OPEN ACCESS
Fractals, Vol. 28, No. 8 (2020) 2040017 (12 pages)
© The Author(s)
DOI: 10.1142/S0218348X20400174

AN USER INTENTION MINING MODEL BASED ON FRACTAL TIME SERIES PATTERN

SHAOFEI WU
Hubei Province Key Laboratory of Intelligent Robots
Wuhan Institute of Technology
Wuhan, P. R. China
School of Computer Science and Engineering
Wuhan Institute of Technology
Wuhan, P. R. China
wasbfc@yeah.net

Received December 29, 2019
Accepted February 24, 2020
Published July 10, 2020

Abstract

Users use the network more and more frequently, and more and more data is published on the network. Therefore, how to find, organize, and use the useful information behind these massive data through effective means, and analyze user intentions is a huge challenge. There are many time series problems in user intentions. Time series have complex characteristics such as randomness and multi-scale variability. Effectively identifying the inherent laws and objective phenomena contained in time series is the purpose of analyzing and processing time series data. Fractal theory provides a new way to analyze time series, and obtains the characteristics and rules of time series from a new perspective. Therefore, this paper introduces the fractal theory to analyze the time series problem, and proposes an improved G-P algorithm to realize the prediction and mining of user intentions. First, the method of array storage instead of repeated calculations is used to improve the method of saturated correlation dimension. Second, the Hurst exponent of the time series is obtained by the variable scale range analysis method. Finally, a fractal model for predicting user intent in short time series is established using the accumulation and transformation method. The experimental results show that the use of fractal theory can effectively describe the relevant characteristics of time series, the development trend of user intentions can be mined from big data, and the prediction model for short time series can be established to achieve information mining of user intentions.

Keywords: Fractal Theory; Time Series; User Intent; Data Mining.

1. INTRODUCTION

We live in the era of data. With the widespread application of the network and the rapid increase in the number of users, data has also grown in an explosive manner. There is bound to be huge value in massive data. Users use the network more and more frequently, and more and more data is published on the network.[1] While users enjoy various services such as dating, games, and instant communication brought by the Internet, they also produce a large amount of data through the network, which is a reflection of the user's true thoughts and ideas. With the continuous penetration of the user-centric concept in various fields and industries, people have gradually realized the importance of acquiring user data in the network and analyzing the intention of the network user.[2] Therefore, how to find, organize, and utilize the useful information behind these massive data through effective means, and analyze the user intention pattern will be a huge challenge.[3,4]

With the popularization of the Internet, the analysis and research of user intentions at home and abroad is also increasing.[5] However, due to the difficulty of data acquisition, early analysis and research on the intent of network users focused on the interaction intent among small-scale users. With the development of information technology and the improvement of database computing and storage capabilities, the problem of obtaining large-scale network data has also been solved, providing conditions for in-depth study of user intention characteristics. Hoang *et al.*[6] analyzed and classified users' motivations for use, and found that users can be classified into three categories: friend interactors, information collectors, and information providers. The user's purpose can be communication, dialogue, self-expression, news release, and information sharing four categories. Deanna *et al.*[7] analyzed the intentional characteristics such as the relationship between the number of users' followers and the number of followers, and found that the user's work and rest time has a great influence on the change in the number of releases. He *et al.*[8] analyzed the user data of the entire network, and obtained the distribution and correlation of the number of user followers and followers on a larger scale. Wilbert *et al.*[9] analyzed a large amount of user data, compared the three parameters of retweets, citations, and followers, and studied the effect of user intentions on the topic propagation process. Naeimeh *et al.*[10] measured the average number of retweets and comments of users and the time distribution, and found that the first three obey the power law distribution. There are also some researchers who conduct analysis and modeling studies from a mathematical perspective.[11–13] Therefore, many researchers have applied mathematics to the prediction and analysis of user intentions. Du *et al.*[14] found that the user's intention is related to the network structure of the user and the number of user friend retweets, and proposed a publishing intention model based on the social network relationship and the degree of user interaction. Chen *et al.*[15] used the fractal theory model to mine and cluster user-focused topics, and compared the clustering results with other models, and found that the fractal theory model can more effectively mine the topics that the user cares about. Li *et al.*[16] analyzed and conducted in-depth research on the intentions of users to forward information, the different intentions of spam users and normal users, and the intent expressed by the five personality traits, which provided a new method for solving related problems in the network.

There will be more time series problems in the user's intention. The time series has complex characteristics such as randomness and multi-scale variability.[17] It can effectively identify the inherent laws contained in the time series[18] and the objective phenomenon is the goal of analyzing and processing time series data. Fractal theory provides a new way to analyze time series, and obtains the characteristics and rules of time series from a new perspective.

Therefore, based on the fractal theory, this paper explores the characteristics of time series, mines the development trend of user intentions from big data, and builds a prediction model for short time series to achieve information mining of user intentions.

The detailed chapters are arranged as follows: Section 2 introduces related concepts. Section 3 proposes a user intention mining model based on fractal time series pattern. Section 4 is the experimental results and analysis. Section 5 is the conclusion.

2. RELATED CONCEPTS

2.1. Definition of Fractal

Fractal originally refers to irregular, fragmented objects. It was first introduced to the scientific field by Mandelbrot to describe more complex processes and graphics. Fractals can be divided into regular fractals and irregular fractals. Roughness and complexity of many things in nature are usually random, such as the changing Brownian motion trajectory, complex tortuous coastlines, and so on. These curves have approximate or statistical self-similarity. These curves belong to the category of irregular fractals.

