## CLAIMS

1. A method for pushing an application message, the method comprising:

detecting that a label of a message of a plurality of messages to be sent is associated to a non-real-time transmission (101);

dynamically determining, based on a timely-response rate and a correction scheme, an optimal time period for each user device in a user device group to respond to the message (102); and

separately sending the message to a corresponding user device in the user device group based on the optimal time period for each user device to respond to the message and based on performance of a mapping operation to improve pushing concurrency (103).

2. The method according to claim 1, wherein dynamically determining, based on the timely-response rate and a correction scheme, the optimal time period for each user device in the user device group to respond to the message comprises:

predetermining a time interval for collecting statistics on the timely-response rate (201, 301);

collecting statistics on a number of messages that are received by and a number of messages that are responded to in time by each user device in each time interval based on the predetermined time interval (202, 302);

determining a timely-response rate of each user device in each time interval based on the number of messages that are received by and the number of messages that are responded to in time by each user device in each time interval (203, 303);

determining the time interval corresponding to a maximum value of the timely-response rate as the optimal time period (204);

correcting the determined timely-response rate of each user device in each time interval by using the predetermined correction scheme, to minimize a fluctuation amplitude of the timely-response rate (304); and

determining the time interval corresponding to a maximum value of the corrected timely-response rate as the optimal time period (305).

3. The method according to claim 2, wherein if there are multiple maximum values of the timely-response rate of the user device, the method further comprises:

separately obtaining time intervals corresponding to the multiple maximum values of the timely-response rate, and determining an earliest time interval as the optimal time period for

the user device to respond to a message; or

determining the time interval with an optimal network status as the optimal time period for the user device to respond to a message.

4. The method according to any one of claims 1 to 3, further comprising:

dividing the user device group into a first user device group and a second user device group based on a predetermined ratio (401);

separately sending, to a corresponding user device in the first user device group, the message to be sent based on an optimal time period for each user device in the first user device group to respond to a message (403); and

randomly sending, to each user device in the second user device group, the message to be sent (403).

5. The method according to claim 4, further comprising:

collecting statistics on a timely-response rate of each user device in the first user device group in the optimal time period and on a timely-response rate of each user device in the second user device group (404); and

re-determining the optimal time period for each user device in the user device group when it is found by comparison that a difference between the timely-response rate of each user device in the first user device group in the optimal time period and the timely-response rate of each user device in the second user device group is less than or equal to a predetermined difference threshold (405).

6. The method according to any one of claims 1 to 5, further comprising:

in response to determining that the user device group has a user device whose timely-response rate in the optimal time period is lower than a predetermined timely-response rate threshold, re-determining the optimal time period for the user device to respond to a message.

7. The method according to any one of claims 1 to 6, further comprising:

detecting that the label of the message to be sent is a message label for real-time transmission; and

sending, to each user device in the user device group in real time, the message to be sent.

8. The method according to any one of claims 1 to 7, wherein the separately sending, to a

corresponding user device in the user device group, the message to be sent comprises the following:

determining an identifier of a sending server corresponding to the identifier of each user device so that the sending server sends, to the corresponding user device, the message to be sent based on an identifier of each user device in the user device group.

9. The method according to any one of claims 1 to 8, wherein the correction scheme is based on a Wilson interval.

10. The method according to any one of claims 1 to 9, further comprising differentiating between working days and holidays.

11. The method according to any one of claims 1 to 10, wherein performing the mapping operation comprises processing a portion of the plurality of messages to be sent that are within respective responsibility ranges, to reduce overlapped tasks while improving pushing concurrency.

12. A device for pushing an application message, located on an application back-end server side, comprising a plurality of modules configured to perform the method of any one of claims 1 to 11.