International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 11 | Nov 2023 www.irjet.net p-ISSN: 2395-0072

Product Comparison Website using Web scraping and Machine learning.

Aswad Shaikh[1], Aniket Sonmali[1], Soham Wakade[1]

[1]*Student at Department of Information Technology, Atharva College of Engineering, Mumbai, Maharashtra, India*

Abstract - *The use of web scraping techniques to collect and compare data from multiple product websites has become increasingly popular in recent years. This research paper explores the development and implementation of a product comparison website using web scraping techniques. The website extracts data from various product websites. The collected data is then compared using a customized algorithm that takes into account various factors such as price, features, and user ratings. The website provides users with a comprehensive comparison of products in a specific category and helps them make informed decisions about which product to purchase. The implementation of this website has shown that web scraping can be an effective tool in collecting and analyzing data for product comparison websites.*

Furthermore, the website developed in this research can be used as a template for developing similar websites in other categories.

Key Words: **Product comparison, Recommendation, Web scraping, Relevance, Machine learning, Price**

1. INTRODUCTION

Product comparison websites have become increasingly popular in recent years, as consumers seek to make informed purchasing decisions in a crowded and complex marketplace. These websites provide users with the ability to compare products across multiple dimensions, including price, quality, and features. However, as the number of products and the amount of information available online continues to grow, the challenge of effectively presenting this information to users becomes more complex. In this research paper, we will explore the topic of product comparison websites and their potential for improving the user experience through the incorporation of machine learning. Specifically, we will examine how machine learning can be used to enhance the functionality of product comparison engines, making them more visually appealing, easier to use, and more accurate in their recommendations. Through our research, we hope to provide insights into the future of product comparison websites and the potential impact of machine learning on their development. We will begin by providing an overview of the current state of product comparison websites, followed by a review of relevant literature in the field of machine learning. We will then present our methodology for incorporating machine learning into our existing product comparison engine and discuss our findings. Ultimately, we hope to demonstrate the value of machine learning in improving the user experience of product comparison websites.

1.1 Motivation

In offline shopping, customers need to walk through a number of stores in order to get the most affordable or the best deal on a product. This sometimes proves hectic for a customer and at also time gets invested. Online shopping has made it easier for customers to browse through different online shopping stores by just a few clicks and without roaming in the market. But as in offline shopping customer has many shops as options, the same problem is faced during online shopping also as there are many online shopping applications available. So, to find the best deal a customer has to browse through a number of online stores. This problem is solved by product comparison engines where a customer gets to see the price, specs, etc. of a product at a single glance that is available on various online stores. Also, with the implementation of machine learning the user is recommended similar products according to the search query and the irrelevant products that appear to the user are sorted by the relevance.

1.2 Problem Statement

Consumers face difficulties in making informed purchasing decisions due to the overwhelming number of options available for a particular product category. This is especially true in the age of e-commerce where there are numerous online stores offering a wide range of products. As a result, consumers often spend a lot of time researching and comparing products across different websites, which can be time-consuming and confusing. A product comparison website aims to solve this problem by providing a one-stop platform for consumers to easily compare features, prices, and reviews of various products across multiple brands and retailers, helping them make informed decisions quickly and easily.

2. LITERATURE REVIEW

1. Lu Jiang, Zhaohui Wu, Jun Liu, Qinghua Zheng (Jan 2009) The key to Deep Web crawling is to submit promising keywords to query form and retrieve Deep Web

content efficiently. To select keywords, existing methods make a decision based on keywords' statistic information deriving from TF and DF in local acquired records, thus work well only in textual databases providing full text search interfaces, whereas not well in structured databases of multi-attribute or field-restricted search interfaces. This paper proposes a novel Deep Web crawling method. Keywords are encoded as a tuple by its linguistic, statistic and HTML features so that a harvest rate evaluation model can be learned from the issued keywords for the un-issued in future. The method breaks through the assumption of plain-text search made by existing methods. Experimental results show that the method outperforms the state of the art methods.

