IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Exploring Diverse Search Algorithm Methods for EdTech Freelance Platforms

[1]Aryan Kokane, [2]Hetasvi Maheta, [3]Rajesh Kumawat, [4]Sahil More, [5]Sandeep Kulkarni

[1,2,3,4] Student, [5]Professor

School of Engineering,
Ajeenkya Dy Patil University, Pune, India

Abstract: Online education has seen the rise of EdTech freelancing platforms, which have become essential tools for learners, educators, and freelancers. These platforms function as dynamic marketplaces, providing a wide range of educational material and services to a worldwide audience. A key factor in their success is their ability to provide effortless content discovery and interaction through advanced search algorithms and optimization of user experience. This study explores the utilization of Term Frequency-Inverse Document Frequency (TF-IDF) algorithms to improve content discovery in EdTech freelance platforms. TF-IDF is a fundamental approach in information retrieval that offers a strong foundation for evaluating textual input and generating customized search results. TF-IDF allows platforms to prioritize material that fits with user preferences and search intent by giving weights to phrases based on their value within each learning resource compared to the total corpus. This study examines the installation, difficulties, and benefits of TF-IDF algorithms in enhancing search accuracy, user satisfaction, and overall platform efficiency. This paper explores the important role of semantic analysis in defining the future of online education by thoroughly examining TF-IDF-based search engines. TF-IDF plays a crucial role in driving innovation and change in EdTech freelancing platforms, as the need for individualized and accessible learning experiences increases. It promotes a worldwide culture of lifelong learning and information sharing.

Index Terms - EdTech freelancing platforms, Search algorithms, TF-IDF, Cosine Similarity, User experience, Content analysis, Online education

I. INTRODUCTION

EdTech freelance platforms are crucial in the current ever-changing online education environment as they provide a diverse range of educational material and services to learners and educators worldwide. These platforms function as dynamic markets where instructors share their knowledge, learners access high-quality resources, and freelancers provide their talents in content production, design, and tutoring. An essential factor in their achievement is the effortless identification of material, facilitated by advanced search engines and improvement of user experience.[12] The essence of content discovery on these platforms is the complex interaction between search algorithms and the improvement of user experience. The algorithms, serving as the platform's engine, examine user searches, content information, and interactions in order to provide customized and pertinent search results. They utilize innovative methodologies like Term Frequency-Inverse Document Frequency (TF-IDF) to extract valuable information from textual material, ensuring precise and contextually appropriate search results.[1] TF-IDF is a crucial approach in information retrieval that offers a strong framework for improving the accuracy and user involvement in search results on EdTech freelancing platforms. TF-IDF applies weights to words based on their relevance in describing content by assessing phrase frequencies within each resource relative to the overall corpus. By doing semantic analysis, computers are able to find and rank resources that are relevant to user preferences, hence improving the overall search experience. An important benefit of TF-IDF-based algorithms is their capacity to adjust to the various and changing requirements of users.[9] TF-IDF, unlike typical keyword-based methods, takes into account the

semantic significance of terms within each resource context, leading to more precise search results. By comprehending the context, we are able to provide customized suggestions that are specifically designed to meet the individual user's interests and learning goals, resulting in a more engaging and immersive learning journey. Moreover, TF-IDF algorithms provide clarity and comprehensibility, enabling people to comprehend and have confidence in search outcomes. Through the use of textual analysis, users may get valuable understanding of the significance of results. This enables them to make well-informed decisions, so increasing their sense of autonomy and ownership in the learning process. Moreover, the interpretability of TF-IDF scores enables the continual optimization process by considering user input and content trends. This ensures a constant enhancement in search accuracy and user pleasure.[7] TF-IDF-based search algorithms are powerful tools that may improve content discovery and user experience on EdTech freelancing platforms. By utilizing semantic analysis of textual data, these algorithms enable learners and instructors to browse educational materials with confidence. However, it is crucial to tackle the issues of scalability and efficiency in order to maintain the effectiveness of TF-IDF algorithms in enabling smooth content search and knowledge acquisition. With the increasing demand for online education, continuous research and innovation in search algorithms will play a crucial role in providing excellent user experiences and promoting platform expansion.

