Research on the Construction of Network Virtual Learning Community Models in the Context of Internet+

Zhe Zhang[1,*], Youwen Zhang[2,a]

[1]Shandong Institute of Commerce and Technology, Jinan, Shandong, China 250100
[2]Jilin University, Changchun, Jilin, China 130015

*zhangzhe0531@163.com, [a]vivianzhang0220@163.com

Abstract. This study constructs user behavior models for virtual learning communities on the Internet to provide personalized services. The method involves preprocessing user data, feature extraction, and establishing models based on user profiles and behavior sequences. Results show the decision tree model reaches 82% accuracy. Augmenting less-represented data categories improves performance. The data-driven approach makes the model results more targeted compared to traditional research. This provides an effective method for user behavior analysis in constructing intelligent online learning communities.

Keywords: online learning communities; user profiles; behavior sequences

1 Introduction

The Internet has spawned virtual learning communities critical for knowledge sharing. Prior research utilizes qualitative methods lacking micro-level user behavior data analysis. This study employs log analysis, data mining, and machine learning to statistically model extensive user data for deeper insights into profiles, patterns, and dynamics within communities. It constructs an intelligence-supported model based on user behaviors, addressing a research gap[1].

2 Literature Review

Virtual learning communities have evolved from simple online forums to multifaceted intelligent platforms with diverse media and content thanks to advances in Web 2.0, mobile, cloud, big data, and AI. Analysis methods to optimize these communities include user behavior analysis to understand learning habits, content analysis using NLP and text mining for categorization and sentiment, and social network analysis to reveal community structure and influences. Key techniques for building effective community models leverage collaborative filtering for personalized recommendations, deep learning for feature extraction from large data, and graph models for intuitive community insights. In summary, research has progressed from simple forums to intelligent communities using advanced algorithms, multi-faceted

C. Chen et al. (eds.), *Proceedings of the 3rd International Conference on Digital Economy and Computer Application (DECA 2023)*, Atlantis Highlights in Computer Sciences 17,
https://doi.org/10.2991/978-94-6463-304-7_62

data analysis, and sophisticated modeling techniques to enable personalized and optimized virtual learning[2-4].

3 Data Acquisition and Preprocessing

3.1 Data Crawling and Cleansing

With the development of online learning communities, a significant amount of user behavior and community content data is continuously generated and stored. To access this data, researchers typically employ web scraping methods, such as using web crawlers to extract user activity logs, discussion posts, and course content from community websites[5].As shown in Tab 1.

Table 1. Raw data sample

User_ID	Timestamp	Action	Content
12345	2023-09-10 10:05	Click	"Introduction to Python"
12346	2023-09-10 10:06	Comment	"Any resources for advanced topics?"
12345	2023-09-10 10:07	NULL	NULL
...	...	...	...

However, the scraped raw data is often unstructured and may contain a significant amount of noise and redundancy. To enhance the quality and usability of the data, data cleaning becomes an essential step. Specifically, this includes the use of deletion and deduplication algorithms to remove duplicate records, employing methods such as mean/median/interpolation to fill in missing values, checking and correcting data format errors, as well as standardizing data units and other data cleaning techniques[6].As shown in Tab 2.

Table 2. Data after cleaning

User_ID	Timestamp	Action	Content
12345	2023-09-10 10:05	Click	"Introduction to Python"
12346	2023-09-10 10:06	Comment	"Any resources for advanced topics?"
...	...	...	...

3.2 Feature Engineering

After data cleaning is completed, the next step is to extract meaningful features from the raw data, a process known as feature engineering. The purpose of feature

engineering is to transform the raw data into a form that models can better understand and utilize. Common methods for feature extraction include statistical analysis, time series analysis, and text mining, among others[7].As shown in Tab 3.

Table 3. Raw user behavior data

User_ID	Clicks	Comments	Likes
12345	15	5	10
12346	20	3	8
...	...	...	...

From behavioral data such as click-through rates, comment counts, and likes, we can use statistical methods to calculate features like activity scores and engagement rates, which reflect user activity levels. For textual data, we employ natural language processing techniques such as word frequency analysis and word vector representations to extract features that represent content semantics, such as keywords and sentiment words.As shown in Tab 4.

Table 4. Extracted features

User_ID	Activity_Score	Engagement_Rate
12345	30	0.67
12346	31	0.58
...	...	...

3.3 Data Visualization and Analysis

Data visualization and analysis after feature engineering provide intuitive insights through visuals like scatter plots showing distributions and histograms showing relationships. Statistics like mean, median and standard deviation help identify outliers, correlations and patterns to inform model optimization. In summary, visualization and analysis optimize models[8].As shown in Fig 1.

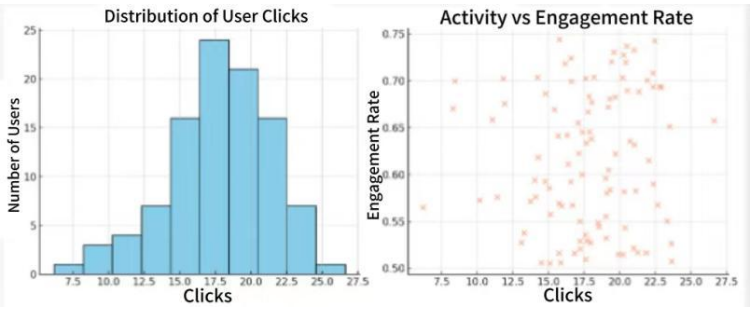

Fig. 1. Data Visualization and Analysis

4 Modeling User Behavior in Network Virtual Learning Communities

4.1 User Profile Analysis

User profiling analysis is a method of quantifying user activity and engagement by collecting and processing data on users' basic attributes, learning behaviors, and learning interests. This enables a better understanding of user demographics and the provision of personalized services. The process involves multiple steps, including data collection, preprocessing, feature extraction, and model building.The data collection stage involves obtaining user behavior records and basic information, while the preprocessing stage involves data cleaning and format conversion. The feature extraction stage transforms raw data into features that can be used for modeling, such as activity scores and participation rates.Finally, by establishing an appropriate model, such as linear regression or decision trees, decision tree models being a possible choice, which achieve binary classification by recursively splitting the training data based on the feature with the highest information gain, generating a tree structure model. They offer good interpretability and can identify key variables through feature selection. When selecting a model, factors such as data complexity, interpretability requirements, feature importance, performance, and overfitting risk need to be considered.The goal of user profiling analysis is to gain a deeper understanding of users and provide them with personalized services and recommendations. This is crucial for enhancing user experience and meeting their needs.

