Anlysis and Design of Activity Degree Monitoring Algorithm

line 1: Yinghao Du
line 2: R&D Department of Tieba
line 3: Baidu Online Network Technology (Beijing) Co.,Ltd.
line 4: Beijing, China
line 5: duyinghao@baidu.com

line 1: Xuebing Wang
line 2: R&D Department of Tieba
line 3: Baidu Times Technology (Beijing) Co.,Ltd.
line 4: Beijing, China
line 5: wangxuebing01@baidu.com

line 1: Zhihai Lei
line 2: R&D Department of Tieba
line 3: Baidu Online Network Technology (Beijing) Co.,Ltd.
line 4: Beijing, China
line 5: leizhihai@baidu.com

line 1: Yiran Li
line 2: R&D Department of Tieba
line 3: Beijing Baidu Netcom Science and Technology Co.,Ltd.
line 4: Beijing, China
line 5: liyiran01@baidu.com

line 1: Bin Hu
line 2: R&D Department of Tieba
line 3: Baidu Online Network Technology (Beijing) Co.,Ltd.
line 4: Beijing, China
line 5: hubin13@baidu.com

line 1: Guang Li
line 2: R&D Department of Tieba
line 3: Baidu Online Network Technology (Beijing) Co.,Ltd.
line 4: Beijing, China
line 5: liguang03@baidu.com

***Abstract*—DAU (Daily Active User) is the number of daily active users, often used to reflect the operation of websites, Internet APPs and games [1]. DAU usually counts the number of users who have logged in or used a product (removing users who are repeatedly logged in) within one day (statistical day), which is similar to the concept of visitors (UV) in the traffic statistics tool. As we all know, the revenue source of some Internet applications lies in the revenue of advertising, and the amount of advertising revenue depends on the size of DAU. Therefore, the design strategy and algorithm to monitor the fluctuations of DAU can better help people analyze and improve our products, thus bringing improvements to the products. Therefore, this paper will design a variety of algorithm construction models to monitor the fluctuation of DAU, and achieve alarm announcement, analysis and location of the abnormal fluctuation of DAU, so as to explore of the value of DAU.**

Keywords—DAU, fluctuated alarm, week-on-week, wave angle, upstream positioning

I. Introduction

At present, the Internet industry is developing rapidly, and applications associated with it have become closely related to everyone and become an indispensable part of people's lives. For some applications that create value through traffic, traffic is everything. Therefore, monitoring and analyzing traffic changes and intrinsic value of traffic are particularly important to better help us analyze and optimize products. . User activity (for example, DAU daily active users, MAU monthly active users and other indicators) as one of the most important indicators to measure traffic changes in Internet products, indicating the frequency of user interaction using products, also reflects satisfaction with the product [2].

In the actual situation, the users' activity is generally regular, whether the product is in the growth period, stable period, or recession period. If there are external factors, such as, the launch of new product functions, holidays, product PUSH, user cheating, hacking, client anomalies, etc., various degrees of fluctuations will be triggered of the user activity indicators. Some of these fluctuations are positive and some are negative. Therefore, in the real world, we hope to monitor and analyze the fluctuations of user activity caused by external factors, and explore the intrinsic value of fluctuations to help people optimize products [3].

This paper will build a monitoring algorithm with DAU as the monitoring target in user activity, and realize the alarm and location of abnormal problems by analyzing the fluctuation of DAU. This paper will introduce the construction idea of the algorithm from the aspects of alarm and positioning. In the aspect of alarm, the general cycle-to-loop ratio algorithm is abandoned, the curve angle algorithm is proposed to compare the week, and the historical backtracking algorithm is introduced to monitor the continuous fluctuation. In terms of positioning, this paper constructs a product interface network and introduces two important parameters of computing interface correlation: conversion rate and impact factor. When abnormal fluctuation of the interface occurs, the interface is backtracked upstream through correlation, and the analysis and location will be ultimately completed.

