The Journal of Supercomputing (2024) 80:6336–6355
https://doi.org/10.1007/s11227-023-05557-w

Personalized agricultural knowledge services: a framework for privacy-protected user portraits and efficient recommendation

Huarui Wu[1,2,3] **· Chang Liu**[1,2,3] **· Chunjiang Zhao**[1,2,3]

Accepted: 3 August 2023 / Published online: 24 October 2023

Abstract

In recent years, the increasing demand for knowledge services and the challenges of information overload have posed significant problems in delivering personalized and efficient agricultural knowledge services. This paper presents a comprehensive framework that addresses the issues of vague user positioning, serious privacy leakage, and low efficiency in personalized knowledge services within the national agricultural knowledge intelligent service cloud platform. The proposed framework utilizes privacy-protected user portraits based on generative adversarial nets (GAN) and leverages the TextCNN-LSTM algorithm for agricultural knowledge service prediction. By embedding labels into the algorithm and employing data obfuscation techniques, the framework achieves accurate inference of user behavior while preserving user privacy. Experimental results demonstrate the effectiveness and accuracy of the proposed framework, highlighting its potential for regional precise positioning and recommendation of personalized agricultural knowledge services. Experimental data shows that the average absolute error and root-mean-square error of this method are 1.1997 and 1.4143, respectively, and compared with MLP, TextCNN, and LSTM models, and it has higher prediction accuracy. In recent years, the increasing demand for knowledge services and the challenges of information overload have posed significant problems in delivering personalized and efficient agricultural knowledge services.

Keywords User portrait · Privacy protection · Knowledge service · Intelligent recommendation · Service matching

✉ Chunjiang Zhao
zhaocj@nercita.org.cn

[1] National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, People's Republic of China

[2] Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, People's Republic of China

[3] Key Laboratory of Digital Village Technology, Ministry of Agriculture and Rural Affairs, Beijing, People's Republic of China

1 Introduction

With the rapid development of information technology and the increasingly diverse needs of farmers, agricultural knowledge services are facing numerous challenges. In the current digital era, the Internet is flooded with a vast amount of agricultural information, leading to the problem of "information overload." Traditional knowledge service systems often struggle to efficiently filter and provide relevant information that meets the diverse, dynamic, and personalized agricultural needs of users [1]. This gap between the overwhelming amount of information and the users' specific requirements calls for the development of personalized agricultural knowledge services.

The significance of personalized agricultural knowledge services lies in their ability to address the limitations of traditional systems and provide tailored and targeted guidance to farmers and agricultural stakeholders. Personalization enables the delivery of relevant and context-specific information to users, taking into account factors such as location, climate, soil conditions, and individual preferences. By providing customized recommendations and advice, personalized knowledge services can assist farmers in making informed decisions and implementing practices that are most suitable for their specific circumstances.

To cope with the challenges posed by information overload and the need for personalization, researchers have proposed various recommendation methods. These methods include user-based collaborative filtering algorithms [2, 3], recommendation algorithms based on association rules [4], and tag-based recommendation algorithms [5, 6]. However, modeling large-scale data remains a challenge, and obtaining satisfactory results can be difficult.

In recent years, researchers have started exploring the application of user portrait technology in the field of agricultural knowledge services. User portrait refers to a tagged user model constructed through qualitative and quantitative analysis of collected user information and behavior data. By analyzing user needs and classifying user groups, personalized recommendations can be provided [7]. User portrait technology has the potential to address the issue of information overload by offering tailored information and services that align with users' specific requirements. Several studies have demonstrated the effectiveness of user portraits in agricultural knowledge services. For example, Zhang et al. [8] developed a recommendation model with multidimensional context integration, considering the individualized needs of agricultural users in different locations and timeframes. Chen et al. [9] studied user portrait and group dynamic user portrait technology, employing a hybrid recommendation algorithm to enable precise marketing on an agricultural product e-commerce platform. Wang et al. [10] focused on identifying the characteristics of farmers with different production performances and developed a farmer portrait model based on production performance segmentation. Zhang [11] studied precise marketing based on user portraits. By extracting user attributes and features, user portraits were constructed. Compared with BP neural network and SVM, the random forest algorithm has the highest prediction accuracy. Through the collection and mining of user behavior

status and other data, feature tags that best represent users are extracted to provide users with high-quality, personalized, and professional knowledge resource services [12]. These studies highlight the potential of user portraits in enhancing the personalization and effectiveness of agricultural knowledge services.

While user portraits offer promising solutions, the construction process often requires integrating a large amount of user information. This raises concerns about data privacy, particularly in the virtual and uncertain cloud space of service platforms. Excessive collection and access to private information such as location tracking, browsing content, and points of interest can lead to data loss and an increased risk of privacy breaches. Agricultural knowledge service platforms also face privacy issues and become potential targets for network attacks [13]. Thus, ensuring data security is crucial for improving the quality of agricultural knowledge services. Balancing the provision of high-quality knowledge services while protecting user privacy poses a significant challenge.

To address these issues, this study focuses on the agricultural personalized knowledge service method based on user portraits. By modeling user portraits, this method aims to accurately infer users' retrieval intentions and knowledge preferences. To tackle the problem of dynamic time decay of user interest, a user portrait model based on weight decay TF–IDF is constructed. Additionally, the TextCNN-LSTM algorithm is utilized to enhance the model's ability to extract temporal and spatial features from data. To mitigate the risk of user data privacy leakage, the GAN network is employed to confuse the embedded vector data. The effectiveness of the proposed method is demonstrated using the userdata set of the national agricultural knowledge intelligent service cloud platform.