II. LITERATURE REVIEW

The expanding EdTech sector has facilitated the emergence of freelancing platforms that link educators with learners. Efficiently optimizing search algorithms on these platforms is essential for achieving effective matching and ensuring a great user experience. This literature study examines several search algorithm techniques that may be used on EdTech freelancing platforms. It takes into account issues such as user preferences, expertise matching, and platform-specific requirements.

Exploring the use of E-commerce in the field of EdTech:
Previous studies on e-commerce platforms provide useful insights that may be used for freelancing in the EdTech industry. Research conducted by Wu et al. and Zhou et al. examines the conflict between search accuracy and user discovery. Enhancing accuracy can enhance click-through rates [2], but it may limit exploration and decrease long-term engagement [2]. When creating their search algorithms, EdTech platforms must carefully evaluate this trade-off.

Recommendation systems and personalization:
Recommended system techniques may be used to customize search results on EdTech freelancing platforms. [4] illustrate the capabilities of Natural Language Processing (NLP) in recommendation algorithms on platforms such as Netflix. Similarly, educational technology systems can employ natural language processing (NLP) to comprehend user queries and customize search outcomes according to learner profiles and previous interactions with educators. Early studies, such as Dias et al. (2008), highlighted the value of personalized recommender systems in e-commerce settings. Their work suggests potential benefits for adapting these techniques in the EdTech sector.[6] Research like Mehta and Gupta (2021) explores how sentiment analysis, alongside techniques like cosine similarity, can be integrated into recommendation systems, offering potential benefits for EdTech platforms where learner preferences and reviews are important.[8]

Expanding Beyond Keyword Matching: Utilizing Expertise Matching and Skill-based Ranking:
Although keyword matching is a factor, EdTech freelancing platforms necessitate a subtler and refined approach. Research, such as the study conducted by [2], explores the significance of data in educational settings. EdTech platforms may utilize data on educators' abilities, experience, and student learning results to develop sophisticated expertise matching algorithms that emphasize achieving effective learning outcomes. Studies like Guo and Yang (2016) demonstrate the value of fine-tuning keyword weighting, using techniques like TF-IDF, to improve the precision of search results on EdTech platforms.[7]

EdTech platforms can push the boundaries of these algorithms by exploring hybrid similarity measures that combine structural data with traditional rating patterns.[5]

Further factors to take into account are user reviews, reputation, and trustworthiness:
Incorporating user evaluations and educator ratings may greatly improve search results. Favorable evaluations can enhance the prominence of educators and establish credibility with students seeking competent professors.

III. METHODOLOGY

3.1 Understanding the implementation of search algorithm in e-commerce platform

3.1.1 Understanding the Particular circumstance:

Our primary focus is on an educational technology (EdTech) platform that functions as a centralized online centre for learning materials, courses, and communication between instructors and learners. EdTech platforms are essential for enabling remote learning, individualized education, and the enhancement of skills.

3.1.2 Analysing content trends:

The primary goal is to examine the prevailing patterns of material on the EdTech platform. This entails analysing the variety of courses available, prevalent subjects, metrics related to user involvement, and user preferences. By comprehending these patterns, we may customize the search functionality to more effectively fulfil the requirements and preferences of platform users.

3.1.3 Utilizing Natural Language Processing (NLP) Techniques:

Natural Language Processing (NLP) approaches empower us to efficiently handle and evaluate textual data. Our approach involves utilizing two unique natural language processing techniques: TF-IDF and Cosine Similarity.

3.1.4 TF-IDF:

TF-IDF, which stands for Term Frequency-Inverse Document Frequency, is a numerical statistic used in information retrieval and text mining. It quantifies the importance of a term in a document by taking into account both the frequency of the term in the document and the frequency of the term in the whole collection of documents.[9]

TF-IDF is a statistical metric employed to assess the significance of a phrase in a text or collection of documents. The algorithm computes a numerical score for each phrase by considering its frequency inside a single text and its scarcity across numerous documents. This methodology enables us to discern crucial phrases and concepts included in the information accessible on the EdTech platform.[10]

3.1.5 Cosine Similarity:

Cosine Similarity is a mathematical measure used to assess the similarity between two text documents by calculating the cosine of the angle formed by their vector representations in a multi-dimensional space. Through the assessment of the resemblance between search queries and course descriptions or other textual information on the platform, we are able to offer consumers search results that are pertinent to their needs.[10]

3.1.6 Development of search feature:

By utilizing the knowledge acquired from analysing content trends and employing natural language processing techniques, we provide a search functionality for the EdTech platform. This functionality enables users to enter search queries and obtain customized and pertinent outcomes by analysing course names, descriptions, teacher profiles, and other textual data accessible on the site.