II. Monitoring Algorithm Analysis

The current monitoring algorithm for DAU is mainly performed by monitoring the change of the cycle-to-cycle ratio of the interface uv by hour. This is a coarse-grained monitoring method, which uses rough numerical differences for data comparison, resulting in a large number of false positives for interface fluctuations. Therefore, this section will start from the coarse-grained DAU monitoring algorithm, gradually optimize and improve the shortcomings and deficiencies in the algorithm, and reduce the false positive rate of the algorithm for the abnormal problem while ensuring that the false negative rate is at a lower level, and introduces an abnormal judgment on the upstream interface to locate the source that causes the interface to fluctuate.

A. Abnormal judgment

1) Week-on-week

The meaning of the cycle-to-cycle ratio can be defined as the comparison of the uv of the current hour interface with the uv of the same hour of the previous day to determine whether the uv of the interface has fluctuated. It is undeniable that the weekly cycle can reflect the fluctuation of the interface to a certain extent, but there will be a large number of false positives. The reason is that the uv of the interface will have a certain degree of natural fluctuation with time, and the cycle-to-cycle ratio will treat this natural fluctuation as an abnormal situation. Therefore, in order to reduce the false positive rate, this paper introduces a week-on-week ratio.

Through the analysis of the data, it can be found that the uv change of the interface exhibits a certain periodicity, that

2330-2186/19/$31.00 ©2019 IEEE
DOI 10.1109/CloudCom.2019.00062

is to say, the change of time as a hidden factor affects the natural fluctuation of uv. This periodicity is expressed in days and in weeks. As shown in the figure, the uv change of an interface, Figure 1 shows the change of the uv-hour level of an interface for two days, and Figure 2 shows the change of the uv-hour level one week apart.

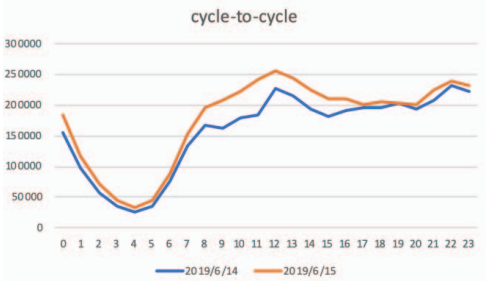

Fig. 1. Cycle-to-cycle uv comparison

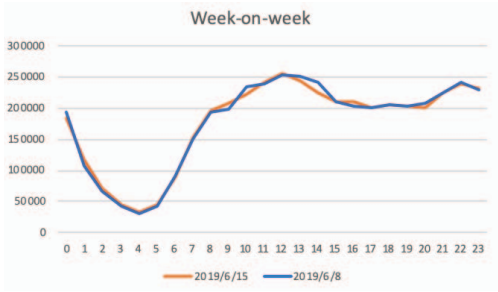

Fig. 2. Week-on-week uv comparison

It can be seen from the comparison between Figure 1 and Figure 2 that the year-on-year comparison with the cycle-to-cycle ratio has better fitting, which can better eliminate the natural fluctuation caused by time, thus achieving the purpose of reducing the false alarm rate.

However, there is only one problem compared with the previous week. If the data of the previous week is abnormal and the data is normal this week, this will cause the normal data to be alarmed as abnormal data. The reason is that the accuracy of the last week's data used as a baseline cannot be guaranteed. Therefore, this paper introduces the week-on-week data of the previous four weeks and obtains the baseline for judging abnormalities based on the data between the four weeks. The easiest way is to use the average of the previous four weeks, but the average is susceptible to extreme values and causes a large deviation. In the case where most of the normal data is abnormal data, the algorithm sets the median value of the data of the previous four weeks as a baseline to make an abnormality determination.

2) Difference calculation

The so-called difference calculation is how to calculate the uv change in the monitoring process of uv to judge whether there is an alarm or not. Some current calculation methods include: simple numerical difference, rate of change, and so on. The method of simple numerical difference is too simple, and the data is not normalized and has no universality. Therefore, the rate of change is chosen as a means of calculating fluctuations. The rate of change θ at this stage is mainly calculated by the method of (1):

$$\theta = (d_t - d_{t-1})/d_{t-1} \tag{1}$$

Among them, d_t、d_{t-1} is the uv value at the current time and the same time last week. The essence of this method is to compare the changes in the values, but only the normalization process, and does not take into account the fluctuations of the interface uv during this hour. Therefore, this section proposes a curve-flip algorithm to calculate fluctuations based on (1). Its calculation method is as shown in (2):