The main contributions of this work are summarized as follows:

1. A user portrait knowledge service framework based on generative adversarial nets (GAN) privacy protection is proposed. This framework addresses the challenges of vague user positioning, serious privacy leakage, and low efficiency of personalized knowledge services in the national agricultural knowledge intelligent service cloud platform.
2. A user profile model based on weight decay TF–IDF is constructed to tackle the issue of dynamic time decay of user interest. This model captures users' evolving preferences over time, ensuring the relevance and accuracy of personalized recommendations.
3. The agricultural knowledge service prediction is performed by embedding labels into the TextCNN-LSTM algorithm. This integration of labels enhances the capability of the intelligent algorithm to predict users' ratings accurately. Moreover, the GAN network is employed to confuse the embedded vector data, thus improving the privacy protection ability of the algorithm.
4. Experimental results demonstrate that the proposed knowledge service framework achieves higher accuracy in predicting user ratings while ensuring user privacy.

The rest of this paper is organized as follows: Section 2 introduces the construction method of the user portrait labeling system. Section 3 presents the framework

for agricultural knowledge services based on user portraits. Section 4 presents the experimental setup. In Sect. 5, the influence of each parameter on the model's performance is discussed, and the prediction performance is evaluated through experiments. Finally, Sect. 6 provides an overview of the main conclusions of this study and outlines potential directions for future research.

2 Construction of user portrait label system

Based on the basic information and log records of registered users of the national agricultural knowledge intelligent service cloud platform, this paper divides the data that constitutes user portraits into static data and dynamic data. Names, genders, ages, and other basic user information that is relatively stable can be self-contained depending on the information entered when registering. Using dynamic context data, hobbies, and behaviors of users, create labels and weights. According to the data sources of user portraits, different attribute characteristics of users can be obtained, which can also reflect the specific agricultural information needs of users to a certain extent. Then, user portraits are comprehensively described from different dimensions according to knowledge service requirements [14]. The finer the dimension division, the more accurate the user characteristics can be described, and the higher the accuracy of the service.

Comprehensively considering the agricultural service scenarios, the actual needs of users, and combining the interactive behavior of users on the platform, the labeling system is constructed from five dimensions: basic information, context attributes, behavior attributes, hobbies, and user access activity. The specific information is shown in Table 1. All dimensional information constitutes the overall user portrait, reflecting the normal interaction behavior and information needs of users in the agricultural knowledge service platform. The national agricultural knowledge intelligent service cloud platform extracts data from sporadic and scattered information establishes real-time labels for data resources and realizes precise services on the platform more accurately and effectively based on user portraits.

Table 1 Agricultural user portrait label system

Datasources	Datacontent
Basic attributes	Name, gender, age, and occupation
Behavioral attributes	Browse, favorite, like, and comment
Situational attributes	Date, location
Hobby	Vegetables, grain, fruit trees, livestock and poultry
User activity	Low activity, high activity

3 An agricultural knowledge service framework based on user portraits

In traditional knowledge service systems, it is often challenging to accurately determine the specific needs and preferences of individual users. For instance, farmers seeking agricultural information may face difficulties in locating relevant and timely resources due to the vast amount of information available. The proposed framework aims to address this issue by constructing user portraits that capture various dimensions of user characteristics, including basic information, context attributes, behavior attributes, hobbies, and user access activity. This comprehensive user profiling enables more accurate positioning of user preferences and needs, resulting in personalized and targeted agricultural knowledge services. In order to solve the problem of accurate characterization of user behavior and accurate inference of knowledge preference, an agricultural knowledge service framework based on user portrait is proposed, as shown in Fig. 1. Firstly, the user portrait knowledge service framework incorporates generative adversarial nets (GAN) privacy protection, which aims to address the issues of vague user positioning, serious privacy leakage, and low efficiency of personalized knowledge services. This integration of privacy protection techniques is specifically tailored to the agricultural knowledge service platform, ensuring that user data remains secure and confidential. Secondly, the construction of the weight decay TF–IDF user profile model is designed to tackle the problem of dynamic time decay of user interest. By incorporating weight decay techniques into the TF–IDF model, we can capture the evolving interests and preferences of users