2. Simon Philip, P.B.Shola, Abari Ovye John (October2014.) Recommender systems are software applications that provide or suggest items to intended users. These systems use filtering techniques to provide recommendations. The major ones of these techniques are collaborative-based filtering technique, content-based technique, and hybrid algorithm. The motivation came as a result of the need to integrate recommendation feature in digital libraries in order to reduce information overload. Content-based technique is adopted because of its suitability in domains or situations where items are more than the users. TF-IDF (Term Frequency Inverse Document Frequency) and cosine similarity were used to determine how relevant or similar a research paper is to a user's query or profile of interest. Research papers and user's query were represented as vectors of weights using Keyword-based Vector Space model. The weights indicate the degree of association between a research paper and a user's query. This paper also presents an algorithm to provide or suggest recommendations based on users' query. The algorithm employs both TF-IDF weighing scheme and cosine similarity measure. Based on the result or output of the system, integrating recommendation feature in digital libraries will help library users to find most relevant research papers to their needs.

3. F. M. Javed Mehedi Shamrat, Zarrin Tasnim, A.K.M Sazzadur Rahman, Naimul Islam Nobel, Syed Akhter Hossain (Jan 2020.) The World Wide Web (WWW) is a web customer server design. It is an incredible framework dependent on complete independence to the server for serving data accessible on the web. The data is masterminded as a huge, circulated, and non-direct content framework known as the Hypertext Document framework. These frameworks characterize some portion of a report as being hypertext-bits of content or pictures which are connected to different records by means of stay labels. HTTP and HTML present a standard method for recovering and introducing the hyperlinked records. Web programs, use web crawlers to investigate the servers for required pages of data. The pages sent by the servers are prepared at the customer side. Presently days it has turned into a significant piece of human life to utilize Internet to obtain entrance data from WWW. The present populace of the world is about 7.049 billion out of which 2.40 billion individuals (34.3%) use Internet [1] (see Figure 1). From .36 billion of every 2000, the measure of Internet clients has expanded to 2.40 billion out of 2012 i.e., an expansion of 566.4% from 2000 to 2012. In Asia out of 3.92 billion individuals, 1.076 billion (i.e.27.5%) use Internet, though in India out of 1.2 billion, .137 billion (11.4%) use Internet. The same development rate is normal in future as well and it isn't far away when one will begin reasoning that life is deficient without Internet. Figure 1: outlines Internet Users in the World by Geographic Regions.

4. Mr. Santosh Kumar Mishra, Mukul Singhal,Dikshant Awasthi, Parshvi Verma,Sumit Kumar Malik (June 2020.)

Search engines have become the dominant model of online search. Large and small ecommerce provide built-in search capability to their visitors to examine the products they have. While most large business are able to hire the necessary skills to build advanced search engines, small online business still lack the ability to evaluate the results of their search engines, which means losing the opportunity to compete with larger business .Due to today's transition from visiting physical stores to online shopping, predicting customer behaviour in the context of e- commerce is gaining importance. It can increase customer satisfaction and sales, resulting in higher conversion rates and a competitive advantage, by facilitating a more personalized shopping process. With the rapid growth of e-Commerce, online product search has emerged as a popular and effective paradigm for customers to find desired products and engage in online shopping. However, there is still a big gap between the products that customers really desire to purchase and relevance of products that are suggested in response to a query from the customer. In this synopsis, we propose a robust way of predicting relevance scores given a search query and a product, using techniques involving machine learning, natural language processing and information retrieval.

3.METHODOLOGY

1) Web Crawler:

The system deals with price comparison engine. The first thing required are to gather large amount of data from different e- commerce websites. It is not possible to manually collect the data from websites. Hence the best way is to create a web crawler that will navigate to these e-commerce websites. The fetched URLs are sent to scrapper for scrapping process.

2) Web Scrapper:

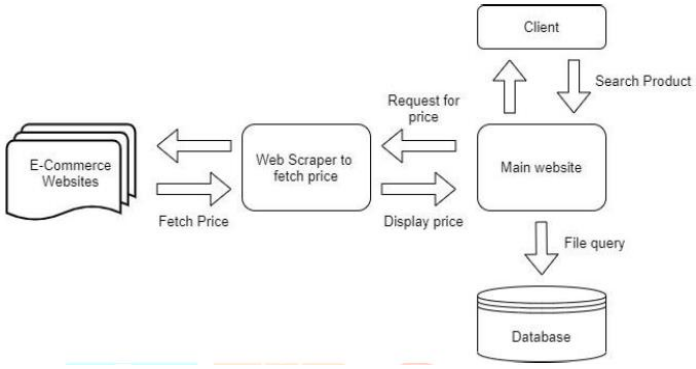
Web Scrapping is used to extract HTML data from URL's and use it for personal purpose. As this is price comparison website, data is scrapped from multiple e-commerce websites.