Mandelbrot gives two mathematical definitions of fractals:[19]

(1) If a set has a Hausdorff dimension DH strictly larger than its topological dimension DT in Euclidean space, then the set is a fractal set, referred to as fractal for short. In general, DH is not an integer but a fraction.
(2) Fractals are shapes that are similar in some way to the whole. This definition emphasizes the self-similarity between parts and wholes (including small parts and large parts) in a graph.

The second definition above emphasizes the characteristics of fractal self-similarity, which is a property that natural things generally have, that is, between parts, between parts and the whole, there is a certain self in terms of function, form, time, and space. Similar characteristics. However, Mandelbrot is not satisfied with either of these definitions, because they exclude certain fractals. Theoretical research and practical application show that the above two definitions really cannot completely cover such a wide range of fractals. In fact, a fractal cannot be defined very well so far, so people try to understand fractals from another angle, that is, to study what a fractal is by studying a series of characteristics of the fractal. Scientist K. Falconner described the fractal as follows:[20]

(1) A fractal set has proportional details at any small scale, or it has a very fine structure.
(2) Fractal set cannot be described by traditional geometric language. It is neither the trajectory of points that meet certain conditions, nor the solution set of some simple equations.
(3) Some self-similar form of fractal set may be approximate self-similarity or statistical self-similarity.
(4) In general, the fractal dimension of a fractal set is strictly greater than its corresponding topological dimension.
(5) In some cases, the fractal set can be defined by a very simple method, which may be generated by the iteration of the transformation.

2.2. Basic Characteristics of Fractals

Self-similarity and scale invariance are the most important features of fractals. They are not only the theoretical basis of fractal research, but also important criteria for fractal analysis. The fractal analysis methods and the definition and calculation of fractal dimensions all start from these two characteristics.

Self-similarity: refers to the local properties or characteristics of the research object similar to the whole. Self-similarity exists in many fields of study, such as materials science, economics, physics, chemistry, biology, astronomy, and so on. Self-similarity is a universal manifestation of material movement and development, which can exist at multiple levels of the material system, which is one of the universal laws of nature.

Scale invariance: Scale invariance refers to the selection of a local area on a fractal, regardless of whether it is enlarged or reduced, its various characteristics (including morphology, structure, nature, function, complexity, irregularity, etc.) No change will occur, so the scale invariance is also called scaling symmetry.[21] For an actual fractal, this scale invariance applies only to a certain range, and beyond a certain range, it is not a fractal.

The characteristics of the two fractals, self-similarity and scale-invariance, are closely related. Things with self-similarity must have an interval that satisfies scale-invariance. Objects with characteristics must have self-similarity. Although there are many kinds of fractal structures, the number of regular fractal is very small, which indicates that the self-similarity and scale invariance of fractal can only exist in a certain scale range, and its structure is not infinite. In fact, many structures with fractal characteristics do not have strict self-similarity, but only have self-similar structures in a statistical sense. The scope of fractal theory is only limited to a certain interval, and for scale-free intervals without self-similarity, the meaning of fractal theory no longer exists.[22]

2.3. Fractal Dimension of Time Series

For a random process, a random variable sequence $X_1, X_2, X_3, \ldots$ arranged at time $t_1 < t_2 < t_3 < \ldots$ with time t as the gate variable is called a time series. The n observations $x_1, x_2, x_3, \ldots$ of the random variable sequence are called n observation samples of the time series.

The data in the actual problem is only limited observation time series of samples, so the main task of the time series analysis is based on the characteristics of observed data for the data to establish a reasonable statistical model as far as possible, and then using the statistical properties of the model to explain the data statistical rule, in order to achieve the purpose of control or forecast. As shown in Fig. 1, the fractal theory is used to analyze the time series of wind speed in the wind field, and the resulting fractal dimension is obtained.

Time series can be divided into two types: stationary time series and non-stationary time series. A random process, if its mathematical expectation and variance do not change with time, and the autocorrelation function is only a function of their time interval and has nothing to do with absolute time, it is called a stationary random process, and the corresponding time series is called a stationary time series. The so-called stationary time series analysis is a method of studying the linear correlation between a time series with different time intervals. The established model is called an autoregressive equation.

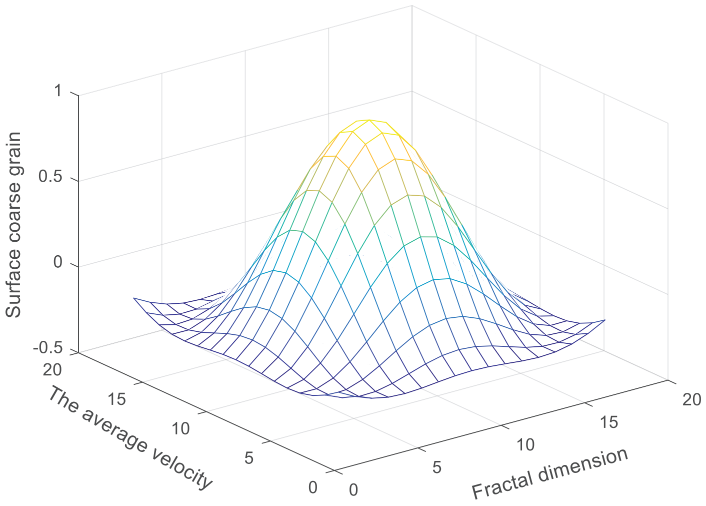

Fig. 1 Analysis of wind field time series using fractal theory.