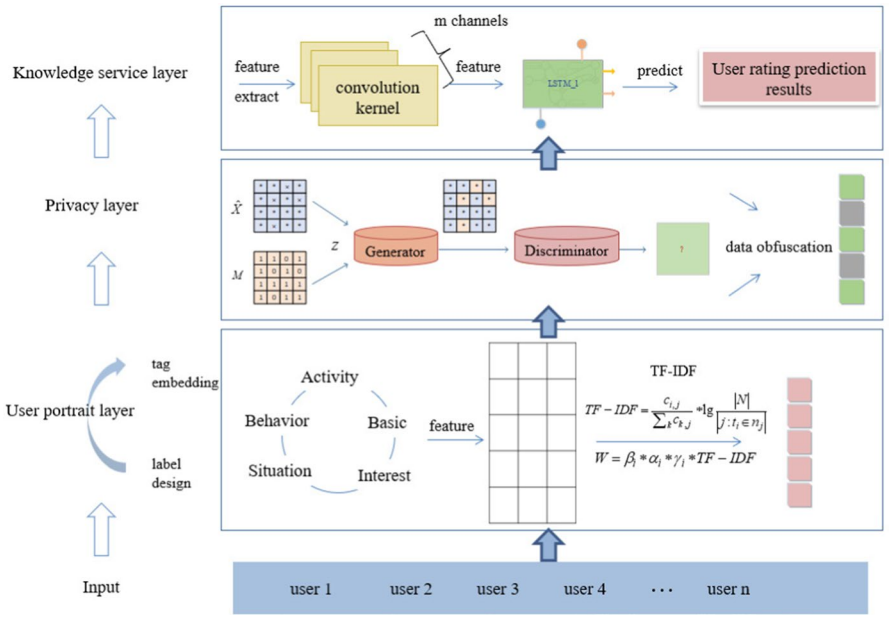


Fig. 1 Framework of agricultural knowledge service based on user portrait

over time, enhancing the accuracy and timeliness of our knowledge service predictions. Thirdly, the agricultural knowledge service prediction utilizes the TextCNN-LSTM algorithm with embedded labels. This combination allows us to leverage the strengths of both convolutional and recurrent neural networks, enabling accurate and context-aware predictions for agricultural knowledge services. The integration of the GAN network to confuse embedded vector data further enhances privacy protection, ensuring that sensitive user information is safeguarded. To clearly illustrate the entire structure, this framework uses a multi-layered notation consisting of three layers: user portrait construction, privacy protection, and knowledge service. Among them, the user portrait construction layer includes label design and label embedding.

3.1 TF–IDF user portrait construction layer based on weight decay

Term Frequency-Inverse Document Frequency (TF–IDF) is a statistical method used to evaluate the importance of feature words in a document set or corpus, and it is also a very effective feature extraction algorithm. With increasing number of occurrences in the document, the importance of the words decreases inversely with frequency. High-frequency words with little discrimination are filtered out by TF–IDF, while representative low-frequency words are retained. In the process of extracting user portrait tags, natural attribute information such as basic information and user categories are used as the basic features of user portrait tags. The levels of tags and user groups can be divided according to the basic features, and the tags will not change for a long time after they are established. The user's interest attributes such as behavioral attributes and situational attributes will change over time and have certain timeliness. Therefore, the time decay of interest tags needs to be considered when calculating tag weights. Based on user behavior weight, times, and time decay information, the following formula is proposed for calculating the weight of TF–IDF tags:

$$\mathrm{TF-IDF} = \frac{c_{i,j}}{\sum_k c_{k,j}} * \lg \frac{|N|}{|j : t_i \in n_j|} \tag{1}$$

$$W = \beta_i * \alpha_i * \gamma_i * \mathrm{TF-IDF} \tag{2}$$

In the formula, $c_{i,j}$represents the number of tags that user i appears in item j, $\sum_k c_{k,j}$ represents the number of tags that user i appears in all items, $|N|$ represents the number of tags that item j appears in all users, $|j : t_i \in n_j|$ represents the total number of all tags, and β_i represents the weight of user behavior, α_i represents the time decay coefficient, and γ_i represents the number of behavior tags. Combined with the actual situation, users' browsing, favorites, comments, and likes on the platform will become less and less relevant to the current reference as time goes by, so the weight will gradually decrease over time. Set according to Newton's law of cooling,$\alpha = e^{-0.1556t}$, where, represents the interval time. Behavior-type weight dimension tables are established according to the importance of the behavior. The weight

of browsing behavior is set to 0.3, the weight of collection behavior is set to 0.5, the weight of comment behavior is set to 1, and the weight of like behavior is set to 1.5.

3.2 GAN-based user privacy protection layer

Privacy protection is of utmost importance when handling user data, especially in the context of agricultural knowledge services. For example, when sensitive data such as user location, browsing content, and points of interest are excessively collected and accessed, there is an increased risk of data loss and privacy breaches. To mitigate this risk, incorporating privacy protection measures is proposed into our framework. To ensure that user data is safeguarded and protected from unauthorized access or disclosure, in the process of building user portraits, it is necessary to use user information as detailed as possible to improve the knowledge intelligent recommendation ability of the national agricultural knowledge service cloud platform. However, it is inevitable to encounter the contradiction between user information utilization and privacy protection [15].

Generative adversarial nets (GAN) are used to confuse embedded vector data, improve the model's ability to protect user privacy, and resolve the contradiction between data privacy protection and platform precision services [16]. The activity data generated by users in the process of surfing the Internet, such as browsing, favorites, comments, and likes, users often choose to disclose such data and submit it to service providers in exchange for high-quality personalized recommendation knowledge services. However, users of data such as gender, age, and occupation will regard it as their own private property and are unwilling to open it to the outside world. Although the user does not publish his private data, the connection between public data and private data usually leads to serious privacy leakage. Based on the GAN model, this paper generates time series data consistent with the distribution of the original data set for data obfuscation, so as to achieve the purpose of privacy protection. A generator, discriminator, and input noise comprise the basic structure, as shown in Fig. 2.

Assume that the original user data set is expressed as x,

$$x = [x_{i1}, ..., x_{ij}, ..., x_{\text{im}}] \in R^{m \times l},\ i \in [1,\ l],\ j \in [1,\ m]$$

where m is the number of basic user information, l is the length of the time series, and x_{ij} is the jth information value corresponding to the ith moment. Here, define a binary matrix T,

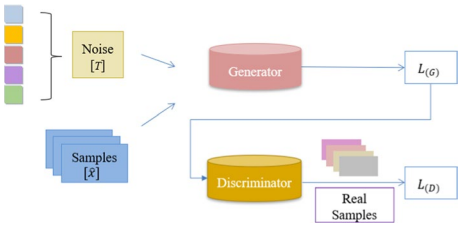


Fig. 2 Basic structure diagram of GAN

$$T = [T_{i1}, ..., T_{ij}, ..., T_{\text{im}}] \in R^{m \times n},\ i \in [1,\ n],\ j \in [1,\ m]$$

which is the same size as the original user data x and only consists of 0 and 1. Then, the reconstructed data $\hat{X}$ can be expressed as:

$$\hat{X} = \begin{cases} x_{i,j},\ T_{i,j} = 1 \\ \text{Nan},\ \text{else} \end{cases} \tag{3}$$

T can mean that a component of x is observed. Values of 1 in T represent observed true data, and values of 0 in T represent missing values in $\hat{X}$. By adjusting the number of 0 in T, construct samples with different ratios of user information protection. Then, the generator performs obfuscation processing according to the observed results, and the obfuscated vector is represented by $\hat{y}$:

$$\widehat{y} = G(\hat{X},\ T,\ (1 - T) \odot Z) \tag{4}$$

In the formula, $\hat{X}$ means missing processing data; T means a binary matrix with the same size as $\hat{X}$; Z means noise; $\odot$ means multiplication of corresponding elements.