3.1.7 User Experience considerations:

During the development process, our main focus is on prioritizing the user experience, with the goal of creating a search interface that is both smooth and intuitive. This encompasses the enhancement of search capability to achieve optimal speed and accuracy, the provision of filters and sorting choices, and the integration of user input to consistently enhance the search feature.

3.1.8 Integration with platform features:

The search functionality is smoothly incorporated into other platform features, including personalized learning routes, course suggestions, and interactive forums. This comprehensive approach improves the overall user experience and encourages active participation with the platform's content and community.

3.1.9 Data Grouping:

The dataset was obtained from an EdTech platform and consists of several data points, including course titles, categories, instructors, course descriptions, ratings, and durations.

3.1.10 Data Pre-processing:

Data pre-processing is an essential and crucial stage in the preparation of unprocessed data for analysis. For our investigation, we employed several pre-processing approaches to purify and enhance the data prior to doing exploratory analysis.

3.1.11 Tokenization:

Tokenization is the process of dividing a sequence of text into smaller units, called tokens. Tokenization is the act of dividing text into smaller pieces known as tokens, which can be words, phrases, or characters. In this context, tokenization refers to the process of dividing textual material, such as course descriptions or module descriptions, into separate units, such as individual words or phrases.

As an illustration, the course description and module descriptions obtained from the database may be divided into lists of words or phrases using suitable tokenization algorithms.

3.1.12 Stop Words Removal:

Stop words are frequently used words that lack substantial significance and are frequently eliminated from text data to enhance the accuracy and efficiency of analysis.

Eliminating stop words guarantees that attention is directed towards phrases and keywords that are particular to the text.

Stop words can be eliminated from tokenized text data by using techniques such as NLTK's stop words corpus or bespoke stop word lists.

3.1.13 Punctuation Removal:

Punctuation marks, such as commas, periods, exclamation marks, and so on, are often insignificant for text analysis and can be eliminated to cleanse the data.

Eliminating punctuation marks streamlines future research and guarantees uniformity in textual data.

This may be accomplished by utilizing Python's string manipulation utilities or regular expressions to substitute or eliminate punctuation marks from the tokenized text input

3.1.14 Exploratory Data Evaluation (EDA):

In order to do exploratory data analysis (EDA) on the given EdTech platform data, we employed several Python modules that are typically utilized for EDA.

3.1.14.1 Data Retrieval and Overview:

We extracted the pertinent data from the MySQL database linked to the EdTech platform, encompassing course descriptions, module descriptions, keywords, and course information.

The data presented included fundamental statistics and information, such as the count of courses, modules, keywords, and so on. We analyzed the data's structure, encompassing data kinds, missing values, and unique values.

3.1.14.2 Data Analysis:

Analyzed data from course descriptions, module descriptions, and keywords to examine their distribution and features.

Summary statistics were computed for many variables, including the length of course descriptions and modules, average keyword weights, and so on.

We employed data visualization techniques such as histograms, word clouds, and bar plots to analyze the distribution of text data and find prevalent phrases, themes, or trends.

3.1.14.3 Keyword Analysis:

We conducted an analysis of the distribution of keywords and their corresponding weights in order to find significant phrases and subjects. The occurrence and diversity of keywords were analyzed across various courses and modules. Keyword distributions were shown through the use of histograms, word clouds, and scatter plots

3.1.14.4 Correlation Analysis:

An investigation was conducted to examine potential relationships between several factors, including course features, module descriptions, and keyword weights.

Correlation coefficients were computed or visualization techniques such as heatmaps were employed to ascertain correlations between variables.

Data cleaning and preprocessing involve the process of removing errors, inconsistencies, and irrelevant data from a dataset in order to prepare it for analysis.

Missing or inconsistent data were detected and resolved using data cleaning and preprocessing methods, such as imputation or elimination of outliers. Prior to doing further investigation, the accuracy and excellence of the data were confirmed.

3.1.15 Data Cleaning and Preprocessing:

Dynamic representations were created using interactive visualization tools like Plotly or Bokeh, enabling viewers to explore the data interactively.

3.1.16 Analysis and understanding:

The summary encapsulated the significant discoveries and understandings derived from the exploratory data analysis.