The number of elements of a time series data is called the dimension. In general, the original time series data has a high dimensionality. On the one hand, because the time series data has a high dimension, on the other hand, because the time series database is often large, that is, the number of time series is large. This brings difficulties to time series pattern discovery, so the original time series needs to be represented at a high level of abstraction in order to find patterns that are easy to understand and explain, that is, to determine appropriate exploration granularity issues. If the excavation granularity used is too small, it is difficult to grasp the operating law of the system as a whole. At the same time, the presence of noise will affect the quality of the model. In addition, the excavation process will increase unnecessary calculation costs.

In the face of massive data, it is very difficult to directly operate a high-dimensional data space. A sequence with n points can be regarded as a point in n-dimensional space. If the n-dimensional point is directly indexed by the spatial access method (SAM) multi-dimensional index structure, it will easily lead to dimensional disaster. Therefore, it is necessary to study the appropriate data representation for dimensionality reduction and effective mining on efficient and convenient representation.

For things with fractal properties, the fractal dimension is an important parameter for its fractal properties. In European-style geometric space, straight lines or curves are one-dimensional, planes or spheres are two-dimensional, and cubes, cuboids, and so on are three-dimensional. These dimensions are integer values; however, for fractal dimensions, such as Koch curve. In order to quantitatively describe non-integer-valued dimensions similar to fractal dimensions, German mathematician F. Hausdorff proposed the definition of Hausdorff dimension.[23] Fractal dimension is one of the most important parameters for describing fractals.

Fractal dimension tells us how shapes or time series fill their space.

This paper generalizes Hausdorff's method to obtain a calculation equation for the fractal dimension as shown in Eq. (1).

$$D = \frac{\ln(N(\varepsilon))}{\ln(1/\varepsilon)} (\varepsilon \to 0). \quad (1)$$

However, it is very difficult to calculate the hydrological time series according to this definition, and the calculation results are not accurate. Therefore, an indirect method is needed to calculate the fractal dimension. For the fractal runoff time series, the relationship between the fractal dimension D and the index Hurst index is shown in Eq. (2).

$$D = 2 - H. \quad (2)$$

In the equation, H represents the Hurst index, and the Hurst number can be calculated by the R/S method.

3. USER INTENTION MINING MODEL BASED ON FRACTAL TIME SERIES PATTERN

The process of user intent data mining is a process of continuously exploring the characteristics of data, establishing and testing models, and discovering the characteristics of customer intent. The effective method of data mining application is: starting from a small, critical problem, establishing a relatively effective model, and continuously testing and improving the model through application practice, and gradually solve the problem for the user. User intention data mining is to digitize the user's intentions, to process and classify the data, and to use a certain association rule to quantitatively analyze the user's intention data. Finally, perform a qualitative analysis based on the data analysis results. The intent data mines valuable information in order to better meet user needs and provide better services. The user intention data application system architecture is shown in Fig. 2. The user behavior data application system framework is divided into a data collection layer, a data processing layer, a data service layer, and a business application layer. The data collection layer mainly collects data generated by users in the network, such as browsing. The data processing layer needs to perform data cleaning, data extraction, and feature extraction on the data. Both the data service layer and the business application layer indicate that through the collection and analysis of data, the user behavior is ultimately predicted.

There are more time series issues in user intent. Fractal theory obtains the characteristics and rules of time series from a new perspective. This paper discusses the characteristics of time series based on fractal theory. Therefore, this paper proposes an improved G-P algorithm for the time series problem of user intention, and then constructs a mining model of user intention.

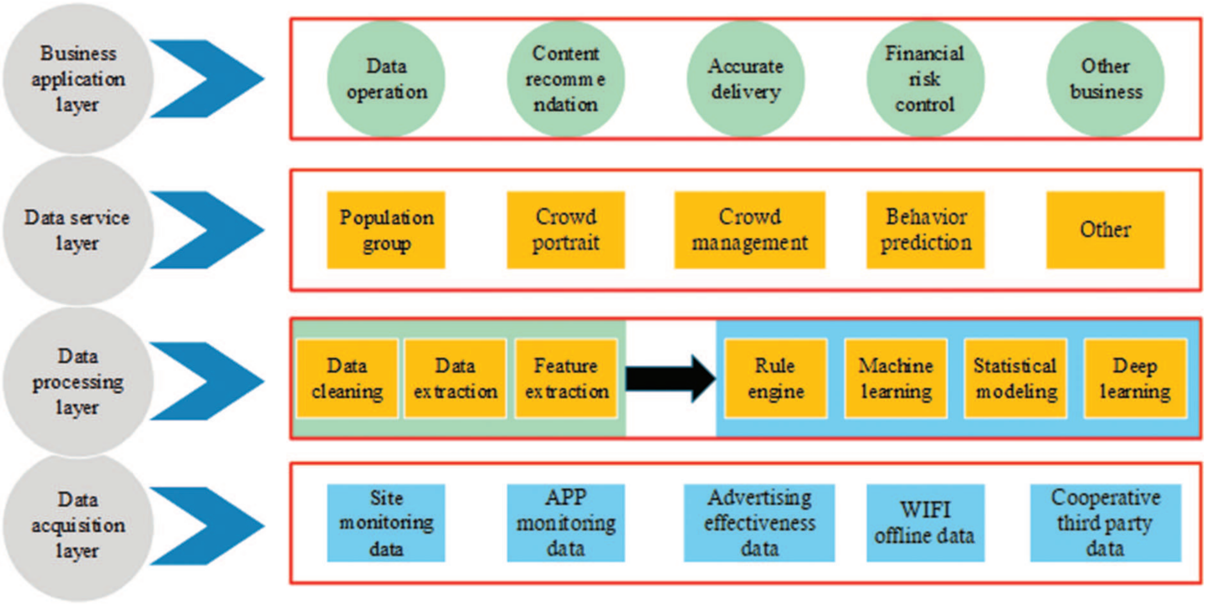

Fig. 2 User intent data application system architecture.