The final output of the generator is a complete vector $\tilde{X}$ after confusion, and the formula is:

$$\tilde{X} = T \odot \hat{X} + (1 - T)\widehat{y} \tag{5}$$

Since some of the complete results output by the generator are real and some are generated, unlike the original GAN network, the discriminator here does not judge whether the entire vector is true or false, but tries to distinguish which are real and which are generated. Train D by maximizing the probability of correctly predicting T, and train G by minimizing the probability of D correctly predicting T. The objective function formula is:

$$\min_G \max_D V(G, D) = E_{\tilde{X},T}[T^T \log D(\tilde{X}) + (1 - T)^T \log(1 - D(\tilde{X}))] \tag{6}$$

The discriminator distinguishes the source of each part of the input data, and the obtained discriminant matrix is represented by $\widehat{T}$. In order to accurately judge each element in T, the cross-entropy loss function is used. The calculation formula is expressed as:

$$\mathcal{L}_D(t, \widehat{t}) = \sum_{i=1}^{d} [t_i log(\widehat{t_i}) + (1 - t_i) \log(1 - \widehat{t_i})] \tag{7}$$

For the estimation of the confused data as close as possible to the original data, the loss function formula of the generator is expressed as:

$$\mathcal{L}_G(t, \widehat{t}) = \sum_{i=1}^{d} [1 - t_i \log(\widehat{t_i})] + \alpha RMSE \tag{8}$$

3.3 Knowledge service layer based on TextCNN-LSTM word embedding

The knowledge service layer in our proposed framework utilizes the TextCNN-LSTM algorithm for predicting agricultural knowledge services, enabling accurate inference of user behavior. To achieve this, word embedding is first performed in the first convolutional layer, where each word is mapped into a word vector and then classified [17]. Considering that a word usually has different importance in documents with different class labels, the classification performance of TextCNN is improved by combining with word embedding, and then LSTM is used for time series prediction, as shown in Fig. 3.

First, according to the input text information and word vector dimensions, TextCNN is used to model the text and label vectors, and capture the semantic correlation among them. The text can be expressed as:

$$X = x_1 \oplus x_2 \oplus \cdots \oplus x_n \tag{9}$$

In the formula, $\oplus$ represents a cascade operation, and a document is represented as a matrix similar to an image matrix, which is convenient for using convolution operations to obtain local features. A convolutional layer consists of a convolution kernel that performs a convolution operation with input samples and the resulting feature map. The length of the convolution kernel is F_1, the width is the same as the input word vector dimension, and the number of convolution kernels is F_n. The convolution kernel performs convolution operation on the input text data along the direction of time series, with a step size of 1. To keep the convolved output in the same dimensionality, zero-padding is used. After the convolution operation, the Tanh activation function is used to increase the nonlinear processing capability of the model. The convolution operation formula is expressed as:

$$x_v = \varphi(W_v{}^T X_{i:i+N_{w-1}} + b_v), \quad 0 < v \le F_n, \quad v \in Z \tag{10}$$

In the formula, φ represents the nonlinear activation function; $X_{i:i+N_{w-1}}$ represents the input time series; W_v and b_v represent the weight and bias items of the convolution kernel, respectively.

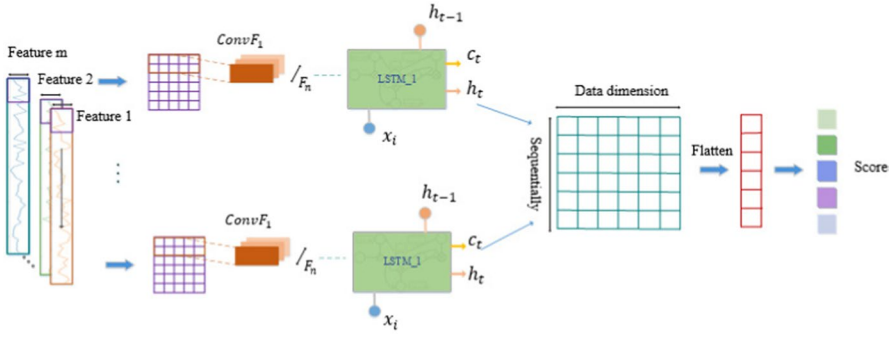


Fig. 3 TextCNN-LSTM structure diagram

The feature map obtained by the convolution kernel is expressed as:

$$x = \{\varphi(W_1{}^T X_{i:i+N_{w-1}} + b_1), \varphi(W_2{}^T X_{i:i+N_{w-1}} + b_2), ..., \varphi(W_{F_n}{}^T X_{i:i+N_{w-1}} + b_{F_n})\} \tag{11}$$

In the formula, W_1, W_2,..., W_{F_n} represent the weights from the 1st to the F_nth convolution kernel, $X_{i:i+N_{w-1}}$ represents the input sequence, and b_1, b_2,..., b_{F_n} represent the weights from the 1st to the F_nth volume, the offset of the product kernel.

