Research on Fund Product Recommendation Based on Investor Profiles

Kui Fu and Zhilin Li(✉)

School of Economics, Wuhan University of Technology, Wuhan, Hubei, China
fukui@whut.edu.cn, 642723324@qq.com

Abstract. In the field of fund product recommendation research, there has been limited study on investor profiling and personalized recommendation, overlooking the significant value of investors' multidimensional characteristics in influencing fund product selection. To address the issue of individual investors' fund product selection, this study proposes a fund product recommendation model based on investor profiles. Real data of investors from Eastmoney.com, a popular mutual fund investment platform in China, was collected through web crawling for experimentation, validating the practicality and effectiveness of the mutual fund product recommendation system based on investor profiling.

Keywords: Investor Profiles · Collaborative Filtering · Fund Product Recommendation

1 Introduction

In the field of financial product marketing, investor profiling can provide a characterization of target users and identify their needs, enabling more accurate support for personalized services or assisting in applications such as credit assessment in the internet finance industry [1]. For instance, Dong Xinxin et al. utilized the K-Means clustering algorithm to mine user profiles and clustered user profiles with service profiles, resulting in an improved user response rate for pension service recommendations [2]. Additionally, Zhao Ming et al. constructed a three-dimensional commercial bank fund customer classification model, achieving precise fund customer marketing. These studies have provided different design approaches for investor profiling [3]. However, existing research on investors largely focuses on single dimensions such as basic data [4], investment behavior [5], risk preference [7], and social behavior [6].

In the field of financial product recommendation, Gan Qiang (2015) proposed a hybrid collaborative filtering algorithm and content-based clustering analysis algorithm. This algorithm process can construct user interest models for P2P online lending products and make product recommendations [8]. Zhou Ying (2014) has also made certain achievements in researching internet financial product recommendation systems. She used machine learning algorithms for clustering analysis of users and employed various recommendation algorithms, including collaborative filtering, heat conduction, and hybrid diffusion algorithms, for financial product recommendations [9]. Overall, research in the academic community on financial product recommendations is mostly

R. Lotfi et al. (Eds.): UPRE 2023, AEBMR 253, pp. 398–408, 2023.
https://doi.org/10.2991/978-94-6463-218-7_45

based on traditional collaborative filtering algorithms, and this study is no exception, as it chooses to improve the user-based collaborative filtering algorithm based on investor profiling.

In conclusion, this study will profile investors from multiple dimensions, including demographic information, behavioral characteristics, and preference characteristics, and based on this, improve the user-based collaborative filtering algorithm to construct a fund product recommendation model based on investor profiles.

2 Modeling of Fund Investors' Profiles

2.1 Indicators System for Fund Investors' Profiles

Characterizing investor profiles involves assigning labels to them. These labels are appropriate identifiers for investors with different characteristics based on specific business needs. Drawing on previous research, the influence factors of fund product attributes and investor behavioral preferences, three primary indicators, four secondary indicators, and 20 tertiary indicators were selected from three dimensions of investor demographic characteristics, behavioral characteristics, and preference characteristics as index system, as shown in Table 1.

2.2 Modeling Investor Profiles Based on Vector Space

Based on the indicator system established in the previous section, this section will mathematically model investors using a binary tuple, as shown below, where Info represents the vector of investor basic attributes, and Fundpre represents the vector of investor product preferences:

$$\text{Investor Persona} = < \text{Info}, \text{Fundpre} > \tag{1}$$

(1) Investor Base Attribute Model

Info = <Sex, Salary, Occuption, Education, Age, Address>, respectively representing gender, annual salary, occupation, education level, age and region.

Investor's basic attribute labels generally do not change dynamically over time, and their weights also remain constant. In this study, the Analytic Hierarchy Process (AHP) is used to assign weights to these labels, as shown in Table 2. The assigned weights will be used in Chapter 4 to calculate investor similarity.

