ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research, volume 233
3rd International Conference on Contemporary Education, Social Sciences and Humanities (ICCESSH 2018)

GM(1,1) Based Analysis of National Research Tendency on Hotspot of Library's Data Service in Last Decade

Yue Chen
Library
Jianghan University
Wuhan, China

***Abstract*—With the amount of papers from 2008-2017 year in China, which retrieved by terms of "data service" and "library", this paper applies metabolic GM (1, 1) prediction model to explore variation trend of the hot-spot in the big data era. The result indicates that the proportion of paper's amount based on "data service" to total refers to "library", keeps ever-increasing in the ration of 18.68% since 2013.**

Keywords—data service; grey system theory; metabolic GM(1,1)

I. INTRODUCTION

Computers play an irreplaceable role in the ternary world formed by humans, machines and objects. And they are silent forces for social research has entered into the fourth paradigm. Both the change. With the digital age evolving to the data age, scientific community and the individuals have a strong thirst for more in-depth, more convenient knowledge services. Under the urge of academic users, the research libraries are actively promoting the user experience of knowledge service, which will help users to expand knowledge discovery, and reveal the knowledge contained in the literature. Knowledge service can build the link among the knowledge and formulate the knowledge maps. Content-based knowledge service is the inevitable choice to break through the bottleneck of knowledge acquisition in the context of machine learning prevalence [1].

What is Data Services in library, it helps researchers and readers with data and statistical resources though out the entire research life-cycle, as both consumers and producers of data. The service items could include the follows:

A. Identify & Access Data and Statistics

- E-Resources and Research Guides
- Research Help: Ask a Librarian
- Instructional Support
- Data Acquisition

B. Create & Manage Their Data

- Research Data Management Guidance
- Digital Text Services

C. Data Literacy Education

- Primary Stage: Data Science
- Practical Stage: Data Related Skills
- Advanced Stage: Data Mining

The advent of data-intensive scientific paradigm has evoked researchers' needs for scientific data service [2]. Management of research data is increasingly seen as an important role for academic libraries and university libraries [3]. Nowadays data of managing scientific research are spread across different systems and lack of being organized. Research information management （RIM） now is drawing attentions for research funding agencies, universities and libraries [4].

This paper took paper amounts which research on "data service" and "library" as analysis and modeling objects. Firstly, took Knowledge Tendency Analysis on the hotspot; secondly, took full text literacy database as data source, collected and analyzed the data since 2008 year by Excel; thirdly, applied gray system theory [5,6] to explore transformation tendency of the hot-spot in the big data era; and summarized the explored research tendency which on the relevant hot-spot.

II. HIT RATE ANALYSIS OF THE HOT-SPOT

Entered the interface of "Knowledge Trend Analysis" by logging in the Wanfang data platform, we took "research data service" as search term, so the hit term could be got from each year during 2008-2017. In this way, we could find out the transformation tendency of the relevant hot-spot with the hit rates, and pave the way for the quantitative analysis and trend mining. The transformation process was showed in "Fig. 1".

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

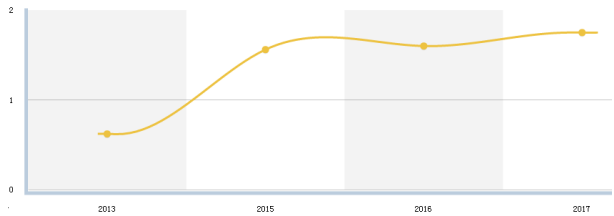

Fig. 1. Knowledge trend analysis curve of hot-spot on “research data service” during 2013-2017.

The paper took CNKI as data source, the sample span was from 2008 to 2017 year. Curve of “Data Service” and “Library” relevant paper quantities during 2008-2017 was showed in "Fig. 2". Taking statistics with these data by Excel, we analyzed the proportion of paper’s amount based on “data service” to total refers to “library”, statistical results were showed in "Fig. 3", we select historical data of 2008-2017 years as modeling objects.