3.1. Calculating Fractal Dimension with Saturated Correlation Dimension Method

When studying time series, the sample data scattered on the time axis cannot describe the changing rules of complex multidimensional systems, and phase space reconstruction provides conditions for it. The saturated correlation dimension method[24] (referred to as the G-P algorithm) is the most commonly used method to determine the correlation dimension in the phase space reconstruction theory. The principle is as follows:

Phase space reconstruction is performed on the time series $\{x_1, x_2, x_3, \ldots, x_N\}$ by taking the delay time t and the embedding dimension m. The reconstructed sequence is $\{X_1, X_2, X_3, \ldots, X_M\}$. Among them, there is a relationship of $X_i = \{x_i, x_{i+t}, \ldots, x_{i+(m-1)t}\}$ $(1 \leq i \leq M)$ between X and x. The variable M is the number of phase points, and $M = N - (m - 1)$. Take the neighborhood radius r and define the correlation integral as shown in Eqs. (3) and (4).

$$C(m,r) = \frac{2}{M(M-1)} \times \sum_{1 \leq i < j \leq M} H(r - ||X_i - X_j||_\infty). \tag{3}$$

The equation for calculating the value of the H (x) function in Eq. (3) is shown in Eq. (4).

$$H(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{4}$$

Observing the original G-P algorithm process, it is found that each time the calculation of the associated integral requires m times of subtraction between time series elements. Then define the function as shown in Eq. (5).

$$Dx(i,j) = |x_i - x_j|, \quad 1 \leq i < j \leq N. \tag{5}$$

This function indicates that the absolute value of the difference between each two time series elements is stored in the array $Dx(i,j)$, and the independent variables i, j are the subscripts of the time series elements. When the G-P algorithm runs, it first fetches all i, j in the domain, and calculates all the values of the array $Dx(i,j)$. Each time the loop calculates $||X_i - X_j||_\infty$, no further calculation is needed, just the magnitude of the absolute value of the m elements in the vector $||X_i - X_j||$, the maximum is $||X_i - X_j||_\infty$. Therefore, when the number of embedded dimensions is K_m and the number of neighborhood radii is K_r, the difference between the number of calculations of the G-P algorithm before and after the improvement is shown in Eq. (6).

$$\Delta n = ((mM(M-1)K_m K_r)/2) - N^2. \tag{6}$$

Due to the number of time series elements $N \gg m$, Eq. (6) is transformed as shown in Eq. (7).

$$\Delta n = ((mK_m K_r)/2)(N - m + 1) \times (N - M) - N^2 \gg 0. \tag{7}$$

That is $\Delta n \gg 0$. Therefore, the number of calculations of the improved algorithm is greatly reduced.

Take K_r different neighborhood radii r to calculate the correlation integral and draw the $\ln(C(m,r))$–$\ln(r)$ curve. The slope of the curve is the correlation dimension D of the time series. Take K_m different embedding dimensions m and repeat the above process to draw a D–m curve. When D does not change with increasing m, the critical value of m represents the minimum number of variables required to describe the time series, and the sequence tends to a finite dimension D_m in the phase space.

3.2. Calculating Hurst Index of Time Series with Variable-Scale Range Analysis

Hurst, a British hydrologist, studied the scale intent of time series data and found that the results of time series records have self-affine characteristics, and thus created a rescaled range analysis method. Abbreviated as R/S analysis method, which can be described with the help of fractal theory. According to the principle of the R/S analysis method,[25] the Hurst index is calculated, and its size can reveal the strength of the time series correlation. The larger the gap between H and 0.5, the stronger the correlation within the time series. The relationship is shown in Table 1.

3.3. Establishing a User Intention Analysis Model for Fractal Time Series

The definition of power series distribution of fractal is shown in Eq. (8).

$$N = C/r^D. \tag{8}$$

In the equation, r is the characteristic line; N is the magnitude of the measured object; D is the fractal dimension; C is the parameter. In the standard

Table 1 Corresponding Table of Hurst Index Value and Time Series Status.

Value Range of *H*	**[0, 0.5)**	**0.5**	**(0.5,1)**	**1**
Time series status	Anti-sustained effect	Irrelevant	Sustained effect	Completely sure

fractal distribution, the fractal dimension is constant. Substitute any two points $(r_i, N_i), (r_j, N_j)$on the straight line into Eq. (8), and arrange the calculation equations of D and C, such as Eqs. (9) and (10) shown.

$$D = \frac{\ln(N_i/N_j)}{\ln(r_i/r_j)}, \tag{9}$$

$$C = N_i r_i^D. \tag{10}$$

The point distribution in the standard fractal $lnN - lnr$ double logarithmic image is a straight line, but in practical applications, the fractal dimension is usually not constant, and the straight line relationship does not appear in the double logarithmic image. In this case, the method of accumulation and transformation can be adopted. With $\{N_i\}$ as the basic column, we find the accumulation and sequence of each order, as shown in Eqs. (11)–(13).