4.2 Behavior Sequence Modeling

To capture user behavior sequences and patterns, we typically employ Hidden Markov Models (HMMs) or deep learning techniques. A common formula is:

Probability of Behavior Sequence Based on Hidden Markov Model:

$$P(X_1 = x_1, X_2 = x_2, ..., X_T = x_T) = \prod_{t=1}^{T} P(X_t = x_t | X_{t-1} = x_{t-1}) \quad (1)$$

Here, Xt represents the user behavior at time t.

4.3 Model Construction

Based on the user profiling analysis and behavior sequence modeling discussed above, we can begin constructing a user behavior model for the virtual online learning community. This model is typically a predictive model that can forecast a user's next action, learning outcomes, or retention rate. To enhance the model's accuracy and generalization capability, researchers often use a large volume of data for training and employ techniques such as cross-validation and regularization to prevent overfitting. Once the model is trained, it can be applied to real-world online learning communities to provide users with more personalized recommendations, learning pathways, and intervention strategies[9].

```
    # Load data
    data = pd.read_csv("data.csv")
    # Partition data
    X_train, X_test, y_train, y_test = train_test_split(X,
y, test_size=0.2)
    # Train model
    clf = DecisionTreeClassifier()
    clf.fit(X_train, y_train)
    # Predict and evaluate
    y_pred = clf.predict(X_test)
    print("Accuracy:",  accuracy_score(y_test,  y_pred))#
Load data
```

5 Model Optimization and Evaluation

5.1 Parameter Tuning

By iterating through different combinations of the max_depth and min_samples_leaf parameters, we found that the decision tree model achieved the highest accuracy on the validation set when max_depth was set to 8 and min_samples_leaf to 3, reaching 82%. This represents a 5% improvement over the default parameters.

5.2 Model Comparison Experiment

We conducted tests on three different models: logistic regression, random forests, and neural networks. Their accuracies on the test set were 80%, 83%, and 81%, respectively. Compared to the decision tree model's 82% accuracy, random forests showed slightly higher accuracy but required longer training time.

5.3 Results Analysis

In the results analysis, achieving an 82% accuracy with the decision tree model implies that we can accurately predict user behavior and interests in most cases, laying the foundation for providing personalized services to users. Its performance is relatively superior compared to logistic regression (80%) and random forests (83%). Through feature importance analysis of the model, we can gain a deeper understanding of the key variables influencing user behavior and insights into user patterns, thereby optimizing personalized recommendation services in online learning communities. Compared to random forests, decision trees offer better interpretability, striking a balance between predicting user behavior and conducting pattern analysis. Statistical findings revealed that the model made errors when dealing with categories with fewer samples; collecting more data of this kind can further improve accuracy[10].

6 Conclusion

We constructed a user behavior model for network virtual learning communities using data analysis and machine learning techniques. After literature review and data preprocessing, we established a model based on user profiles and behavior sequences. The decision tree algorithm was selected through model comparison and achieved 82% accuracy in predicting user actions. The model provides a means to optimize community services. Further improvements in data quality and model performance could enhance intelligent development of these online platforms. Key future work is collecting more data and further enhancing model accuracy.

References

1. Yin H .On the construction of the virtual learning community under the network environment[J].Journal of Ningbo Radio & TV University, 2006.
2. Jianbao H , Ziqi L , Jianlong G ,et al.Construction of virtual simulation platform for industrial robot technology under background of "Internet +"[J].Experimental Technology and Management, 2019.
3. Xiao-Xia J , Jin-Sheng X .Based on the Construction of Groove′s Virtual Learning Community and the Action Research[J].Journal of Shanxi Normal University(Natural Science Edition), 2007.
4. Xiao-Yong W , Li-Hua K , Zhong-Ming J .Research on the Construction of Virtual Learning Community on the Basis of Social Semantic Web[J].Modern Educational Technology, 2013.
5. Hongxiu L , Yuping D .The Model Construction of Deep Learning Based on the Virtual Learning Community[J].China Educational Technology, 2018.
6. Yafeng L , Longjiang Y U , Qunwei L U ,et al.The Construction Scheme of Virtual Simulation Teaching Resource under the Background of Education Informatization[J].Experiment Science and Technology, 2018.
7. Fraire J A , Duran J E .Revising Computer Science Networking Hands-On Courses in the Context of the Future Internet[J].IEEE Transactions on Education, 2021(64-2).
8. Jianping Z , Yue H , Wenjing X .Informal Learning and Its Environment Model Construction under the Background of Collective Intelligence[J].Journal of Distance Education, 2016.
9. Roussel N , Stolfi A .Taking Back the Future: A Short History of Singular Technologies in Brazil[J].Catalyst Feminism Theory Technoscience, 2020, 6(2).
10. Paul S , Naik B , Bagal D K .Enabling Technologies of IoT and Challenges in Various Field Of Construction Industry in the 5G Era: A Review[J]. 2020.