$$\theta = tan^{-1}\left(\frac{d_t - d_{t-1}}{x_{st}}\right) - tan^{-1}\left(\frac{d'_t - d'_{t-1}}{x_{st}}\right) \tag{2}$$

d_t、d_{t-1}、d'_t、d'_{t-1} respectively indicate the value of uv at the moment of the day, the value of uv at the previous moment of the day, the value of the same time on the same day of last week, and the value of the previous day of the same day last week. x_st is the standard unit of the axis by calculating the maximum and minimum values of the uv at that day. The variation angle can be calculated by (2), as Curve angel.

Equation (2) calculates the angle between the two lines in Figure 3, which represents the fluctuation of the uv change rate at this time of the week compared to the last week. When this fluctuation exceeds the specified threshold, an alarm is issued through the algorithm to achieve the effect of monitoring fluctuations.

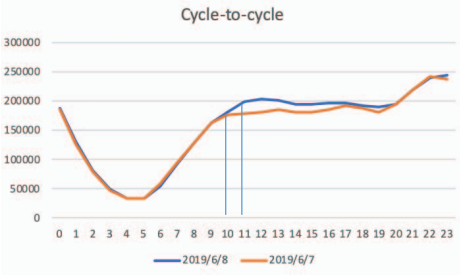

Fig. 3. Curve angel

3) Historical data backtracking

The occurrence of anomalies in real life can be roughly divided into two situations: one is an emergency problem, which is characterized by a sharp change in uv in a short period of time; the other is an abnormality in non-emergency small fluctuations. The characteristic of this situation is that the change in a short time is small, but it varies greatly. The method of the year-on-year proposed in this paper in A-1 can only detect the dramatic changes of uv in the period of time, and make an abnormal alarm for urgent emergencies. In order to address persistent non-emergency small fluctuation anomalies, this section proposes a method of backtracking historical data for monitoring.

Set the threshold for determining whether the fluctuation is abnormal or not is t, the cumulative fluctuation a of the first 8 hours of the interface is obtained by (3):

$$a = a_1 + a_2 + \cdots + a_8 \tag{3}$$

Therefore, when the judgment interface of $t < c \pm a$ has a continuous non-emergency small fluctuation abnormality. Where $a_1 \sim a_8$ is the fluctuation of the interface for the first 8 hours, calculated by (2); c indicates the natural change of the product, when a product c in the rising period is a positive number, the product c in the stationary period is 0, the product c in the falling period is a negative number.

B. Abnormal positioning

When it is detected that the uv change of an interface is abnormally fluctuating, it is usually necessary to analyze and locate the problem. This section will design and analyze the problem through the upstream and downstream relationship between the interfaces, so as to facilitate better optimization and improvement of the product.

The sample product, Figure 4, is a common product core behavioral process that describes the main behavioral processes of this product user. The user entry is page 1, and the behavior a1 can be performed in page 1, or can be entered into page 2 through behavior b1, and behavior 3 is entered in page 3. These pages essentially represent one interface, and the user jumps through the interface through operations. According to the jump relationship of the interface, the upstream and downstream of the interface can be determined, thereby constructing a directed graph representing the interface relationship.

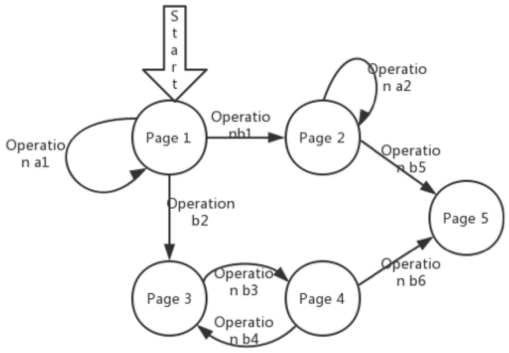

Fig. 4. User Behavior State Model

To quantify the relationship between interfaces, this section introduces three related concepts: conversion rate, impact factor, and correlation:

a) conversion rate: The upstream and downstream conversion rates of the interface indicate the impact of the upstream interface on the downstream interface. The specific performance is shown in Figure 5. For the user operating a1, 10% of the operations are performed, that is, the uv of the interface a1 has 10 % is up to b1, at which time it is defined that interface a1 is upstream of b1. The dashed line in Figure 4 indicates a weak association and the solid line indicates a strong association. The definition of weak association is that there is no necessary connection between the two operations. For example, two operations of the same level in a certain page, a1, b1, remain in the original page after the operation a1, and then perform the b1 operation. The uv of these two interfaces have a certain relevance, but the interfaces are not directly related, so they are related weakly.