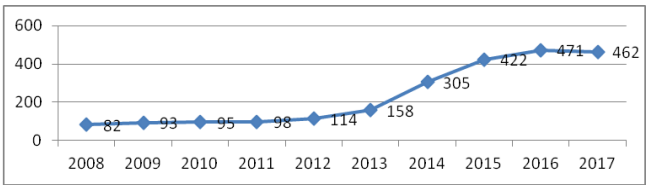

Fig. 2. Paper quantities of “data service” and “library” during 2008-2017.

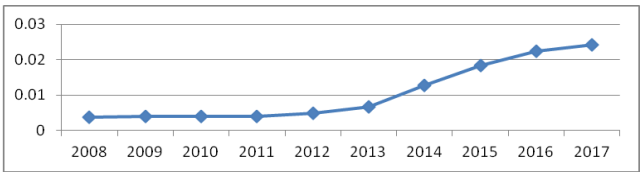

Fig. 3. The Proportion Of Paper’s Amount Based On “Data Service” To Total Refers To “Library” During 2008-2017.

III. MODELING WITH GRAY SYSTEM THEORY

In the natural and social systems, the problems of uncertainty are widespread. Systems with lots of samples can be analyzed by probability and statistics. In other cases, fuzzy mathematics can be used to solve the problem. However, due to the lack of samples, poor information and lack of experience, there are also problems with uncertainty, which can be described in gray model(GM), when the processes are complex and indescribable with precision and accuracy through the use of mathematical models. The gray system theory was originally proposed by Chinese researcher Deng (1982), and developed by Liu Sifeng.

According to the central limit theorem, probability theory and mathematical statistics which with large sample data, the traditional statistical model needs to meet the requirements of large sample. However, due to accelerated development of the IT, academic research on data service was proposed on the last decade. In practical applications, the traditional mathematical statistical model has certain limitations to the amount of samples, but fortunately, gray theory model requires only a small number of samples, simple calculation, but have higher adaptability and is more reliable. In this way, the gray theory makes up for the problem of fewer samples in the modeling process under the conditions of "small sample and poor information". For this reason, the approach based on gray system theory is appropriate for forecasting the proportion of paper’s amount based on “data service” to total refers to “library”.

The processes of proportion prediction with GM(1,1) model are presented as follows:

- Assume that $x^{(0)}=\{x^{(0)}(1),x^{(0)}(2),\mathrm{L}\ ,x^{(0)}(n)\}$, which collected from the historical proportion of the first n years, is a given sequence of raw data.
- Applying 1-AGO(Accumulated generating operation) on $x^{(0)}$, provided that $x^{(1)}=\{x^{(1)}(1),x^{(1)}(2),\mathrm{L}\ ,x^{(1)}(n)\}$, where $x^{(1)}(i)=\sum_{k=1}^{i}x^{(0)}(k),i=1,2\mathrm{L}\ ,n$.
- Apply a consecutive neighbor mean generation to $x^{(1)}$. Let $z^{(1)}(k)=\frac{1}{2}x^{(1)}(k-1)+\frac{1}{2}x^{(1)}(k),k=2,3,\mathrm{L}\ ,n$, then it follows that $z^{(1)}=\{z^{(1)}(2),\ z^{(1)}(3),\ \mathrm{L}\ ,\ z^{(1)}(n)\}$.
- Construct a white differential equation based on grey system theory: $\frac{dx^{(1)}}{dt}+ax^{(1)}=u$, where a means the developing coefficient, u the grey input. Solving the equation, the model response can be given by

  $$x^{(1)}(t)=(x^{(1)}(t_0)-\frac{u}{a})e^{-a(t-t_0)}+\frac{u}{a},$$

  where t_0 is the initial time moment. If we sample the equal-time-interval discretely, the time response obtained, which follows that