(2) Investor Product Preference Model

The product preference vector is the core of the fund product recommendation service based on investor profiles, and its main task is to analyze investor behavior data. In this paper, the product preference of investors is expressed and modeled in the form of a "label-weight" list. The investor product preference model based on the vector space model can be represented as a collection of binary tuples in the form of a label-weight list, with the following structure:

$$\text{Fundpre} = ((k_1, w_1), (k_2, w_2), \ldots (k_n, w_n)) \tag{2}$$

Table 1. Fund Investor Profile Indicator System

Primary Indicators	Secondary indicators	Tertiary Indicators
A1 Demographic Features	A11 Basic Attributes	A111 Annual Income
		A112 Occupation
		A113 Education Level
		A114 Age
		A115 Region
		A116 Gender
A2 Behavioral characteristics	A21 Product interaction	A211 Monitoring
		A212 Favorites
		A213 Holdings
		A214 Reviews
	A22 Social behavior	A221 Number of Followings
		A222 Number of Followers
		A223 Length of Participation
A3Feature of preference	A31 Product preference	A311 Fund Variety
		A312 Fund Risk
		A313 Fund Rating
		A314 Fund Company
		A315 Fund Manager
		A316 Fund Theme
		A317 Investment style

Table 2. Investor base attribute weight results

	A111	A112	A113	A114	A115	A116
Weights	0.3281	0.2571	0.1732	0.1090	0.0793	0.0533

where ki represents the preference label, wi represents the weight of the preference label, and n represents the number of preference labels. The TF-IDF method is used for weight calculation. TF-IDF uses statistical methods to evaluate the importance of a certain feature in a collection of items. It is a weighted technique, and in this paper, it is applied to calculate the weights of investor preferences. If investor j has commented on a product label more frequently in their product collection Cj, and this product label accounts for a smaller proportion in the entire product collection's labels, then it is considered that the investor has a higher preference for this product label. In this paper, TFi,j refers to the frequency of the product label ki appearing in the investor's commented

product collection Cj. A larger TFi,j indicates that the investor has a greater preference for that product label, and that the product label can better represent the investor's preference. IDFi,j refers to the frequency of the product label ki appearing in the entire product collection Cj. A larger IDFi,j indicates that the product label ki contributes less to differentiating investor preferences, which helps eliminate interference from popular product preference labels in differentiating investors.

Therefore, in this paper, the weight preference w_i of the product preference label k_i for investors is calculated using formula (3):

$$w_{i,j} = TF_{i,j} \times IDF_i \tag{3}$$

$$IDF_i = \log \frac{N}{n_i} \tag{4}$$

$$TF_{i,j} = \frac{f_{i,j}}{f_{c,j}} \tag{5}$$

where N is the number of products in the product set C, ni is the number of times label i appears in N, and thus the IDF of k_i can be represented by Eq. (4); $f_{i,j}$ represents the frequency of label k_i appearing in the product set C_j, f_{Cj} represents the total number of labels appearing in the product set C_j, and $TF_{i,j}$, as shown in Eq. (5), represents the term frequency of k_i in the product set C_j. Therefore, $U = ((k_1,w_1),(k_2,w_2),...(k_n,w_n))$ represents the vector space features of investors, and similarly, the fund product space vector model can be constructed using $P = ((K_1,W_1),(K_2,W_2),...(K_n,W_n))$.

2.3 Dynamic Investor Profile

Based on time series theory, this paper proposes a formula for calculating tag weights in a vector model based on investor product preference. By combining behavioral data of the same tag from different time periods and using an exponential smoothing model to predict the tag weight value at the next moment, the purpose of updating investor profiles is achieved. The specific calculation formula is as follows:

$$F_{t+1} = \frac{1}{T}(Y_t - Y_{t-T}) + F_t \tag{6}$$

F_{t+1} represents the predicted value, which is an estimation of Y_{t-T} at time F_t. Substituting F_t into the equation, we have:

$$\begin{aligned} F_{t+1} &= \frac{Y_t}{T} + \frac{F_t}{T} + F_t \\ F_{t+1} &= \frac{Y_t}{T} + (1 - \frac{F_t}{T}) \end{aligned} \tag{7}$$

Let $\alpha = \frac{1}{T}$ smoothing constant, take into the upper formula:

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t \tag{8}$$

From the above equation, it can be observed that with the availability of the observation and forecast values from the previous time step, it is possible to make predictions for the next time step without relying on all historical data. This method utilizes exponential smoothing forecast technique to predict the weights of indicators for the next time step. By adjusting the weights of indicators, the current weight of the label can be obtained [13].

3 Recommendation Model Based on Investor Profiling

3.1 Recommendation Framework Based on Investor Profiling

The overall framework of the fund product recommendation system based on investor profiling is shown in Fig. 1.