b) impact factor: The impact factor of an interface indicates the extent to which the downstream interface is affected by an upstream interface. When there are multiple association operations on the upstream of an interface, different upstream association operations have different impacts on the downstream user behavior. The impact factor quantitatively describes the impact of an upstream interface on the downstream interface. As shown in Figure 5, the upstream interface of interface b3 has two b2 and b4. Through data statistics and analysis, 80% of uv of b3 interface comes from b2, 20% comes from b4, so the influence factor of b2 on b3 is 0.8, b4 The impact factor for b3 is 0.2.

c) correlation: The correlation between the upstream and downstream interfaces quantitatively represents the magnitude of the correlation between the two interfaces. The correlation is mainly calculated through the conversion rate x, the impact factory and the correlation degree c of the upstream and downstream of the interface. The degree of association indicates the relationship between two interfaces. For example, the weak association and strong association mentioned in a are represented by the size of the c value, and the degree of association between any two fixed upstream and downstream interfaces is a constant. Therefore, the correlation between the two interfaces can be calculated by (4):

$$cor = ax + by + c \tag{4}$$

a and b are two constants that represent the effect of conversion and impact factors on correlation.

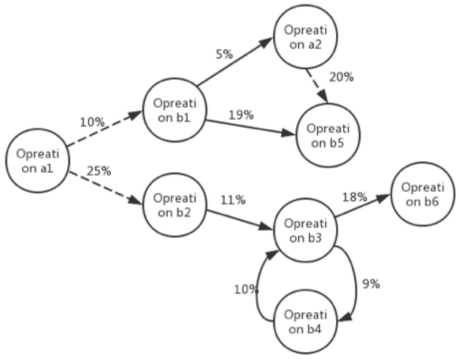

Fig. 5. User behavior conversion model diagram

The threshold value for judging whether the interface b6 is abnormal or not is set to be backtracked to the upstream interface b3 of b6 in order to locate the location where the abnormality occurs, and the threshold of the b6 abnormality caused by b3 is calculated by (5):

$$t_{n+1} = \frac{t_n}{cor_{b3b6}} = \frac{t_n}{ax+by+c} \tag{5}$$

a、b are two constants. c represents the degree of association between the two interfaces; x indicates the conversion rate between the two interfaces; y indicates the impact factor between the two interfaces; t_n indicating the threshold for the downstream interface to determine the fluctuation abnormality; and t_{n+1} indicates thresholds that fluctuation anomalies in the downstream interface are cause fluctuation anomalies in upstream interface.

Through (5), the fluctuation threshold for determining whether the b6 abnormality is caused by b3 can be calculated. If the fluctuation range of the b3 interface exceeds, the fluctuation abnormality of b6 can be considered to be caused by the fluctuation of the b3 interface. Then, according to (5), it is calculated whether the b3 fluctuation is caused by b2 or b4, and so on, and backtracking, and finally, the source interface that causes abnormal fluctuation can be located.

III. Monitoring Algorithm Design

Through the analysis in the second section, this section uses the hourly interface data as the monitoring target to build the DAU monitoring algorithm. The steps are as follows:

Step 1: First, the interface is initially filtered using the current coarse-grained method. Calculate the change of the interface uv in this hour, compare with the uv change of the same hour of the previous day and the uv of the same hour of the previous day of the previous day. If it is more than the threshold of the previous day or last week, then finer evaluation and positioning will be done; otherwise, such fluctuations are considered natural fluctuations;

Step 2: Obtain the uv value within the same hour of the same day of the previous four weeks, and use the median value to construct a baseline to determine whether there is an abnormality in the fluctuation of the previous time. The angle of the fluctuation change is calculated using the curve angle algorithm proposed in II-A-2 and compared with the threshold. When the change does not exceed the threshold, it is considered that no abnormality causing severe fluctuation occurs at this time, so the process proceeds to step 3 to perform the continuous fluctuation abnormality test; when the change exceeds the threshold, it is considered that the abnormal fluctuation occurs at this time, and then proceeds to step four for analysis and positioning.