Following the TextCNN component, the LSTM model learns the relationship between the long- and short-term patterns in the time series data. It consists of an input layer, a hidden layer, and an output layer, incorporating three gating units (input gate, forget gate, and output gate) and a memory unit. The input gate, forget gate, and output gate control the flow of historical information, and the memory unit retains important information from previous time steps. The LSTM model effectively captures temporal dependencies in the data, facilitating accurate predictions [18, 19].

Finally, the TextCNN-LSTM model produces a score output, which is converted into a single unit using the Flatten operation. This step transforms the multidimensional data processed by the model into a one-dimensional representation.

4 Experimental setup

4.1 Experimental data set description

The data comes from the user log records of the national agricultural knowledge intelligent service cloud platform. Data set components include user id, project id, user location, gender, occupation, age, behavior type, and behavior frequency, and each records a user's scoring data for the project. The score is divided into 5 levels, with 1 being the lowest and 5 being the highest. The training set is used for training the model of knowledge service, and the test set is used to verify its performance. In the test set, the user's rating of the item is hidden, and only basic information, behavior, and basic information are retained. See Tables 2 and 3 for the data distribution of the training set and test set, and see Table 4 for the detailed description of each field.

Table 2 Distribution of training set

	User_id	Item_id	Area	Sex	Profession	Age	Behavior_type	Behavior_count	Score
Min	0	0	1	1	1	1	1	1	1
Max	221000	14000	8	2	8	7	4	9	5

Table 3 Distribution of test set

	User_id	Item_id	Area	Sex	Profession	Age	Behavior_type	Behavior_count
Min	0	0	1	1	1	1	1	1
Max	221000	14000	8	2	8	7	4	9

4.2 Experimental settings and evaluation metrics

The hardware environment of this experiment is Intel(R) Core(TM) i7-10700F CPU, 32 GB memory, and NVIDIA GTX1660 graphics card. A Windows 10 operating system is used in the software environment. The programming environment is the Tensorflow framework, Python 3.6. RMSE and MAE are used as evaluation indicators to verify the effectiveness of the method proposed in this paper, and the formula is as follows [20]:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |d_i|}{n} \tag{12}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2} \tag{13}$$

In the formula, d_i represents the difference between the predicted value and the actual value. MAE and RMSE measure the deviation of the prediction results. The smaller the value, the higher the prediction accuracy and the better the recommendation algorithm.

5 Experimental results and analysis

5.1 Analysis of TF–IDF user portrait results based on weight decay

To assess the impact of promoting agricultural knowledge services and determine whether it leads to an increase in users, we need to first understand the daily active user count on the platform and track changes in user growth over time. In this paper, user activity labels are divided into two dimensions: high-active users and low-active users. The platform provides users with agricultural knowledge services, which are mainly reflected in the form of questions and answers. Based on activity information analysis of user comments, Fig. 4 shows that the number of comments and the number of comments within 30 days have a positive relationship.

According to Fig. 4, the number of platform users participating in comments gradually decreases as the number of comments increases, and an inflection point appears 15 times. Therefore, from the perspective of the distribution of user comment behavior, users who comment more than 15 times are defined as high-active users, and those who comment less than or equal to 15 times are defined as

Table 4 Dataset field description

Attributes	Field range	Content description
Area	"1–8"	1: "Northeast Region"; 2: "North China Region"; 3: "Central China Region"; 4: "East China Region"; 5: "South China Region"; 6: "Northwest Region"; 7: "Southwest Region"; 8: "Hong Kong, Macao and Taiwan regions"
Gender	"1–2"	1: "Male"; 2: "Female"
Profession	"1–8"	1: "Fruit Farmer"; 2: "Vegetable Farmer"; 3: "Cotton Farmer"; 4: "Forest Farmer"; 5: "Student"; 6: "Agricultural Technician"; 7: "Businessman"; 8: "Others"
Age	"1–7"	1: "Under 18"; 2: "18–24"; 3: "25–34"; 4: "35–44"; 5: "45–54"; 6: "55–64"; 7: "65+"
Behavior	"1–4"	1: "Browse"; 2: "Favorite"; 3: "Comment"; 4: "Like"

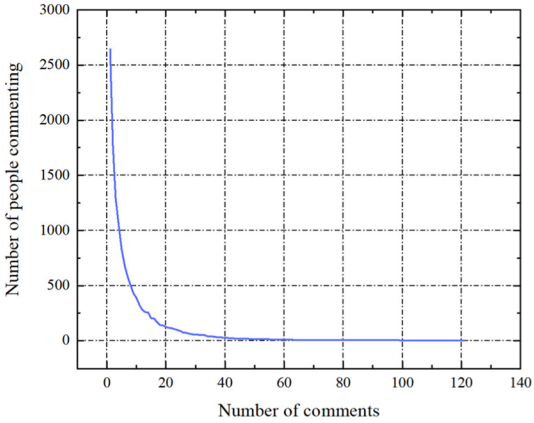


Fig. 4 Relationship between the number of comments and the number of comments within 30 days

low-active users. Through the analysis of user activity, the platform can take corresponding measures and optimize policies for different users to reduce user loss. For users who are not active, optimize and adjust the webpage according to the functions and paths of their access to the site, so as to increase the attractiveness of the system to users.

As shown in Table 5, some label weight results are obtained according to the TF–IDF label weight method based on weight decay using user behavior categories, behavior weights, time decay factors, and behavior times.