$$\begin{aligned} \{S1_i\} &= \{N_1, N_1 + N_2, N_1 + N_2 \\ &\quad + N_3, \ldots\}, i = 1, 2, \ldots, n, \end{aligned} \tag{11}$$

$$\begin{aligned} \{S2_i\} &= \{S1_1, S1_1 + S1_2, S1_1 + S1_2 \\ &\quad + S1_3, \ldots\}, i = 1, 2, \ldots, n, \end{aligned} \tag{12}$$

$$\begin{aligned} \{S(j+1)_i\} &= \{Sj_1, Sj_1 + Sj_2, Sj_1 + Sj_2 \\ &\quad + Sj_3, \ldots\}, i = 1, 2, \ldots, n. \end{aligned} \tag{13}$$

The fractal dimension D of the corresponding accumulation and sequence is determined by $(r_i, Sj_i), (r_{i+1}, Sj_{i+1}), i = 1, 2, \ldots, n-1$, respectively. The fractal dimension of each segment is relatively close, that is, the accumulative sum sequence $\{Sj_i\}$ which has a good fitting result of the fractal distribution is used as the simulation and prediction sequence. Because the future development trend is most closely related to the last fractal dimension used in modeling, this fractal dimension is selected as the fractal dimension used for prediction, and Sj_k is calculated according to Eqs. (11), (12), and (13). Then, calculate N_k from Eqs. (14) and (15), and complete the prediction. The specific process is shown in Fig. 3.

$$S(j-1)_k = Sj_k - Sj_{k-1}, \tag{14}$$

$$N_k = S1_k - S1_{k-1}. \tag{15}$$

4. RESULTS AND DISCUSSION

4.1. User Data Standardization

This paper uses the digital library to simulate the user's intention data mining. The digital library analyzes user intent data, analyzes user habits, and provides users with personalized services. Before data association, the data must be standardized, and the array storage method should be used

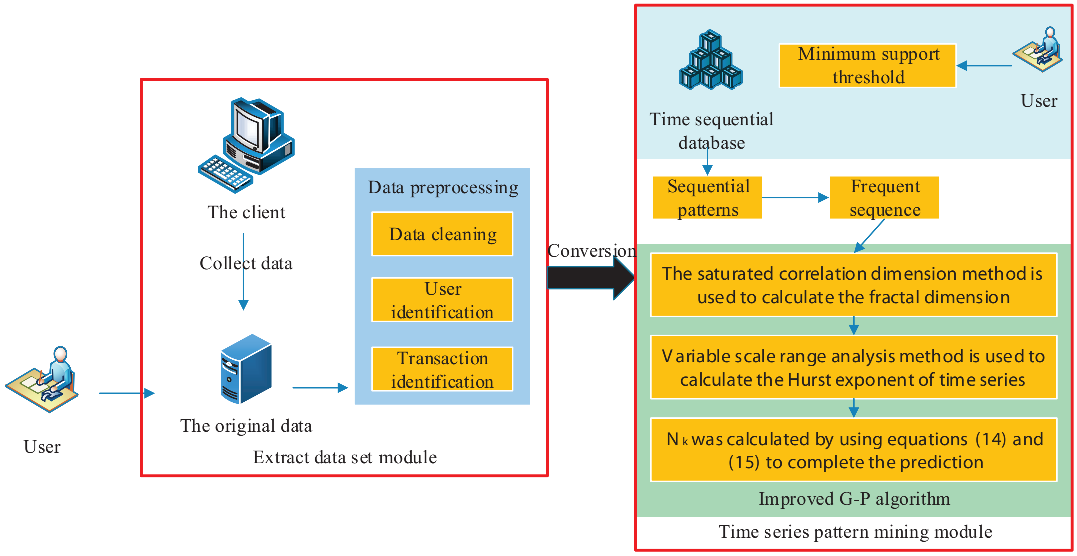
Minimum support threshold
User
Time sequential database
Sequential patterns
Frequent sequence
The saturated correlation dimension method is used to calculate the fractal dimension
Variable scale range analysis method is used to calculate the Hurst exponent of time series
N k was calculated by using equations (14) and (15) to complete the prediction
Improved G-P algorithm
Time series pattern mining module
Conversion
The client
Collect data
Data preprocessing
Data cleaning
User identification
Transaction identification
User
The original data
Extract data set module

Fig. 3 Flow chart of user intent mining algorithm based on fractal time series.

instead of repeated calculations to improve the saturated association dimension method. First collect data samples and perform data mining on 2000 digital library users. The number of transaction item sets is 2000. Each item includes user account, password, age, gender, education, major, search history category, borrowed book classification number, download, and resource situation.

Let the user information domain $X = \{x_1, x_2, x_3, \ldots x_n\}$ be the classified objects, and each object is measured by m indicators $x_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots x_{im}\}, i = 1, 2, \ldots, n$, where x_1 is the user name, x_2 is the password, x_3 is the academic degree, x_4 is the specialty, and x_5 is the middle picture classification number where the user has searched the journal a lot. The symbol x_6 indicates the library's literature index, x_7 indicates the user's download records at different times, and x_8 indicates the user's downloaded material name, subject category, etc.

The original data matrix can be obtained as shown in Eq. (16).

$$\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix}. \quad (16)$$

After the original data is obtained, the data can be subjected to standard deviation transformation, as shown in Eq. (17).

$$x'_{ij} = (x_{ij} - \bar{x}_j)/s_j \quad (i = 1, 2, \ldots, b;\ j = 1, 2, \ldots, m). \quad (17)$$

Among them, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and $s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$.