Step 3: Calculate the angle of the fluctuation of the eight hours before the current hour using the curve angle algorithm. Also calculate according to the calculation method of step two. Use the median value of the uv value within the same hour of the same day of the previous four weeks to get the first eight hours per hour. The angle of the fluctuation is accumulated and compared with the threshold. If the cumulative fluctuation angle exceeds the threshold, it is considered that the previous 8 hours have a continuous fluctuation abnormality, and the step 4 is analyzed and positioned. If the cumulative fluctuation angle is not If the threshold is exceeded, the fluctuation of the interface for the hour is considered to be natural fluctuation.

Step 4: Use the upstream backtracking algorithm introduced in Section II-B Anomaly Location to calculate the threshold of the upstream interface abnormality through the threshold of the abnormal fluctuation interface and analyze the upstream interface. If the process proceeds from step 2 to step 4, the upstream interface is determined by using the method for determining the abnormality in step 2. If the process proceeds from step 3 to step 4, the upstream interface is determined by using the three methods for determining the abnormality. By analogy, the backtracking is reversed, and finally the source interface that causes the interface uv to fluctuate abnormally is located.

The above steps are shown in Figure 6:

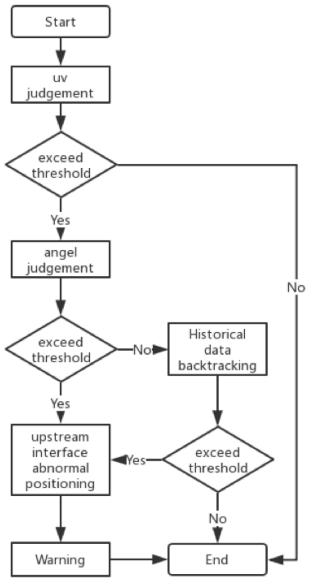

Fig. 6. Schematic diagram of the process

IV. Test Analysis

A. Test design

This test uses the model described above and applies it to the actual scenario. This article uses Baidu Post Bar as the application scenario, and selects the data of 100 interfaces running on Baidu Post Bar online for 30 days to test the pros and cons of the proposed model. The test will compare the performance of the following five models: 1. The current monitoring model based on the change of the cycle-to-cycle ratio uv; 2. The monitoring model based on the recent week-on-week uv change; 3. The most recent week-on-week data is used as the monitoring model for the judgment; 4. The monitoring model based on the change of the median angle of the past four weeks is used as the basis for judging; 5. Based on the model 4, the monitoring model of the historical backtracking algorithm is introduced. Based on the above five monitoring models, an upstream backtracking algorithm is introduced to locate the anomaly, and the accuracy of the positioning is verified by the test data.

This paper will use the precision P (accuracy rate), recall rate R (recall rate), and 、F_1、$F_\beta$criteria to make a judgment. The four standards are defined as follows [4]:

As shown in Table 1, the confusion matrix for the monitoring results of interface anomalies:

TABLE I. Monitoring result confusion matrix

The actual situation	Monitoring result	
	abnormal	Non-exception
abnormal	TP (true exception)	FN (false normal)
Non-exception	FP (false exception)	TN (true)

Then we can get the following formula:

Precision P (accuracy):

$$P = \frac{TP}{TP+FP} \tag{6}$$

Recovery rate R (recall rate):

$$R = \frac{TP}{TP+FN} \quad (7)$$

F_1, The essence is the harmonic average of the precision and recall ratio:

$$F_1 = \frac{2\times P\times R}{P+R} \quad (8)$$

F_β, the essence of which is the weighted harmonic average of the precision and recall:

$$F_\beta = \frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R} \quad (9)$$

$\beta > 0$ measures the relative importance of the recall rate to the precision. $\beta = 1$ means standard F_1; $\beta < 1$ means a greater impact on accuracy. For the anomaly monitoring model, the recall rate should take a greater weight than the precision, so in the F_β, the value of β is chosen to be 2[5].