The framework of the mutual fund product recommendation system based on investor profiling includes three main components: 1) Designing an investor profiling indicator system based on investor basic data and behavioral data; 2) Investor profiling modeling; 3) Mutual fund product recommendation.

Phase 1: Quantify investor reviews, calculate the similarity of investor ratings, and combine it with similarity weighted by the investor profiling model to form a comprehensive investor similarity. Based on the nearest neighbor investors, preliminary screening of mutual fund products that target investors may be interested in. Through predicted ratings, filter and generate a candidate list of top 2N mutual fund product recommendations.

Phase 2: The top 2N mutual fund product recommendations generated in the first phase are re-ranked using a recommendation approach based on investor profiling. Based on the similarity between the mutual fund product attribute vector and the investor product preference vector, mutual fund products with higher matching scores are recommended. The core content of this phase is to predict the label weights based on time series for the investor's historical preference labels, calculate the cosine similarity between the

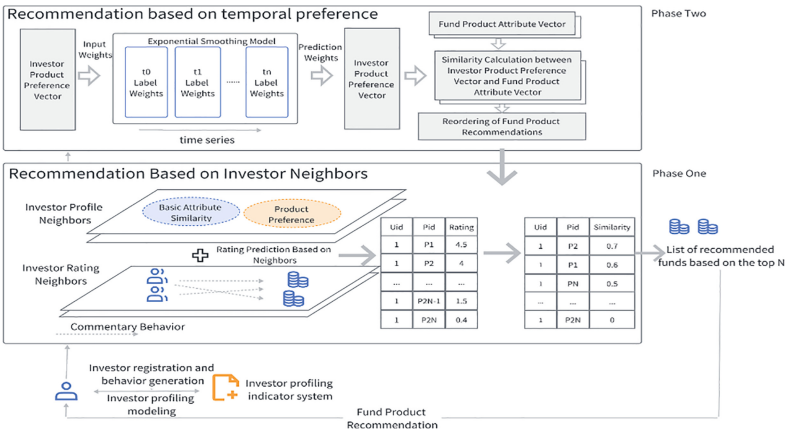

Fig. 1. Fund Product Recommendation Framework based on Investor Profiling.

updated investor preference and mutual fund product attributes, and filter the candidate mutual fund product recommendation list for the second time. This generates the final TopN mutual fund product recommendations for the target investor.

3.2 Investor Profiling-Based Recommendation Process

3.2.1 Recommendation Results Based on Investor Neighbors

The foundation of the collaborative filtering algorithm based on investors is to construct a rating matrix of investors' preferences for products. The main idea is to calculate the similarity between different investors, find the most similar group of investors, and recommend products to the target investor that have been purchased by these similar investors but not yet by the target investor, thus achieving the goal of personalized recommendation. Neighbor search is the core step of the collaborative filtering algorithm, and similarity calculation is the means to implement neighbor search. The nearest neighbor recommendation steps based on the comprehensive similarity of investors are shown in Fig. 2.

(1) Data Processing for Investor Rating Similarity.

First, obtain investor ratings and fill in the sparse rating matrix. In this study, sentiment ratings of investors towards fund products are obtained as an auxiliary tool for predicting ratings of unknown fund products in the recommendation system.

Combining the Chinese sentiment lexicon for the financial domain [12] with open-source interfaces to evaluate investors' preferences towards fund products. The comments of investors on fund products are segmented and sentiment analysis is conducted using jieba tokenizer and SnowNLP. The results of sentiment values calculated by SnowNLP range from 0 to 1, where a value closer to 1 indicates a stronger positive sentiment and a value closer to 0 indicates a stronger negative sentiment. Finally, the ratings are mapped to a range of [0, 5] for subsequent recommendation.

Next, the rating matrix is filled in. In this study, the Slope One algorithm proposed by Xiaodong et al. is used to fill in the rating matrix, which can ensure the diversity of filled values and avoid recommendation errors caused by single filling [10, 11].

