Learning Analytics System

JI ZHUORAN, WANG HAICHENG, WANG ZIXU

I. Introduction

In this era, so called information era, the size of the knowledge and information are bursting everyday. People who needs to survive must adapt to the era by learning quickly and efficiently. Online learning platform Therefore, learning how we learn become a heated topic and research area has also been focusing on the topic for several years. In 2013, Stanford University kicked out the research in this field and till now the researches have accumulated many great methodologies and concepts. From only analyze the different background of the learning at the beginning[1], now they are expanding the analysis to surrounding platform like social platform and deepening analysis using more tools like RNNs[2]. On the other side, the industries also speed up to provide better tools for understanding learning process in personal level.

II. Problem Definition

For instructors who teach undergraduate or graduate courses and need to trace the performance of the student in order to make sure the whole class can follow up the contents, the learning analytics system is an integrated system that can collect the data from the student which can reflect the learning activities, analyze the data automatically by data mining and machine learning to find the hidden pattern which can relate the learning activity with the outcome of the student.

Unlike the traditional method, which is time consuming and restricted to the experience of the instructs since that only the raw data are given and this data need be managed manually, our system will provide a user-friendly interface, and most of the data will be visualized by chart. The core subsystem, analysis system, not only can provide the visualization of the performance of the student but also can predict the final result of a particular student with the help suggestion according to the history data or shared database. Instructors can offer their student better help with the assist of this system.

III. Sub-System

A. UI

The layer which interacted with the user directly. The main function of the UI is to display the contents of the system and make it easier for users to interact with the system.

B. Presentation System

The presentation sub-system will receive the request from the UI and ask the data management system for the analysis result, finally return the response to the UI after formatting.

C. Analysis System

The function of the analysis sub-system is analyzing the data and sending back the data to the data management system.

D. Data management system

The sub-system which manage the raw data. It is the abstraction of the database system, provide the interface for raw insertion, data requesting and analysis result storage.