In Fig. 5, an outline of the national agricultural knowledge intelligent service cloud platform is presented. The platform leverages user data, including behavior categories, behavior weights, time decay factors, behavior times, and tag weight results obtained through the TF–IDF label weight method based on weight decay. The platform utilizes these data to build user portraits and enhance its precision and personalized service capabilities. One of the key features of the platform is the visualization and analysis of user data. Numerical user data, such as monthly visit frequency and fixed-time visit volume, are analyzed and represented using various visualization techniques, including pie charts, column charts, and bar charts. These visualizations provide valuable insights into user behavior patterns and preferences, enabling a deeper understanding of user needs. By visualizing and analyzing user data, the national agricultural knowledge intelligent service cloud platform

Table 5 Partial results of label weights

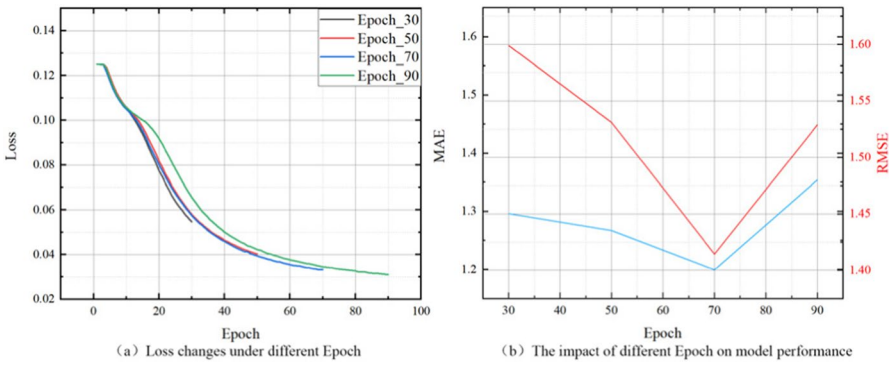
User_id	Item_id	Behavior_type	Behavior_count	tfidf_ratio	Act_weight_plan	Time_reduce_ratio	Act_weight
492	310	1	1	0.149055	0.3	0.5366	0.0239
492	399	1	1	0.159331	0.3	0.0174	0.0008
492	508	1	1	0.178222	0.3	0.2109	0.0112
492	652	1	2	0.178222	0.3	0.2109	0.0225
492	310	1	1	0.149055	0.3	0.5366	0.0239

Fig. 5 Schematic diagram of user portrait application system

facilitates effective decision-making and resource allocation in the agriculture domain. It allows agricultural service providers to gain a comprehensive understanding of user interests, behavior patterns, and information needs. This knowledge can be used to optimize knowledge service delivery, tailor recommendations to individual users, and improve the overall user experience on the platform.

5.2 Results of personalized service of agricultural knowledge

In order to verify the effectiveness of the method proposed in this paper, the model Epoch parameters were analyzed first, and the experimental results are shown in Fig. 6.



(a) Loss changes under different Epoch

(b) The impact of different Epoch on model performance

Fig. 6 Influence of Epoch on the experimental results

Figure 6a shows that as the number of iterations increases, the loss value gradually decreases and tends to be stable at about 70 times. The model effect is evaluated according to Eqs. 12 and 13, and the changes of MAE and RMSE with Epoch have plotted relationship, as shown in Fig. 6b. With the increase in Epoch, the prediction effect of the model fluctuates. When the Epoch is small, it brings more randomness and it is difficult to converge, thus reducing the prediction accuracy. As the number of iterations increases, however, the model will overfit and increase a lot of training time, which will reduce prediction accuracy. A larger number of epochs makes gradient descent more accurate and reduces the occurrence of training fluctuations. It can be seen from the figure that for the proposed method, the Epoch size is set to 70, and the two evaluation indicators get the optimal value.

Then, in order to explore the performance of the model, it is compared with the methods commonly used in recent years. A comparison of multi-layer perceptrons (MLP), TextCNN, LSTM, and TextCNN-LSTM models is shown in Fig. 7.

Figure 7 shows the comparison curves of the experimental results of the four models in the test data set and two evaluation indicators. The abscissa indicates the different models selected, and the ordinate indicates the evaluation index value. According to the line chart, MLP and LSTM have similar results for both evaluation indicators, and their performance is lower than that of TextCNN. Among the four models, TextCNN-LSTM achieved the best prediction effect. On the one hand, the TextCNN model was used to extract detailed features of the spatial scale, and on the other hand, the LSTM model was used to fit the temporal and nonlinear relationships of complex multidimensional data. Additionally, fusion technology can effectively improve prediction accuracy and solve the problem of high data dimensionality and long-term dependence on time series, which supports the notion that fusion technology has great potential in the field of recommendation.

In addition, in order to verify the GAN-based privacy protection methods proposed in this paper, the gender, age, occupation, region, and other user data of the national agricultural knowledge intelligent service cloud platform was obfuscated. Taking RMSE as the evaluation index, discuss the data confusion effect of GAN, where Eq. 13 represents the error value between the confusion value and the real value.

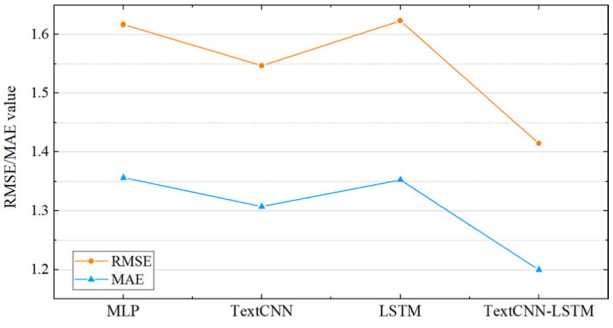


Fig. 7 Comparison results of the prediction accuracy of the four models on the test set

In this paper, "confusion rate" is used to denote the degree of obfuscation or obfuscation applied to the embedded vector data using a GAN network. The confusion rate represents the extent to which the original user data is obfuscated or made less distinguishable to protect user privacy. A higher confusion rate indicates a higher level of obfuscation, leading to a greater degree of data privacy protection. On the test set, the confusion rates are set to 0.1−0.9, respectively. Under different confusion rates, the GAN prediction results are shown in Fig. 8.