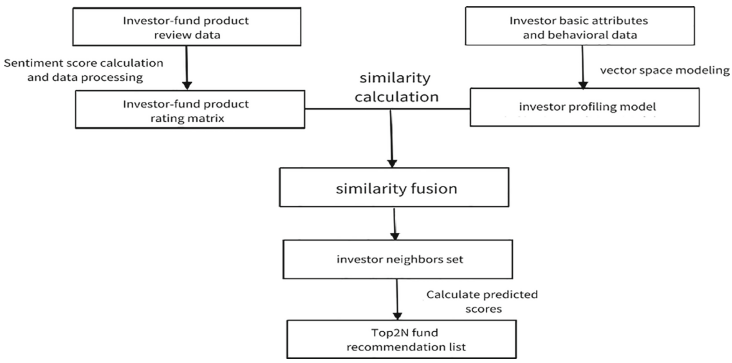

Fig. 2. Steps for Fund Product Recommendation based on Investor Neighbors

After filling in the investor-product rating matrix, the similarity score $\mathrm{sim}_{\mathrm{CF}}(u, v)$ between investor u and investor v based on sentiment ratings is calculated using cosine similarity, as shown in the following formula:

$$Sim_{CF}(u,v) = \frac{\sum_{i \in I_{uv}} (r_{u,i} - \vec{r}_u)(r_{v,i} - \vec{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{u,i} - \vec{r}_u)^2} \sqrt{\sum_{i \in I_{uv}} (r_{v,i} - \vec{r}_v)^2}} \tag{9}$$

(2) Data Processing for Investor Profile Similarity.

① Investor basic attribute similarity

The similarity between investors u and v in terms of basic attribute dimensions can be represented as:

$$\mathrm{Sim_I(u, v)} = \mathrm{t_1} * \mathrm{sim(u, v, a_1)} + \mathrm{t_2} * \mathrm{sim(u, v, a_2)} + \ldots\ldots + \mathrm{t_6} * \mathrm{sim(u, v, a_6)} \tag{10}$$

where SimI(u,v) represents the similarity between investors u and v, and t is calculated based on the weights of each basic attribute indicator obtained in Chapter 3; a_1 to a_6 represent the values of the 6 attributes, and sim(u,v,a_i) represents the similarity between investors u and v on the corresponding attribute. If the label value is consistent, the similarity is 1, otherwise 0.

② Investor Product Preferences Similarity

The labels of investor profiling are sorted according to their weights, and the top 20 labels are selected for similarity calculation, as shown in formula(11), where Tag(m) and Tag(n) represent the label sets of investors u and v, respectively, based on the weights assigned to them.

$$sim_P = (u,v) = \frac{|Tag(m) \cap Tag(n)|}{\sqrt{|Tag(m)||Tag(n)|}} \tag{11}$$

Therefore, the similarity of investor profiles can be represented as:

$$\mathrm{Sim_{IP}(u,v) = Sim_I(u,v) + Sim_P(u,v)} \tag{12}$$

(3) Calculation of Comprehensive Investor Similarity.

Based on the investor profiling and improved collaborative filtering algorithm, suitable coefficients are used to fuse them in order to obtain the comprehensive similarity of investors, as shown following:

$$\mathrm{Sim(u, v)} = \beta \mathrm{Sim_{CF}(u, v)} + (1 - \beta)\mathrm{Sim_{IP}(u, v)} \tag{13}$$

When an investor joins the platform but has not engaged in any commenting behavior, the similarity calculation only considers the basic attribute data of the investor.

(4) Calculate the similarity between the target investor and other investors through step (3), and find the top K nearest neighbors V_k with the highest similarity, i.e., search

for the desired set of investors VK = {v_1, v_2,…, v_k},within the entire investor set U, where sim(u, v_1) > sim(u, v_2) >… > sim(u, v_k).

$$P_{u_k} = \bar{r}_u + \frac{\sum_{v \in U_k} Sim(u, v)(r_{vk} - \bar{r}_v)}{\sum_{v \in U_k} Sim(u, v)} \tag{14}$$

According to the scoring formula, calculate the ratings of the fund products that the target investor's nearest neighbors have evaluated, and select the top 2N products based on their ratings as the recommendation candidate set.