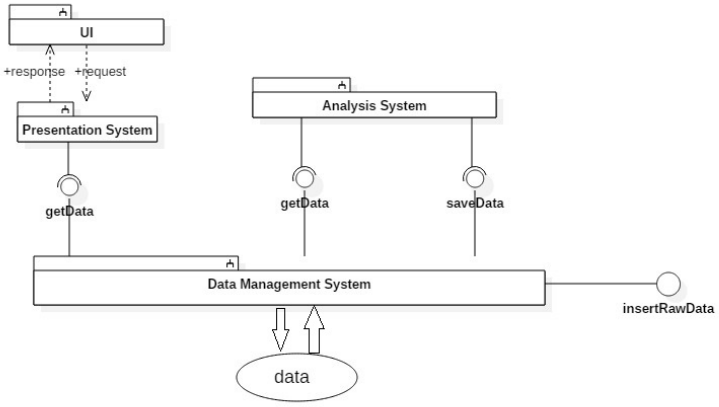

Fig. 1. Sub-system relation

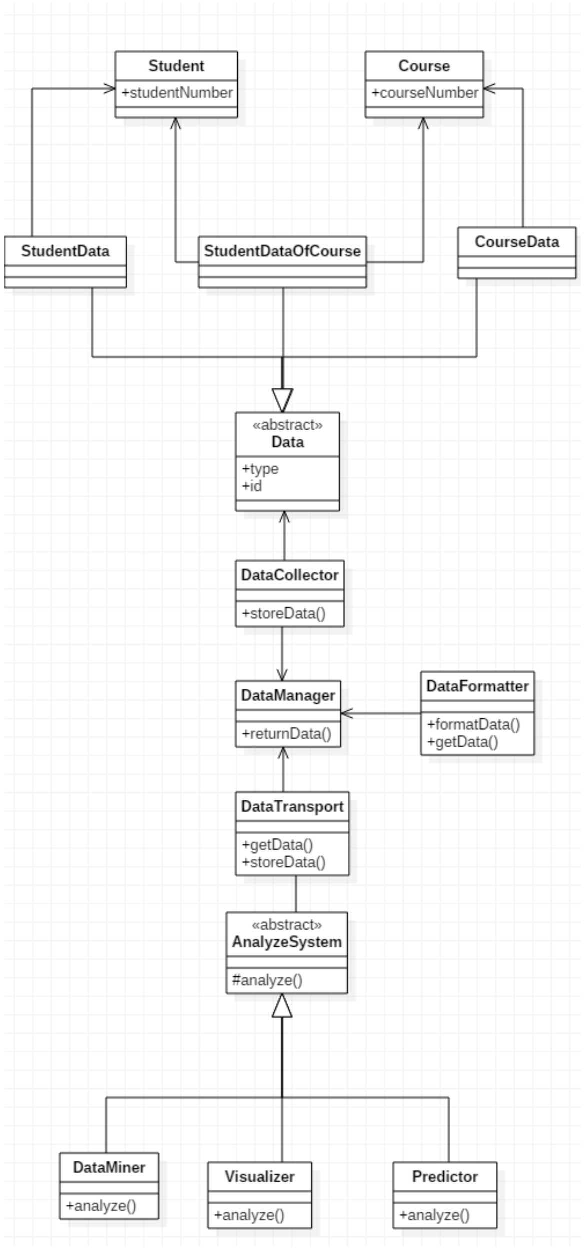

Fig. 2. Sub-system relation

IV. Design Class Diagram

V. Server Model

A. Django

—————- ————— zixu —————- ————
—

B. HTTP

—————- ————— zixu —————- ————
—

VI. Database Model

The core choice of the database model is related to analyzing on demand or per-computation. When a client sends requests to the server, there are two ways for the server to get the result to response. The first one is analyzing on demand, which means that the analyzing will be performed only after receiving the request, while the other is pre-computation, which means that all of the analysis has been finished before, and after receiving the request, the server just search the stored result from the database directly without any computation. The most significant advantage of analyzing on demand is that the users can get the most up to date analysis, while for pre-computing, the response time will be shorter.

For a learning analytics system, real-time data is not that necessary, in other words, even if the analysis of the previous data is acceptable, as long as the data is updated enough. The reason why we need to pre-compute the result is that, in our analytics methods, there are lots of massive computing tasks like data mining or machine learning, which may need long time to find out the result. However, most of users seem to favor immediately response. Another benefit of pre-computation is that the severs can give better concurrence, without massive computing tasks, more resources can be used to response the requests, conversely, to serve same number of clients, smaller scale server cluster can be adopted than the computing-on-demand model, which means that the expenditure of the hardware will be significantly cut down.

Consider the above discussion, in our system, most of the analysis of the row data will be performed at spare time. To achieve the separation of the computation and response, two databases should be build. The first will store the result and accept the new coming data from the students, we call it DB_A, while the second should store the data of previous day for analyzing, we call it database DB_B.

To update the DB_A, 24 hours accessibility will be sacrificed, there will be 1 hours maintenance time at night, in which both database will be offline. Within the maintenance time, these two databases cannot be accessed from the outside. The reason why the maintenance time is needed is that the analysis stored in DB_A should be updated, and the contents of DB_B will be replace by the contents of DB_A, both of these two operations should not be interpret to avoid violate consistent. The scarification of 24 hours accessibility is not a great loss compared to the improvement of response time.

