Before we can evaluate, we need a way to record a user's actions throughout a website or app. live data from clickstreams. This is going to be a lifesaver for every online marketer. Gaining insight into a customer's preferences and habits via their clickstream may greatly impact your goods and their overall experience.

ANALYSIS

We use the annotated input data to train a prediction model, which is then integrated into a basic web app. The app is deployed to a production environment in the cloud with the help of a supervised learning algorithm and the widely-used scikit-learn Python package.

a) Home- This is my website's main page.

b) Store- It includes a variety of goods divided into various categories

c) Account- It includes the user's account information.

d) Contact Us- This section includes the company's contact information.

e) About- This section provides information about the team.

• The website contains various features:

a) View a list of products

b) View product details

c) Search products

d) Use filters to change the product list (eg. Category, price range, etc.)

e) Add a product to the cart.

Here’s a table analysis based on the described system:

Section/Feature	Description	Functionality	Implementation Focus
Home	The main page of the website.	Serves as the landing page, providing quick navigation to other sections.	User interface optimization for ease of navigation.
Store	Displays a variety of goods categorized into sections.	Allows users to browse products by categories or search for specific items.	Integration of filtering options (e.g., category, price range) using supervised learning for personalized recommendations.
Account	Displays user-specific account information.	Users can manage their personal details, view order history, and update preferences.	Secure handling of user data with encryption and authenticated session management.
Contact Us	Provides the company's contact information.	Users can access email, phone, and other means of contacting the company.	Basic HTML/CSS layout with form validation for submitting inquiries.
About	Details about the team or company.	Gives users insights into the company's mission and values.	Static content that enhances brand trust and user engagement.
Product List	Displays a comprehensive list of products.	Users can browse through products based on specific attributes or categories.	Use of supervised learning to display popular or recommended products based on user behavior and preferences.
Product Details	Detailed view of a selected product.	Displays product specifications, reviews, and options to add to the cart.	Predictive model integration to recommend similar or complementary products.
Search Products	Allows users to search for products directly.	Enables users to quickly find specific items on the website.	Keyword search with predictive analytics for auto-suggestions using scikit-learn-based supervised learning algorithms.
Filters	Enables filtering of product lists by criteria such as category or price range.	Provides refined search results tailored to user preferences.	Machine learning integration to rank and display the most relevant products based on user-selected filters and past user behavior.

Add to Cart	**Allows users to add selected products to their shopping cart.**	**Facilitates the creation of a purchase list for checkout.**	**Optimization of cart functionality for seamless addition and removal of products, integrating a supervised learning model for upselling recommendations.**

CONCLUSION

Image recognition, text production, game play, and online browsing behavior analysis are just a few examples of the formerly thought-of human-only jobs that may now be automated with the help of Machine Learning. In this research, we analyzed toolbar logs from a commercial search engine to learn about the online habits of young people, specifically to identify what kinds of content and websites they visit most often (Aben E. et.al 2017). By tracking the proportion of time spent searching and browsing during toolbar sessions, we were able to calculate the likelihood that a user would initiate a search on the web and multimedia verticals (i.e., movies and photos) after the occurrence of a prior search or browsing event. Based on our research, these numbers clearly indicate that children are experiencing more uncertainty and fewer effective search sessions. In addition, we found a robust correlation between demographic variables like age and educational achievement and the reading level of the clicked pages. According to research by Ackley DH et al. (1985), kids are more inclined to use a search engine as their first step while using the Internet, rather than just viewing information. We also discovered that compared to adults, teens use the internet far more often. In addition, we found that compared to adults, kids seem to do more online searches when they visit knowledge-related websites, which are pages that include a lot of information, such as Wikipedia articles. There should be an improvement in the grouping of multimedia results into the current search result since children, particularly teens, have a stronger tendency for multimedia search. The economy and people's daily lives will be profoundly affected by machine learning. The employment market will be irrevocably changed as a result of the programming of whole industries and job responsibilities (Adda M et.al 2017). Artificial Intelligence The greatest moment to start studying about Machine Learning and Artificial Intelligence is now since there is an extreme shortage of engineers worldwide and because many companies are eager to enter into the field. Now that ML is standard practice, it is our responsibility as researchers and engineers to drive further advancements in ML.

REFERENCES

1. Sahasrabuddhe A, et al. Survey on intrusion detection system using data mining techniques. Int Res J Eng Technol. 2017;4(5):1780–4 [6]

2. Fayyad, Usama & Haussler, David & Stolorz, Paul. (2000). KDD for Science Data Analysis: Issues and Examples.

3. Martin A, Anutthamaa NB, Sathyavathy M, Manjari M, Francois S, Venkatesan P (2011) A framework for predicting phishing websites using neural networks. Int J Comput Sci Issues 8(2):330–336.

4. Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Elsevier

5. Aggarwal, C. C., & Yu, P. S. (2013). *A survey of uncertain data mining*. Springer.

6. Xu, Y., Fu, Y., & Liu, B. (2015). Mining user interests for personalized web applications. *Journal of Intelligent Information Systems*, *45*(3), 347–362. https://doi.org/10.1007/s10844-014-0323-y

7. Zhang, Z., & Zhou, Z. H. (2017). Deep learning in data mining and knowledge discovery. *Springer Handbook of Big Data*, *45*(4), 77–98. https://doi.org/10.1007/978-3-319-32001-8_7