3.2.2 Recommendation Results Based on Investor's Temporal Preferences

Integrated with the construction of investor's profile and weight updating methods in Chapter 3, this section further filters the previously obtained Top2N recommendation results based on the match degree between fund products and investors. The specific steps of the algorithm for recommendation are as follows:

(1) Let the representation of the attribute vector of the fund product to be recommended be: $\overrightarrow{P} = (W_1, W_2, W_3, ..., W_n)$.
(2) Investor's product preference vector: The main reflection of investor's fund product preference behavior on the online investment communication platform is the fund products that investors have commented on. Based on the comments, n most weighted product preference labels are extracted to represent the investor's preferences. The investor's product preference vector can be represented as: $\overrightarrow{U} = (w_1, w_2, w_3, ..., w_m)$
(3) Investor's product preference labels and weights based on time series forecasting, inputting the initial value of investor's preference label vector: $\underset{t0}{\overrightarrow{U}} = (w_{1,t0}, w_{2,t0}, w_{3,t0}, ..., w_{m,t0})$. The time series of preference labels for the wm preference label from t0 to Tt periods, integrated based on time series, can be represented as: $timeseries(w_m) = w_{m_{t0}}, w_{m_{t1}}, ..., w_{m_{tT}}$. Using the integrated time series of wm preference labels as the initial values for the observation sequence, predict the forecasted weights for the next time step in the top2N product set.

Utilizing cosine similarity, calculate the similarity between the updated investor preference vector and the attribute vector of the top2N fund products. Select the top n products with the highest similarity as the recommendation results.

4 Experimental Results and the Analysis

4.1 Experimental Design

In this study, several fund forums on Eastmoney.com were selected as the crawling targets. Comment data was crawled randomly from January 2020 to December 2021, totaling 89,589 comments. Duplicate comments (7,973 in total) were removed, and investors were assigned unique IDs. Fund codes were obtained from the six-digit numerical IDs

included in the URLs of the forums. The crawled content included poster's username, poster's forum membership age, post title, post content, post time, fund forum link, and fund forum title, among others. Each investor's profile included their nickname, generated investor ID, comment content, and fund name. Textual product label data, such as fund code, fund abbreviation, fund manager, fund company, fund type, risk level, fund theme, investment style, and fund rating, were obtained from iFind software by Tonghuashun.

The data were filtered to select 302 investors with abundant comment data, 1,608 fund products, and 7,030 ratings, with each investor rating at least 20 products. For experimental purposes, 80% of the data was used as the training set, and the remaining portion was used for testing. Ten-fold cross-validation was performed to ensure model accuracy.

4.2 Analysis of the Results

In this comparative experiment, the recall rate of three models, namely traditional collaborative filtering (CF), user-based collaborative filtering with user profiling (UCF), and the proposed investor-profile-based recommendation algorithm (IPCF), was compared. Additionally, the relationship between time series and accuracy was evaluated.

① The recall rate results

The recall rate was calculated for the three models, and based on the trend of recall rate changes observed from Fig. 3, it was found that all three algorithms showed an initial increase followed by a tendency to stabilize. Among them, the algorithm models with the highest recall rate in descending order were: investor-profile-based recommendation algorithm, user-based collaborative filtering with user profiling, and traditional collaborative filtering. This indicates that the algorithm model proposed in this study has a clear advantage in terms of recall rate.

② In order to examine the impact of time series on the algorithm models, the relationship between time series and accuracy was studied for two recommendation algorithms. The experimental results are shown in Fig. 4.

The recommendation based on static profiles essentially belongs to a content-based recommendation approach. In the initial stage of the recommendation model, historical data of all investors are provided to the model for training. Therefore, compared to the recommendation results based on investor profiles with time series forecasting steps, static profile recommendation may exhibit better performance. This is because

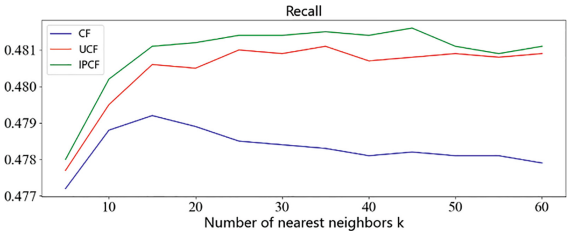

Fig. 3. Recall Rate Recall Rate Curves for Different Algorithms

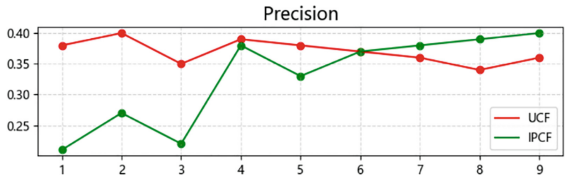

Fig. 4. Comparison of Recommendation Accuracy based on Time Series Preferences

the training data for investor profiles is segmented into time periods. As seen from the graph, the accuracy of static profile recommendation is higher than that of dynamic recommendation in the first six months. However, as time goes on, investors' interests and preferences change continuously, and the recent prediction accuracy shows an upward trend. When the training set accumulates over time, the trend becomes stable. Overall, the accuracy of the proposed recommendation model in this study is not particularly ideal, which may be related to the small dataset and the analysis of accuracy of static profile recommendation. As can be seen from Fig. 4, there is significant fluctuation in the experimental data, which may be due to short-term market volatility in the dataset, as well as rapid updates of data in the investment exchange platform, which could also affect the accuracy results due to the sentiment atmosphere of investors in the "Fund Bar".