3) Relevance Filter:-

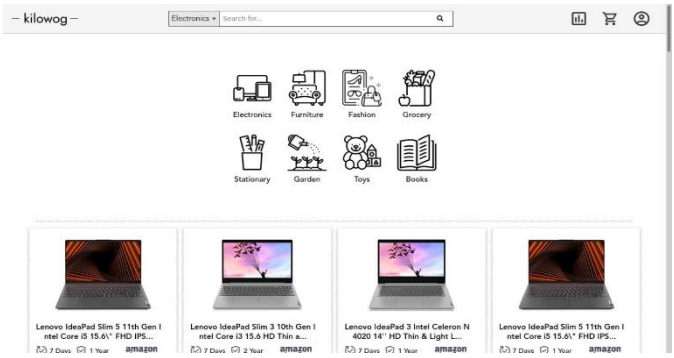
Relevance filtering is a technique used in our system to filter out irrelevant results and only present the most relevant ones to the user. It involves using algorithms and heuristics to score and rank search results based on their relevance to the user's query. This helps to improve the overall search experience and increase the likelihood of finding the desired information.

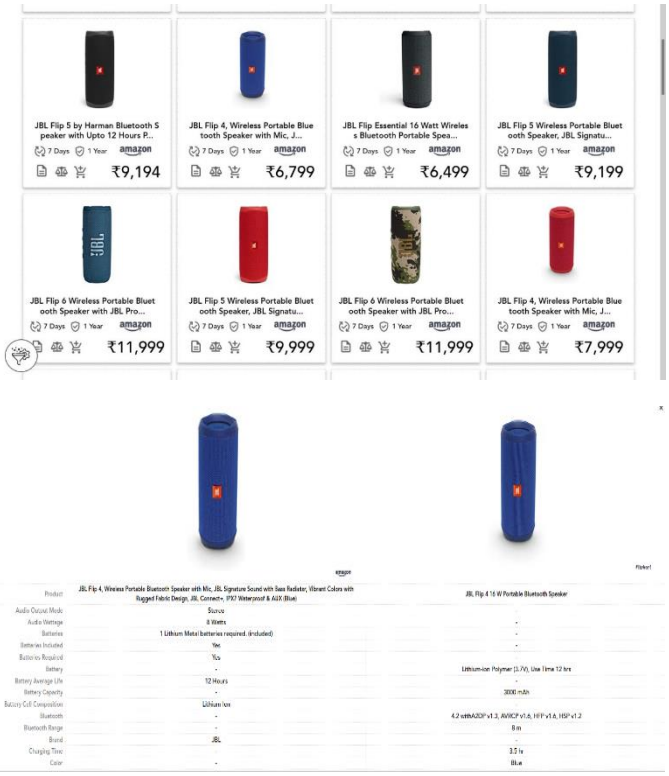
4)Recommendation System:-

A recommendation system is a type of information filtering system that provides personalized recommendations to users based on their preferences, past behavior, and other relevant factors. It is commonly used in e-commerce, social media, and content streaming platforms to suggest products, content, or other items of interest to the user.



4. IMPLEMENTATION





5. CONCLUSIONS

In conclusion, the creation of a product comparison website with relevance and recommendation features using machine learning (ML) is a complex and challenging task that requires expertise in various programming languages and technologies. In this project, we utilized JavaScript, MongoDB, Python, and Rust to develop a website that can provide users with accurate and relevant product comparisons and recommendations based on their preferences.

The website's relevance and recommendation features were implemented using ML algorithms that analyzed user behavior and preferences to provide personalized recommendations. The use of ML technology has enabled the website to become more accurate and effective over time as it learns from user interactions and feedback.

Overall, this project demonstrates the power and potential of ML in developing intelligent and personalized websites that can enhance the user experience and provide valuable insights to businesses. As more data becomes available and more advanced ML algorithms are developed, we can expect even greater improvements in website recommendations and relevance in the future.

6. FUTURE WORK

There are several potential avenues for future work that can be explored to enhance the product comparison

website with relevance and recommendation features using machine learning.

One area for future work is the implementation of deep learning algorithms, such as neural networks. These algorithms can enable the website to learn more complex patterns and relationships in user behavior and preferences, leading to more accurate and effective recommendations. By incorporating deep learning algorithms, the website can better understand the nuances of user behavior and preferences, which can lead to more personalized and relevant recommendations.

In addition, the integration of external data sources, such as social media and other online platforms, can provide additional insights into user behavior and preferences. By incorporating these external data sources, the website can gain a more comprehensive understanding of user preferences and behaviors, which can be used to further enhance its recommendation algorithms.