  $$x^{(1)}(k'+1)=(x^{(1)}(1)-\frac{u}{a})e^{-ak'}+\frac{u}{a},$$

  where k' is the sampled time moment, and valued as positive integer from 1.
- Perform a least squares estimate for the parameters (a and u). The simulated parameters ($\hat{a}$、$\hat{u}$) can be obtained that

  $$\begin{bmatrix}\hat{a}\\ \hat{u}\end{bmatrix}=\left(B^TB\right)^{-1}B^TY_n,$$

  where

  $$B=\begin{pmatrix}-z^{(1)}(2), & 1\\ -z^{(1)}(3), & 1\\ \mathrm{L} & \mathrm{L}\\ -z^{(1)}(n), & 1\end{pmatrix},$$

  $$Y_n=(x^{(0)}(2),\ x^{(0)}(3),\ \ldots\ ,\ x^{(0)}(n))^T.$$

  At this time,

If $|\hat{a}| \le 0.3$, the relative mid or long term forecasting can be conducted.

If $0.3 < |\hat{a}| \le 0.5$, the short term prediction is preferred.

If $0.5 < |\hat{a}| < 0.8$, put the $x^{(1)}$ value back to $x^{(0)}$ value, then back to the step 2).

If $|\hat{a}| \ge 0.8$, should update the raw data, the smoothing process can be used here for example, then back to the step 2).

- Put the simulated parameters ($\hat{a}$、$\hat{u}$) back to the differential equation in step 4), the response is

$$\hat{x}^{(1)}(k'+1) = (x^{(1)}(1) - \frac{\hat{u}}{\hat{a}})e^{-\hat{a}k'} + \frac{\hat{u}}{\hat{a}}$$, and $\hat{x}^{(1)}(1) = x^{(0)}(1)$.

When the time with $k' = 1, 2, \mathrm{L}\ , n-1$, the estimated sequence is called a simulated value of grey model;

When the time with $k' \ge n$, the estimated sequence is called a forecasting value of grey model.

- Evaluate the errors and precision. Assume that $\varepsilon(k')$ is the relative error, $\bar{\varepsilon}$ is the average relative error, and the precision is τ:

$$\varepsilon(k') = \frac{x^{(1)}(k') - \hat{x}^{(1)}(k')}{x^{(1)}(k')} \times 100\%$$, $$\bar{\varepsilon} = \frac{1}{n-1}\sum_{k'=2}^{n}|\varepsilon(k')|$$, $\tau = 1 - \bar{\varepsilon}$.

- Examine the evaluated precision, if less than the pre-set threshold (in the other words, it's not meeting a predetermined requirement), it needs to reduce possible errors, which caused in reciprocating operations, by constructing the remnant GM(1,1) model. The detail process follows that establish a GM(1,1) model using the error sequence

$$\varepsilon'^{(1)}(k') = x^{(1)}(k') - \hat{x}^{(1)}(k')$$

- firstly, and then add the estimated sequence $\hat{\varepsilon}'^{(1)}$ of remnant GM(1,1) model to the $\hat{x}^{(1)}$, repeat the modification process, until the precision meet the requirement.
- Restore the final estimated sequence through inverse accumulating (IAGO), the number of IAGO times should be equal to the AGO times. Therefore,

$$\hat{x}^{(0)}(k') = \hat{x}^{(1)}(k') - \hat{x}^{(1)}(k'-1)$$

is the sequence of proportion predicted (Y_e).

- Construct a metabolic GM(1,1) model, which built on the following new sequence

$$x^{(0)} = \{x^{(0)}(2), \mathrm{L}\ , x^{(0)}(n), x^{(0)}(n+1)\}$$

obtained by inserting $x^{(0)}(n+1)$ and deleting $x^{(0)}(1)$. As a matter of fact, as time goes on, some stochastic interferences or driving forces are concerned with the development of grey system once a metabolic GM(1,1) model established, therefore, the higher accuracy can be achieved.