Figure 8 shows that as the confusion rate increases, the RMSE value of the data set increases, and accuracy slowly decreases. When the confusion rate is greater than 0.6 and the data are seriously lacking, the sample information is less, and it is difficult to restore the real sample data. Therefore, in order to solve the contradiction between data privacy protection and precise platform services, the confusion rate should not exceed 0.6 under the condition of ensuring the accuracy of user data.

Finally, under different confusion rates, the prediction results of the proposed prediction method on the test set are verified. The simulation experiment results are shown in Table 6, and the corresponding curves are shown in Fig. 9.

It can be seen from Fig. 9 that the overall trends of the four models in the two evaluation indicators are basically the same. The performance of the TextCNN-LSTM model is the best, followed by the TextCNN model. When the mixing efficiency is lower than 0.3, the performance of LSTM and MLP is similar, and the mixing efficiency is above 0.3, and LSTM outperforms MLP. When the confusion rate is 0, it means the user's real data set. As the confusion rate increases, the prediction accuracy of the models gradually decreases. When the confusion rate is below 0.4, the predicted values of the models closely resemble the results of the unconfused data, with the proposed prediction model showing better performance. However, when the confusion rate exceeds 0.4, the data information loss becomes substantial, posing challenges for accurate prediction. The results highlight that agricultural user portraits can effectively deliver personalized agricultural information to

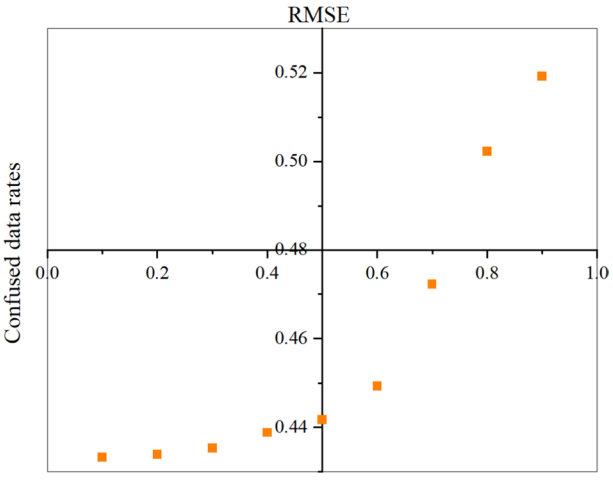


Fig. 8 RMSE value of the paired data under different confusion rate processing

Table 6 Performance comparison of each model under different confusion rates

Model	Evaluation index	Confusion rate									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MLP	MAE	1.3565	1.3571	1.3597	1.3982	1.3989	1.4097	1.4208	1.4309	1.4495	1.4955
	RMSE	1.6169	1.6177	1.6387	1.6389	1.7201	1.7205	1.7624	1.7729	1.8242	1.831
TextCNN	MAE	1.3071	1.3084	1.3156	1.3301	1.3325	1.3401	1.3507	1.3507	1.3599	1.3997
	RMSE	1.5469	1.5473	1.5577	1.5632	1.5894	1.5901	1.6105	1.6322	1.6923	1.7215
LSTM	MAE	1.3529	1.3555	1.3597	1.3599	1.3618	1.3742	1.4024	1.4309	1.4495	1.4711
	RMSE	1.6236	1.6277	1.6387	1.6389	1.7001	1.7014	1.7124	1.7529	1.7532	1.7931
TextCNN-LSTM	MAE	1.1997	1.1998	1.2087	1.2287	1.2496	1.3089	1.3397	1.3507	1.3599	1.3615
	RMSE	1.4143	1.4143	1.4587	1.4687	1.5181	1.5801	1.6005	1.622	1.6223	1.6242

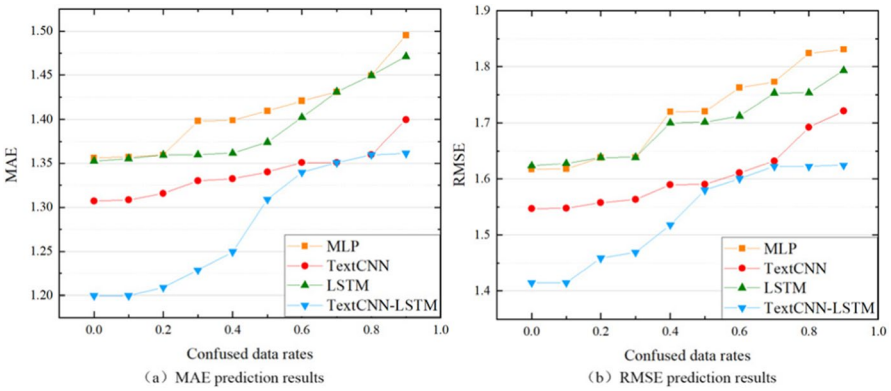


Fig. 9 Performance comparison of each model under different confusion rates

practitioners, providing them with timely and accurate agricultural knowledge. The agricultural information personalized recommendation system leverages user behavior to update user portraits regularly, ensuring that the user attributes and interests are accurately reflected. By collecting and analyzing user behavior, the agricultural knowledge intelligent service cloud platform overcomes constraints of time and space, allowing users to access and publish content anytime and anywhere.