After the transformation, the unit dimensions of all variables are removed, and the mean value is 0, and the standard deviation is 1, and then the range transformation is performed as shown in Eq. (18).

$$x'_{ij} = \frac{x_{ij} - \min\{x_{ij}|1 \le i \le n\}}{\max\{x_{ij}|1 \le i \le n\} - \min\{x_{ij}|1 \le i \le n\}}. \quad (18)$$

After the transformation, all variables take values between [0, 1].

4.2. Correlation Dimension and R/S Analysis of Time Series

Use the improved G-P algorithm to calculate the correlation dimension on the time series, draw the $\ln(C(m,r)) - \ln(r)$ image, calculate the slope D of the trend line of each straight line in the image, and draw the D_m image as Fig. 4 shows. It can be seen from Fig. 4 that with the increase of the embedding dimension m, the slope D of the curve gradually increases at first, and then it tends to be consistent after the development. Starting from $m = 8$, D basically does not change with the increase of m, and the stable value $D_8 = 1.465$, indicating that the time series tends to a finite dimension of 1.465 in the phase space, that is, the saturation correlation dimension is 1.465.

An R/S analysis is performed on the time series, and the image of $\ln(\mathrm{R/S}) - \ln(t)$ is shown in Fig. 5. It can be seen from Fig. 5 that the Hurst index $H = 0.917$ of the time series is estimated by the R/S analysis method, and the correlation coefficient $R = 0.989$. The trend of the time series curve obtained

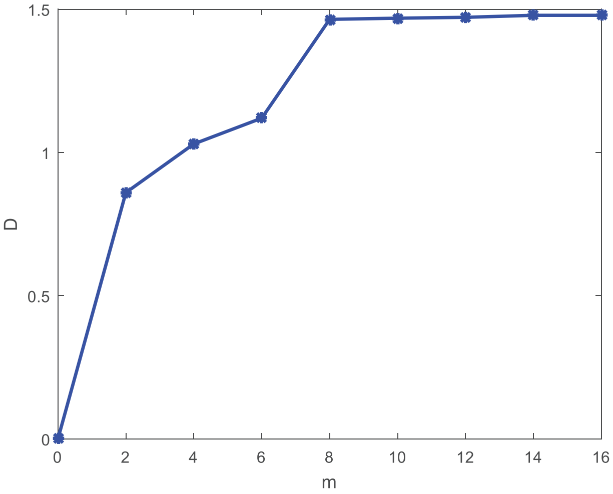

Fig. 4 The image of relationship between D and m.

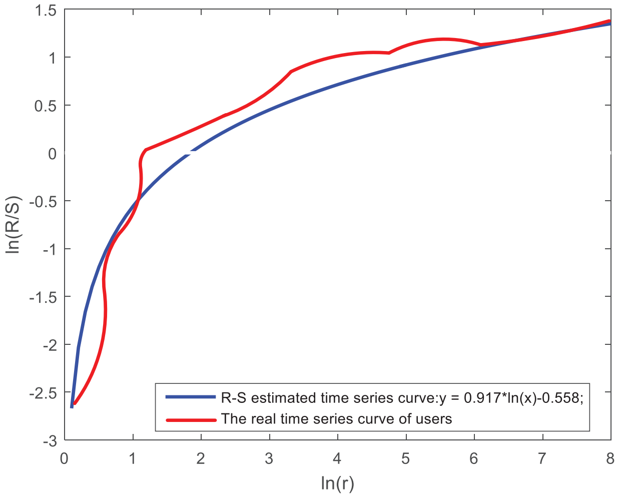

Fig. 5 The image of relationship between $\ln(r)$ and $\ln(\mathrm{R/S})$.

Table 2 **Data Correlation Results of the Original G-P Algorithm.**

Association Rule ID	Frequent Item Sets	Minimum Confidence
Profession	Computer, electronics, communications	0.037
Education	Master PhD	0.031
Search preferences	Electronic journals	0.042
Download Resources	TP class	0.034

Table 3 **Data Correlation Results of the Improved G-P Algorithm.**

Association Rule ID	Frequent Item Sets	Minimum Confidence
Profession	Computer, electronics, communications	0.039
Education	Master PhD	0.033
Search preferences	Electronic journals	0.053
Download Resources	TP class	0.039

by the R/S analysis method is more consistent with the curve of the original time series, which indicates that the time series has a strong continuous effect. It is predicted that the series will continue to develop according to the existing state.

In order to further compare the improved saturation correlation dimension method with better correlation characteristics, the data in the library is analyzed to obtain the data correlation results obtained by the original G-P algorithm and the improved G-P algorithm, as shown in Tables 2 and 3 shows.

From the data in Tables 2 and 3, it can be concluded that both the original G-P algorithm and the improved G-P algorithm complete the association of library user data. In the case of the same data sample, the minimum confidence calculated by the improved G-P algorithm is higher. Therefore, it can be concluded from the minimum confidence that the performance of the improved G-P algorithm is better. As shown in Tables 2 and 3, the frequent item sets associated with this user's profession are computer, electronics, and communications. Judging from the user's access habits, the user has the highest minimum confidence among the three majors, and it can be roughly judged that the user belongs to one of the three majors. In terms of academic qualifications, the minimum confidence of the original G-P algorithm and the improved G-P algorithm are not much different, and the minimum confidence obtained is low, indicating that the probability of judging that the user is a master and a doctor is not large from the user's access habits. The user's education status is not obvious according to the visit. From a search preference perspective, the user has a preference for electronic journals for a period of time, with the smallest confidence level being the highest. In terms of download resources, TP-type e-journals are downloaded more. According to the saturated correlation dimension method, users' professional and academic qualifications, search preferences, and download resources can be prejudged, so that the corresponding resources can be supplemented and supplied in a timely manner. And according to the user's retrieval situation, record the user's search preferences in order to provide users with personalized services.