B. Test results

The data of this experiment is 30 days of data of 100 interfaces from Baidu Tieba, and the total amount of data is 3000. The following is the confusion matrix of five models:

Model 1 is a monitoring model based on the change of the weekly cycle ratio uv. The experimental results show that the model has a high recall rate but a poor precision.

TABLE II. Model 1 Monitoring Results Confusion Matrix

The actual situation	Monitoring result	
	abnormal	Non-exception
abnormal	335	22
Non-exception	127	2527

Model 2 is a monitoring model based on the recent week-on-week uv change. The experimental results show that compared with model 1, the recall rate has decreased and the precision has increased.

TABLE III. Model 2 Monitoring Results Confusion Matrix

The actual situation	Monitoring result	
	abnormal	Non-exception
abnormal	329	17
Non-exception	45	2609

Model 3 is a monitoring model based on the change of the angle between the previous week and the week. The experimental results show that compared with Model 1 and Model 2, the recall rate has decreased and the precision has increased.

TABLE IV. Model 3 Monitoring Results Confusion Matrix

The actual situation	Monitoring result	
	abnormal	Non-exception
abnormal	311	35
Non-exception	27	2627

Model 4 is a monitoring model based on the change of the median angle of the past four weeks. The experimental results show that compared with the second and third models, the model has improved the precision and recall rate.

TABLE V. Model 4 Monitoring Results Confusion Matrix

The actual situation	Monitoring result	
	abnormal	Non-exception
abnormal	331	15
Non-exception	11	2643

Model 5 introduces a historical backtracking algorithm based on the model 4. The experimental results show that compared with the above model, the precision of the model is basically unchanged, and the recall rate is improved.

TABLE VI. Model 5 Monitoring Results Confusion Matrix

The actual situation	Monitoring result	
	abnormal	Non-exception
abnormal	342	4
Non-exception	13	2641

The precision of the five models, the recall rate R, and the results are as follows:

TABLE VII. The Results of all Models

	Precision rate P	Full rate R	F_1	F_β
Model 1	0.73	0.96	0.83	0.90
Mode 2	0.88	0.95	0.91	0.94
Model 3	0.92	0.90	0.91	0.90
Model 4	0.97	0.96	0.96	0.96
Model 5	0.96	0.99	0.97	0.98

It can be found from the above experimental results that Model 5 has the best performance when introducing various technical means, so the DAU monitoring algorithm will be constructed based on Model 5.

V. CONCLUSION

This paper constructs a DAU monitoring algorithm for monitoring, alarming, analyzing and locating DAU fluctuations. The purpose is to explore the intrinsic value of DAU fluctuations. In this paper, we propose the method of week-on-week, curve angle algorithm, historical backtracking algorithm and upstream backtracking algorithm to optimize the model precision and verify the model without using the model. It proves that the DAU monitoring model with multiple algorithms has better performance. The shortcomings of the model determined in this test are that the threshold selection for judging abnormal fluctuations is mostly determined by experience, and the interface threshold is only decomposed and set, and the threshold is not determined for the specific interface.

REFERENCES

[1] Gjoka M, Sirivianos M, Markopoulou A, et al. Poking facebook: characterization of osn applications[C]//Proceedings of the first workshop on Online social networks. ACM, 2008: 31-36.

[2] Nazir A, Raza S, Chuah C N. Unveiling facebook: a measurement study of social network based applications[C]//Proceedings of the 8th ACM SIGCOMM conference on Internet measurement. ACM, 2008: 43-56.

[3] Fields T, Cotton B. Social game design: Monetization methods and mechanics[M]. CRC Press, 2011.

[4] Batista G E, Prati R C, Monard M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM SIGKDD explorations newsletter, 2004, 6(1): 20-29.

[5] Tripathy A, Agrawal A, Rath S K. Classification of sentiment reviews using n-gram machine learning approach[J]. Expert Systems with Applications, 2016, 57: 117-126.