Finally, the implementation of user feedback mechanisms, such as rating and review systems, can enable the website to learn from user interactions and continuously improve its recommendations over time. By incorporating user feedback, the website can adapt to changing user preferences and behaviors, leading to more accurate and effective recommendations.

Overall, future work on this project can focus on incorporating advanced machine learning techniques, integrating external data sources, and implementing user feedback mechanisms to enhance the website's accuracy and effectiveness in providing relevant and personalized product recommendations.

REFERENCES

[1] Jiang, L., Wu, Z., Zheng, Q., & Liu, J. (2009, September). Learning deep web crawling with diverse features. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 572-575). IEEEM. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Philip, S., Shola, P., & Ovye, A. (2014). Application of content-based approach in research paper recommendation system for a digital library. *International Journal of Advanced Computer Science and Applications*, *5*(10).

[3] Shamrat, F. J. M., Tasnim, Z., Rahman, A. S., Nobel, N. I., & Hossain, S. A. (2020). An effective implementation of web crawling technology to retrieve data from the world wide web (WWW). *International Journal of Scientific & Technology Research*, *9*(01), 1252-1256.

[4] Mr. Santosh Kumar Mishra, Mukul Singhal,Dikshant Awasthi, Parshvi Verma,Sumit Kumar Malik," Predict the relevance of search results from Ecommerce sites", 29 June 2020.

[5] Javed, U., Shaukat, K., A. Hameed, I., Iqbal, F., Mahboob Alam, T. & Luo, S. (2021). A Review of Content-Based and Context-Based Recommendation Systems. International Journal of Emerging Technologies in Learning (iJET), 16(3), 274-306. Kassel, Germany: International Journal of Emerging Technology in Learning. Retrieved April 14, 2023 from https://www.learntechlib.org/p/219036/.

[6] Zisopoulos, Charilaos & Karagiannidis, Savvas & Demirtsoglou, Georgios & Antaris, Stefanos. (2008). Content-Based Recommendation Systems.

[7] Singrodia, V., Mitra, A. and Paul, S., 2019, January. A review on web scraping and its applications. In 2019 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.

[8] Vargiu, E. and Urru, M., 2013. Exploiting web scraping in a collaborative filtering-based approach to web advertising. Artif. Intell. Res., 2(1), pp.44-54.

[9] Alam, A., Anjum, A.A., Tasin, F.S., Reyad, M.R., Sinthee, S.A. and Hossain, N., 2020, June. Upoma: A Dynamic Online Price Comparison Tool for Bangladeshi E-commerce Websites. In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 194-197). IEEE.

[10] Julian, L.R. and Natalia, F., 2015, November. The use of web scraping in computer parts and assembly price comparison. In 2015 3rd International Conference on New Media (CONMEDIA) (pp. 1-6). IEEE.

[11] Hillen, Judith. \"Web scraping for food price research.\" British Food Journal (2019).

[12] Milev, Plamen. \"Conceptual approach for development of web scraping applications for tracking information.\" Economic Alternatives 3 (2017): 475-485.

[13] Himawan, Arif, Adri Priadana, and Aris Murdiyanto. \"Implementation of Web Scraping to Build a Web-Based Instagram Account Data Downloader Application.\" IJID (International Journal on Informatics for Development) 9, no. 2 (2020): 59-65.

[14] Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

[15] Kim, SW., Gil, JM. Research paper classification systems based on TF-IDF and LDA schemes. Hum. Cent. Comput. Inf. Sci. 9, 30 (2019). https://doi.org/10.1186/s13673-019-0192-7

[16] Rabiyatou DIOUF, Edouard Ngor SARR,Ousmane SALL,Babiga BIRREGAH,Mamadou BOUSSO,Sény Ndiaye

[17] MBAYE,"Web Scraping: State-of-the-Art and Areas of Application " -2019 IEEE International Conference on Big Data (Big Data).

[18] David Mathew Thomas, Sandeep Mathur"Data Analysis by Web Scraping using Python"-Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019].

[19] Riya Shah, Karishma Pathan, Anand Masurkar, Shweta Rewatkar, Prof. (Ms.) P.N.Vengurlekar "Comparison of Ecommerce Products using web mining"-International Journal of Scientific and Research Publications, Volume 6, Issue 5, May 2016.

[20] Kursa, Miron & Rudnicki, Witold. (2011). The All Relevant Feature Selection using Random Forest.