The data was analyzed to identify patterns, trends, and linkages, and from this analysis, interpretations and actionable suggestions were derived.

3.1.17 Libraries Used:

NumPy: For mathematical assessments and gathering control.

Pandas: For information control and appraisal.

Seaborn and Matplotlib: For information depiction.

Word Cloud: For making word hazes to imagine text information.

NLTK: For standard language dealing with tries, for example, tokenization and stop word launch.

Scikit-learn: For executing TF-IDF and Cosine similarity calculations.

MySQL Connector-Python and MySQL: For speaking with and investigating the MySQL information base.

Cup and Holder SQL Alchemy: For building the Jar application and talking with the SQL information base.

By driving raised exploratory information evaluation utilizing the alluded to libraries, we acquired colossal experiences into the
EdTech stage information, which can illuminate dynamic cycles, include refreshes, and generally stage upgrades.

TF-IDF (Term Frequency – Inverse Document Frequency) is an overall elaborate procedure in typical language dealing with (NLP) for checking on the importance of words in a report or corpus. In our EdTech stage, we finished TF-IDF to isolate the meaning of terms in course titles and depictions, helping with content evaluation and the improvement of a pursuit highlight. Here is a positive clarification of TF-IDF nearby the circumstance utilized:

3.1.18 TF-IDF Clarification:

Term Frequency (TF):

Term Frequency measures the frequency of a term (word) within a document relative to the total number of terms in that document. It indicates how often a particular term appears in a document.

The formula for calculating TF is:

TF = (Number of times term appears in a document) / (Total number of terms in the document)

TF assigns higher weights to terms that appear more frequently within a document.

Inverse Document Frequency (IDF):
Inverse Document Frequency measures the importance of a term across the entire corpus by penalizing terms that are too common. It reflects how unique or rare a term is across documents in the corpus.

The formula for calculating IDF is:

IDF = log_e(Total number of documents / Number of documents containing the term)

IDF assigns higher weights to terms that are rare across documents but occur frequently in specific documents.

TF-IDF Weight:

TF-IDF weight combines both TF and IDF to determine the significance of a term in the corpus. It indicates how important a term is to a document relative to the entire corpus.

The formula for calculating TF-IDF weight is simply the product of TF and IDF:

TF-IDF = TF * IDF

Higher TF-IDF weights indicate terms that are both frequent within a document and rare across the corpus, making them more relevant to the document.

3.II. Implementation in Code:

In our codebase, we used the TfidfVectorizer from the sklearn.feature_extraction.text module to figure TF-IDF scores for course titles, portrayals, and module depictions. The vectorizer consequently handles the assessment of both TF and IDF scores thinking about the information text information.
This is the means by which TF and not permanently set up in our code:
Term Frequnecy (TF): The TfidfVectorizer accordingly figures TF scores for each term thinking about its recurrent inside each narrative
Inverse Document Frequency (IDF): Proportionally, the vectorizer processes IDF scores for each term thinking about its unprecedented case across the corpus.
At long last, the TF-IDF weight for each term is gotten by growing its TF and IDF scores, making a weighted portrayal of terms that mirrors their significance inside the corpus.
By utilizing TF-IDF, we had the decision to see key terms and their importance in the EdTech stage dataset, connecting with sensible substance evaluation and the progress of a convincing pursue highlight.

IV. RESULT:

Through the use of TF-IDF and Cosine Similarity in our EdTech platform, we have witnessed substantial enhancements in content analysis and search capabilities. The following are the primary outcomes of our methodology:

Enhanced Content Analysis:
The utilization of TF-IDF enabled us to precisely evaluate the significance of words in course names and descriptions. By taking
into account both the frequency of terms inside documents (TF) and their relevance throughout the whole corpus (IDF), we may

determine the important terms and their significance in our dataset.
This enhanced our comprehension of the content patterns on our platform, allowing us to customize our offers and suggestions
to more effectively cater to the requirements of learners and instructors.

Improved Search Functionality:
The utilization of Cosine Similarity allowed us to accurately quantify the similarity between course descriptions and module
descriptions. By calculating the cosine similarity of document vectors, we may determine which courses and modules are most
pertinent to a particular search query.
This led to an enhanced search experience for users, enabling them to rapidly and effectively find relevant material.