6 Conclusions

Agricultural knowledge service frameworks based on user portraits were proposed to address problems of inadequate descriptions of knowledge information for specific needs of users, privacy leakage, and low accuracy of personalized recommendations of agricultural information resources. In order to realize the accurate inference of the user's retrieval intention and knowledge preference, the TF–IDF label calculation method with weight decay is adopted in the process of user portrait construction. In the process of user preference prediction, TextCNN-LSTM is used for feature extraction and prediction. At the same time, GAN data obfuscation is used for encryption processing to reduce the risk of user privacy leakage. Experiments were carried out using the data set established by the national agricultural knowledge intelligent service cloud platform. Compared with MLP, TextCNN, and LSTM, the model proposed in this paper demonstrates higher prediction accuracy, affirming the utility of the method introduced in this paper for practical agricultural knowledge services. It has laid a foundation for the users of the national agricultural knowledge intelligent service cloud platform to provide safe and accurate push and personalized services in different scenarios.

In future work, we will further consider using regional characteristic crops combined with platform user-selected interest content to jointly form new user initial interest tags as the initial user portrait of new users in the recommendation system. Due to the lack of new user data, the recommendation system usually faces the

cold start problem when performing personalized user services. Due to the strong regional characteristics of crops, and in the absence of new user behavior data, agricultural production is determined according to the user's location. Based on regional characteristics, a group user portrait is established to solve the cold start problem of the recommendation system.

Acknowledgements This work was supported by the Science and Technology Innovation 2030—"New Generation Artificial Intelligence" Major Project (2021ZD0113604), subject name: National Agricultural Knowledge Intelligent Service Cloud Platform Technology Integration and R &D, and China Agriculture Research System of MOF and MARA Grant CARS-23-D07.

Author contributions HW involved in resources, investigation, and writing—review and editing. CL involved in methodology, conceptualization, validation, formal analysis, visualization, software, and writing—original draft. CZ involved in resources, investigation, and writing—review and editing.

Funding Supported by the Science and Technology Innovation 2030—"New Generation Artificial Intelligence" Major Project (2021ZD0113604), the China Agriculture Research System of MOF and MARA Grant CARS-23-D07. Chunjiang Zhao is the corresponding author of this paper.

Data availability The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the privacy policy of the authors' institution.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Da'u A, Salim N (2020) Recommendation system based on deep learning methods: a systematic review and new directions. Artif Intell Rev 53(4):2709–2748
2. Jiang L, Cheng Y, Yang L, Li J, Yan H, Wang X (2019) A trust-based collaborative filtering algorithm for e-commerce recommendation system. J Ambient Intell Humaniz Comput 10(8):3023–3034
3. Alhijawi B, Kilani Y (2020) A collaborative filtering recommender system using genetic algorithm. Inform Proc Manage 57(6):102310
4. Xiao J, Wang M, Jiang B, Li J (2018) A personalized recommendation system with combinational algorithm for online learning. J Ambient Intell Humaniz Comput 9(3):667–677
5. Zhao J, Zhang Q, Sun Q, Huo H, Xiao Y, Gong M (2021) Folkrank++: an optimization of Folkrank tag recommendation algorithm integrating user and item information. KSII Trans Internet Inform Syst TIIS 15(1):1–19
6. Wu Y, Xi S, Yao Y, Xu F, Tong H, Lu J (2018) Guiding supervised topic modeling for content based tag recommendation. Neurocomputing 314:479–489

7. Labaj M, Bieliková M (2013) Tabbed browsing behavior as a source for user modeling. In: User Modeling, Adaptation, and Personalization: 21th International Conference, UMAP 2013, Rome, Italy, June 10-14, 2013 Proceedings 21, pp 388–391. Springer
8. Zhang H, Qin X, Zheng H (2020) Research on contextual recommendation system of agricultural science and technology resource based on user portrait. In: Journal of Physics: Conference Series, vol 1693, p 012186. IOP Publishing
9. Xiao C, Xinfei C (2022) Research on the precise marketing method of agricultural products e-commerce platform based on user recommendation algorithm. In: 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), pp 519–522. IEEE
10. Wang B, Shi Y, Mu W, Feng J Modeling of farmers' production performance portrait based on gasa-wfcm clustering. Available at SSRN 4196752
11. Zhang M (2022) Research on precision marketing based on consumer portrait from the perspective of machine learning. Wireless Commun Mobile Comput, 2022
12. Yao W, Hou Q, Wang J, Lin H, Li X, Wang X (2019) A personalized recommendation system based on user portrait. In: Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, pp 341–347
13. Huang W, Liu B, Tang H (2019) Privacy protection for recommendation system: a survey. In: Journal of Physics: Conference Series, vol 1325, p 012087. IOP Publishing
14. Gu H, Wang J, Wang Z, Zhuang B, Su F (2018) Modeling of user portrait through social media. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp 1–6. IEEE
15. Miao R, Li B (2022) A user-portraits-based recommendation algorithm for traditional short video industry and security management of user privacy in social networks. Technol Forecast Soc Chang 185:122103
16. Yang P, Gui X, Tian F, Yao J, Lin J (2013) A privacy-preserving data obfuscation scheme used in data statistics and data mining. In: 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing, pp 881–887. IEEE
17. Guo B, Zhang C, Liu J, Ma X (2019) Improving text classification with weighted word embeddings via a multi-channel textcnn model. Neurocomputing 363:366–374
18. Yu Y, Si X, Hu C, Zhang J (2019) A review of recurrent neural networks: Lstm cells and network architectures. Neural Comput 31(7):1235–1270
19. Gers FA, Schmidhuber J, Cummins F (2000) Learning to forget: continual prediction with LSTM. Neural Comput 12(10):2451–2471
20. Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE). Geosci Model Dev Discuss 7(1):1525–1534