4.3. Comparative Analysis of Time Efficiency of G-P Algorithm Before and After Improvement

The G-P algorithm and the improved G-P algorithm are used to calculate the correlation integral for a certain time series of the same data amount, and to count the running time of the program. Among them, the embedding dimension $m = 2, 3, \dots, 20$, the number of neighborhood radius values $K_r = 20$, the statistical time is the average value of the program running five times under the same conditions, and the number of time series samples N is 100. 200, 400, 1000, and 2000, let optimization ratio = run time after improvement / run time before improvement, the statistical results are shown in Table 4.

As can be seen from Table 4, when the number of sample data is 100, the optimization ratio is 0.017. When the sample data is 200, the optimization ratio is 0.009. When the sample data is 400, the

Table 4 Comparison of G-P Algorithm Running Time Before and After Improvement.

	N	**100**	**200**	**400**	**1000**	**2000**
Time/s	Original G-P algorithm	2.261	4.158	19.250	233.111	1269.667
	Improved G-P algorithm	0.052	0.079	0.308	2.098	7.618
	Optimization ratio	0.023	0.019	0.016	0.009	0.006

optimization ratio reaches 0.006. It can be seen that the running time of the improved G-P algorithm is much shorter than that before the improvement. The original complex program can quickly obtain the results. As the sample data increases, the smaller the optimization ratio, the more significant the improvement effect of the algorithm. When the number of samples is increased to 1000, the time taken to use the improved method is greater than 1 min, and it is very easy to cause the program to be interrupted and the results cannot be obtained. The time taken to use the improved G-P algorithm is only 2.098 s. When the number of samples is 2000, the time it takes for the improved G-P algorithm to obtain results is still less than 10 s. It can be seen that the improved G-P algorithm in this paper has higher time efficiency than the original G-P algorithm.

4.4. Comparative Analysis of the Performance of the Improved G-P Algorithm and Different Algorithms

In order to seriously improve the effectiveness of the improved algorithm in this paper, different algorithms are used to analyze the intent data of library users' search literatures, and to predict the literatures finally selected by users. The evaluation standards used in this paper use accuracy, coverage, and recall to measure their performance in off-line calculations. The evaluation rules for each indicator are as follows:

Precision: The first 30 documents calculated from the training sample are compared with those of the user's actual behavior in the test sample to calculate the proportion of correct documents. The evaluation method is shown in Eq. (19).

$$P = tp/(tp + fp). \tag{19}$$

Among them, the symbol tp is the number of documents that the model calculates correctly, and the symbol fp is the number of documents that the model does not calculate correctly.

Table 5 Forecast Results of Different Algorithms.

Methods	**Precision**	**Coverage Rate**	**Recall**
Naive Bayes	0.724	0.691	0.146
KNN algorithm	0.783	0.734	0.151
Literature[9]	0.698	0.581	0.132
Literature[14]	0.761	0.629	0.147
Literature[15]	0.838	0.791	0.169
Literature[16]	0.901	0.878	0.195
Original G-P algorithm	0.857	0.819	0.186
Improved G-P algorithm	0.927	0.901	0.237

Coverage rate: The popularity of the 30 recommended documents calculated from the training samples is compared with the popularity of the entire collection of documents to evaluate the computing model's ability to mine non-popular documents. It can be calculated using the information entropy equation in information theory, as shown in Eq. (20).

$$H = -\sum_{l=1}^{n} p(l)\log(p(l)). \tag{20}$$

Among them, the symbol $p(l)$ is the sum of the popularity of document l divided by the popularity of the entire document.

Recall: Compare the literature of the actual behavior of the user in the test sample with the list of literature recommendations calculated by the algorithm through the training sample to calculate the accuracy of the recommendation list. The evaluation method is shown in Eq. (21).

$$R = tp/(tp + fn). \tag{21}$$

Among them, the symbol fn is the number of uncalculated documents in the recommendation list.

The improved G-P algorithm and comparison algorithm in this paper are used to calculate the training samples, and the recommended list obtained is verified by the evaluation equation. The results are shown in Table 5 below.

From the experimental results in Table 5, it can be seen that by comparing with the algorithms such as Naive Bayes and KNN algorithm in the ideal state, the accuracy and recall of the improved G-P algorithm in this paper have achieved good results, it is 0.927 and 0.901, which can better analyze and predict the user's intention. The higher the coverage rate, the better the algorithm can mine non-popular literature. From the results in Table 5, it can be seen that the improved G-P algorithm proposed in this paper achieved the highest coverage, which was 0.237. Through the above analysis, the effectiveness and accuracy of the improved G-P algorithm can be verified, and the user's intention can be predicted and mined.

5. CONCLUSION

Time series mode is an important research content in data mining. One of its main purposes is to use the value of one or more time series in the past to discover the value of a future series or predict its development trend. This paper introduces the fractal theory to analyze the time series in the field of data mining. The fractal characteristics of the time series are studied. The fractal theory can be used to find the law of change from the time series with fractal characteristics, predict the development trend, and achieve the prediction of user intentions. And dig. First, the method of array storage instead of repeated calculations is used to improve the method of saturated correlation dimension. Secondly, the Hurst exponent of the time series is obtained by the variable scale range analysis method. Finally, a fractal model for predicting user intent in short time series is established using the accumulation and transformation method. The experimental results show that the use of fractal theory can effectively describe the relevant characteristics of time series, the development trend of user intentions can be mined from big data, and the prediction model for short time series can be established to achieve the information mining of user intentions; in order to solve short time series, the model establishment and prediction problem provide a new method.