Enhancing User Engagement and Satisfaction:
The incorporation of TF-IDF and Cosine Similarity algorithms resulted in increased user engagement and satisfaction on-
site. Users may effortlessly discover courses and modules that are in line with their interests and learning goals, resulting in
higher user retention and loyalty.
Customized suggestions derived from text analysis and search outcomes significantly improved the overall user experience
promoting a feeling of personalization and pertinence.

Potential improvements:
In the future, our intention is to investigate supplementary NLP approaches and algorithms in order to expand our content
analysis and search capability even further. This involves integrating sophisticated machine learning algorithms to enhance
comprehension and suggest instructional material more effectively.
In addition, our goal is to utilize user input and interaction data to consistently improve and optimize our search functionality,
guaranteeing its durability and ease of use in the long run.

V. CONCLUSION:

Our utilization of TF-IDF and Cosine Similarity algorithms has effectively improved the functioning and user experience of our EdTech platform. By doing a thorough study of the material using TF-IDF, we obtained significant insights into the significance of phrases in course names and descriptions. This allowed us to develop a better understanding of content patterns and customize our offers appropriately.
In addition, the use of Cosine Similarity has greatly enhanced the search functionality of our platform, enabling users to effortlessly find pertinent courses and modules. By quantifying the resemblance across document vectors, we offered consumers precise and tailored search outcomes, therefore amplifying user involvement and contentment.
In summary, the integration of TF-IDF and Cosine Similarity has established a strong basis for our platform's data-centric strategy for tailored learning and recommendation. In the future, we are dedicated to improving and perfecting these methods, as well as investigating other NLP algorithms, to consistently improve the user experience and stimulate innovation in the online education industry.

VI. REFERENCES

[1] Sintia, Sintia & Defit, Sarjon & Nurcahyo, Gunadi. (2021). Product Codefication Accuracy With Cosine Similarity And Weighted Term Frequency And Inverse Document Frequency (TF-IDF). Journal of Applied Engineering and Technological Science (JAETS). 2. 62-69. 10.37385/jaets.v2i2.210.

[2] Zhou, W., Lin, M., Xiao, M., & Wang, Z. (2020). Beyond the search bar: The value of search quality on e-commerce platforms. In International Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global (International Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global). Association for Information Systems.

[3] Zhou, Wei and Lin, Mingfeng and Xiao, Mo and Fang, Lu, Exploitation and Exploration: Improving Search Precision on E-commerce Platforms (January 7, 2021). Available at SSRN: https://ssrn.com/abstract=3762144 or http://dx.doi.org/10.2139/ssrn.3762144

[4] Chiny, Mohamed & Chihab, Marouane & Bencharef, Omar & Younes, Chihab. (2022). Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms. 15-20. 10.5220/0010727500003101.

[5] Dawei, W., Yuehwern. Y., Ventresca, M., 2020. Improving neighbor-based collaborative filtering by using a hybrid similarity measurement. Expert Systems with Applications

[6] Dias, M. B., Locher, D., Li, M., El-Deredy, W., Lisboa, P.J., 2008. The value of personalised recommender systems to e-business: a case study. In Proceedings of the 2008 ACM conference on Recommender systems (pp. 291– 294). ACM

[7] Guo. A. and Yang, T., 2016. Research and improvement of feature words weight based on TF-IDF algorithm. In 2016 IEEE Information Technology, Networking, Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms 19 Electronic and Automation Control Conference, pp. 415–419, Chongqing, China.

[8] Mehta, R., Gupta, S., 2021. Movie Recommendation Systems using Sentiment Analysis and Cosine Similarity. International Journal for Modern Trends in Science and Technology, 7(01): 16-22.

[9] Muthurasu, M., Rengaraj, N., Mohan, K. C., 2019. Movie Recommendation System using Term FrequencyInverse Document Frequency and Cosine Similarity Method. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-6S3.

[10] Yunxiang, L., Qi, X., Zhang, T., 2020. Research on Text Classification Method based on PTF-IDF and Cosine Similarity. Journal of Information and Communication Engineering: Volume 6 pp. 335-338 (Issue 1).

[11] Amrizal, V. (2018). Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim). Jurnal Teknik Informatika. https://doi.org/10.15408/jti.v11i2.8623

[12] Arazy, O. and Woo, C., 2007. Enhancing information retrieval through statistical natural language processign: A study of collocation indexing. Mis Quarterly, pp.525-546.