In the primitive GM(1,1) model modeling, the past data from the real time $t = n$ is used. However, the development of any grey system, as time goes on, will continue to have some random disturbance factors into the system, the progression of the system impacted. Therefore, with the primitive GM(1,1) model, the higher accuracy is only with a few recent data, deviate from reality, poorer effect of the prediction.

In order to impair the disturbance of the future random disturbance on grey system and improve the forecasting accuracy, the GM(1,1) model is reformed.

In the original data $x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \mathrm{L}\ , x^{(0)}(n)\}$, the latest information $x^{(0)}(n+1)$ is placed and the oldest data $x^{(0)}(1)$ and $x^{(0)} = \{x^{(0)}(2), x^{(0)}(3), ..., x^{(0)}(n+1)\}$ is removed. Same with the above steps, the model grey metabolic GM (1,1) model is established, and a series of forecasting data is inferred out, at the same time, the precision of metabolic GM(1,1) is higher than that of primitive GM(1,1) model.

IV. CONCLUSION

Due to the late start of the academic research on library's data services, it can be seen from figure 1 and figure 2 that the research on library data service has rapidly developed in the last five years, so we have few numbers of the related papers.

In this way, there's only gray model fits for the prediction which requires less sample, simple calculation, and makes up the problems in paper quantity modeling under "little sample and poor information" conditions.

Firstly, we take GM(1,1) proportion modeling with historical data of 2008-2017 years, the average simulation error up to 17.54%, obviously, it does not meet the conventional requirements of precision. Considering with theorem of "recent data is most useful" from gray system theory, we could establish a metabolic GM(1,1) model, metabolize the last data successively. In this way, the average simulation error is 15.19% when modeling with historical data of 2009-2017 years; the average simulation error is 17.21% when modeling with historical data of 2010-2017 years; the average simulation error is 16.5% when modeling with historical data of 2011-2017 years; the average simulation error is 12.76% when modeling with historical data of 2012-2017 years; the average simulation error is 6.8% when modeling with historical data of 2013-2017 years.

In order to improve accuracy, we finally selected historical data from 2013-2017 years, coefficient vector of differential equation was calculated out as

$\hat{a} = [a, u]^T = [-0.186832, 0.0118]^T$, and time response function of differential equation is

$$\hat{X}^{(0)}(k+1) = 0.069839 \exp(0.186832 * k) - 0.063158 .$$

There were 5 sampled data, so we took 3-step forecasting, the predicted results were 0.030290, 0.036513 and 0.044013. In other words, proportion of paper's amount based on "data service" to total refers to "library" each year in 2018, 2019 and 2020 year would be 3.029%, 3.6513% and 4.4013%.

According to the simulated and predicted results, we could find that proportion of "data service" is taking a steady sustained growth, nearly at the rate of 18.68% since 2013 year.

Scientific data is gradually becoming an important information resource supporting scientific activity of university in the digital research environment, and scientific data management and service are gradually becoming an important content of subject service in university library.

ACKNOWLEDGEMENT

This research was sponsored by Hubei province education department (Humanities and Social Science Projects NO. 18g044).

REFERENCES

[1] Yin Zhang. The Trend and Practice of Data Service in American Research Libraries, Library and Information Service [J], 2017(9): 33-41.

[2] Xin Huang, Zhonghua Deng. The Service Models of Library's Sientific Data Under the View of "Internet+", Information Studies：Theory & Application [J], 2017(3): 75-80.

[3] Chen Yuanyuan, Ke Ping. The Enlightenment of Research Data Services in European Academic Research Librariesto Chinese University Libraries. Library and Information Service [J], 2017(6): 73-78.

[4] Yang Helin. The Development of Data Services in University Libraries for RIM and Its Enlightenment. Library and Information Service [J], 2015(21): 83-89.

[5] Deng Julong. Grey Control System [M]. Wuhan：Huazhong Institute of Technology Press, 1985.

[6] Liu Sifeng, Xie Naiming. Gray System Theory and Its Applications [M].Beijing：Science Press, 2008(4).