ACKNOWLEDGMENT

This work was supported by Natural Science Foundation of Hubei Province of China (Grant No. 2018CFB681).

REFERENCES

1. R. M. Safari, A. M. Rahmani and S. H. Alizadeh, User behavior mining on social media: A systematic literature review, *Multim. Tools Appl.* **78** (2019) 33747–33804.
2. S. Wenninger, D. Link and M. Lames, Data Mining in Elite Beach Volleyball — Detecting tactical patterns using market basket analysis, *Int. J. Computer Sci. Sport* **18**(2) (2019) 1–19.
3. A. J. Amutha, R. Padmajavalli and D. Prabhakar, A novel approach for the prediction of treadmill test in cardiology using data mining algorithms implemented as a mobile application, *Ind. Heart J.* **70**(4) (2018) 511–518.
4. M. Dateu and K. Seidel, Image information mining: Exploration of Earth observation archives, *Geograph. Helv.* **58**(2) (2018) 154–168.
5. S. Wu, A traffic motion object extraction algorithm, *Int. J. Bifur. Chaos* **25**(14) (2015) 1540039.
6. T.-A. Hoang and E.-P. Lim, Modeling topics and behavior of microbloggers: An integrated approach, *ACM Trans. Intell. Syst. Technol.* **8**(3) (2017) 1–37.
7. S. Wu, M. Wang and Y. Zou, Research on internet information mining based on agent algorithm, *Fut. Gen. Comput. Syst.* **86** (2018) 598–602.
8. H. Xixu, C. Leiting and Z. Min, A method to construct Weibo user behavior relationship network using dynamic cognition, *J. Univ. Electron. Sci. Technol. China* **47**(2) (2018) 262–266.
9. S. Wu, M. Wang and Y. Zou, Bidirectional cognitive computing method supported by cloud technology, *Cogn. Syst. Res.* **52** (2018) 615–621.
10. N. Laleh, B. Carminati and E. Ferrari, Risk assessment in social networks based on user anomalous behaviors, *IEEE Trans. Dependable Secure Comput.* **15**(2) (2018) 295–308.
11. D. Josheski, E. Karamazova and M. Apostolov, Shapley–Folkman–Lyapunov theorem and Asymmetric First price auctions, *Appl. Math. Nonlinear Sci.* **4**(2) (2019) 331–350.
12. C. Rojas and J. Belmonte-Beitia, Optimal control problems for differential equations applied to tumor growth: State of the art, *Appl. Math. Nonlinear Sci.* **3**(2) (2018) 375–402.
13. S. Shirakol, M. Kalyanshetti and S. M. Hosamani, QSPR analysis of certain distance based topological indices, *Appl. Math. Nonlinear Sci.* **4**(2) (2019) 371–386.
14. S. Wu, Nonlinear information data mining based on time series for fractional differential operators, *Chaos* **29** (2019) 013114.
15. C. Jianjiang, Y. Yuan and Xu Yingqiang, An Analytical model of loading-unloading contact between rough surfaces based on fractal theory, *J. Xi'an Jiaotong Univ.* **52**(3) (2019) 98–110.

16. L. Sijia, T. Juncui and Y. Boli, Multifunctional and multiband fractal metasurface based on inter-metamolecular coupling interaction, *Adv. Theory Simul.* **2**(8) (2019) 1–12.
17. L. Zhou, G. Du and D. Tao, Clustering multivariate time series data via multi-nonnegative matrix factorization in multi-relational networks, *IEEE Access* **6** (2018) 74747.
18. S. Epskamp, L. J. Waldorp and R. Mõttus, The Gaussian graphical model in cross-sectional and time-series data, *Multivar. Behav. Res.* **53**(2) (2018) 1–28.
19. H. Mao, L. Jiao and S. Gao, Surface quality evaluation in meso-scale end-milling operation based on fractal theory and the Taguchi method, *Int. J. Adv. Manuf. Technol.* **91**(1–4) (2017) 657–665.
20. Y. Yang, T. Sun and L. Zhang, Fractal mechanism of spatial distribution of arable land quality in Beijing–Tianjin–Hebei region, *Trans. Chin. Soc. Agric. Machinery* **48**(2) (2017) 165–171.
21. Y. Yuan, L. Gan and K. Liu, Elastoplastic contact mechanics model of rough surface based on fractal theory, *Chin. J. Mech. Eng.* **30**(1) (2017) 1–9.
22. R. Li, Q. Wang and X. Wang, Relationship analysis of the degree of fault complexity and the water irruption rate, based on fractal theory, *Mine Water Environ.* **36**(1) (2015) 1–6.
23. J. Ding, J. Wang and Z. Ye, Fast growth entire functions whose escaping set has Hausdorff dimension two, *Chin. Ann. Math.* **40**(4) (2019) 481–494.
24. J. Zhou and X. Wu, Study on the property of correlation dimension of sleep apnea syndrome electroencephalogram, *J. Biomed. Eng.* **34**(2) (2017) 168–172.
25. M. V. O. Araujo and A. B. Celeste, Rescaled range analysis of streamflow records in the São Francisco River Basin, Brazil, *Theor. Appl. Climatol.* **135**(1) (2018) 1–12.