5 Conclusions

Currently, there is limited academic research in the field of investor profiling and personalized mutual fund product recommendation, both domestically and internationally. The mutual fund product recommendation model based on investor profiling designed in this study can model investor profiling and judge their product preferences, providing personalized mutual fund product recommendations. In future research, the mutual fund product recommendation strategies can be further improved by considering the potential returns of mutual fund products in the future, and recommending appropriate purchase timing or investment portfolios to investors may be a future direction for improvement.

References

1. Shen, J. B. (2017). Application of User Profile in Internet Finance. Modern Business, 2017(33), 57–58. (CNKI:SUN:XDBY.0.2017–33–026)
2. Dong, Xinxin, Li, Chunshan, Chu, Dianhui. [IEEE 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC) - Guangzhou, China (2017.12.12–2017.12.15)] 2017 IEEE Internation[J]. :1251–1255.
3. Zhao, M., Li, X., Li, X. T., & Wu, D. (2013). Classification Study of Business Bank Fund Customers Based on Cluster Analysis. Management Review, 25(07), 38-44. DOI: https://doi.org/10.14120/j.cnki.cn11-5057/f.2013.07.006.
4. Liu, J. L. (2017). Logistic Regression Analysis of Investor Risk Preference in Hengtai Securities. Inner Mongolia Science and Economy, 18, 37–38. (CNKI:SUN:NMKJ.0.2017–18–020)

5. Wang Z , He P L , Guo L S , et al. Clustering anaslysis of customer relationship in securities trade[C]// International Conference on Machine Learning & Cybernetics. IEEE, 2005.
6. Yu, C. M., Tian, X., Guo, Y. J., & An, L. (2018). User Profile Research Based on Behavior-Content Fusion Model. Library and Information Service, 62(13), 54-63. DOI: https://doi.org/10.13266/j.issn.0252-3116.2018.13.008.
7. Tejeda-Lorente, Á., Bernabé-Moreno, J., Herce-Zelaya, J., Porcel, C., & Herrera-Viedma, E. (2019). A risk-aware fuzzy linguistic knowledge-based recommender system for hedge funds. Procedia Computer Science, 162(C), https://doi.org/10.1016/j.procs.2019.12.068.Gan, Q. (2015). Design and Implementation of P2P Online Lending Product Recommendation System Based on Hybrid Algorithms [Dissertation]. Beijing: University of Chinese Academy of Sciences (School of Engineering Management and Information Technology).
8. Zhou, Y. (2014). Personalized Recommendation Research Based on User Behavior Analysis of Securities Wealth Management Products [Dissertation]. Chengdu: University of Electronic Science and Technology of China.
9. Lemire, D., & Maclachlan, A. (2005). Slope One predictors for online rating-based collaborative filtering. In Proceedings of the Fifth SIAM International Conference on Data Mining (pp. 471–480).
10. Zhang, P., & Ge, X. (2016). K-nearest neighbor Slope One algorithm with fused label similarity. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition, 28(4), 518–524. (CNKI:SUN:CASH.0.2016–04–012)
11. Yao, J., Feng, X., Wang, Z., Ji, R., & Zhang, W. (2021). Tone, emotion, and market impact: Based on financial sentiment lexicon. Journal of Management Sciences in China, 24(5), 26-46. DOI: https://doi.org/10.19920/j.cnki.jmsc.2021.05.002
12. Wu, D. (2008). Dynamic index smoothing prediction method and its application. Journal of Systems & Management, 2008(02), 151–155. CNKI:SUN:XTGL.0.2008–02–005
13. Gardner, E. S., & Diaz-Saiz, J. (2007). Exponential smoothing in the telecommunications data. International Journal of Forecasting, 2007(1), https://doi.org/10.1016/j.ijforecast.2007.05.002.