VII. Analysis Method

A. Data Collection

1) Personal Level
    - a) Learning time log
    - b) Learning process log
    - c) Forum usage log (how many times asking or discussing questions on the forum)
    - d) Learning platform social time log
    - e) Connection log and discussing log
    - f) Grade
        - i) Assignment
        - ii) Quiz
        - iii) Class and discuss participation
        - iv) Presentation
        - v) Project score
    - g) Attendance rate
        - i) Class
        - ii) Library
        - iii) Lab
2) Course Level
    - a) Total on-line learning time log (aggregate all personal data)
    - b) Aggregate students process data (statistics)
    - c) Forum activity (aggregate using log), post data (following discuss times, close post statistic ↦ how many people treat it as useful)
    - d) Social platform activity (aggregate using log), connection map
    - e) Average Grade of Whole Class
        - i) Assignment
        - ii) Quiz
        - iii) Class and discuss participation
        - iv) Presentation
        - v) Project score

B. Visualization

1) Basic concept: The main purpose of visualization is to make the data more explicit using chart or visualized objects by visualization technology and computer graphics technology. In our design, the core feature is data normalization. To avoid the bias of human intuition, principle and novel normalization methods will be introduced to reflect the accurate performance of student.

2) Input: Behavioral data: In the visualization stage, the most contributive data is the direct data (i.e. the grade), while other data are feed into the prediction stage. However, other data, such as attendance rate, on-line activities will also be included in the chart, and the most novel part of the visualization is that the output of the prediction will be feed into the visualization system as the input.

3) Output: For each student, one chart for the grade tendency will be generated, and most of the prediction result, such as the possibility of fail, the cause of this situation, and the suggested help method will also be included in the report, from which instructors can easily know the which student need help, why this student need help and what can be done to help this student to learn better.

The visualization of the overall performance of the whole class will also be generated, the reason is that if all students get bad grade, there is a possibility that it is the course itself should be improved, especially when these students get good grades in other courses. In this case, the instructors should consider whether the difficulty of this course does not suit students in this level, or whether the teach method is not acceptable by these students.

4) Normalization method:

1) Interval-scaled variables: continues data that on a roughly linear scale, such as the grade. For example: The score after normalization is

$$newScore_i = \frac{oldScore_i - m_f}{s_f} \tag{1}$$

where

$$s_f = \frac{1}{n}(|oldScore_1 - m_f| + |oldScore_1 - m_f| + |oldScore_n - m_f|) \tag{2}$$

and

$$m_f = \frac{1}{n}(oldScore_1 + oldScore_2 + ... + x_n) \tag{3}$$

After the normalization, every data will be around the zero line, and the newScore is the relative score compared with the average performance of the whole class. The instructors will not be influenced by the bias due to intuition.

2) Binary variables: the variable which is binary, such as the attendance records, 0 means absence and 1 means present. We assume that most student are diligent and will show up most time, the data will be asymmetric, hence Jaccard coefficient will be used. The relative value is $d(i, avg) =$

$$1 - \frac{number \quad of \quad 1's \quad in \quad i \wedge avg}{number \quad of \quad 1's \quad in \quad i \vee avg} \tag{4}$$

where the average can be calculated by threshold method, for example, if 80 percent students are present, then the average case is present (i.e. 1).

3) Mixed type variables

C. Prediction

1) Basic concept: The main purpose of this prediction part is to make reasonable prognosis automatically. As we can image a season teacher can make an accurate guess of a particular student based on his or her current study, the aim is to simulate those seasoned teachers to help predict for larger scale of student automatically, which can save the time for human teacher to absorb data and analyze manually and also enhance the prediction accuracy.

2) Input: Behavioral data: all the data except the direct result (all type of grade) For example, the tutorial attendance in traditional data and learning progress log in on-line data. Intuitively human teachers make reasonable guess based on those data so we assign much importance to those behavioral data closely related to the learning process. Also we use some data like social platform surrounded the learning platform to try to figure out more reliable indicators to help us enhance accuracy.

3) Output: expected result $\mapsto$ final grade, we map all results to five classes, A, B, C, D, F, with grade decreasing one after another. In another word, this part actually classify different learning styles to different grade class. It could be more tolerated compared to the model directly predict the result in percentage system and more diverse compared to binary classification.

4) Analysis methodmachine learning: We use both traditional statistical learning and deep learning in our analysis. Because we are not familiar to the real learning analytics data, we are trying to pick several popular analysis models which can handle linearly separable data and other data with more complicated nature.

We adopt a powerful python machine learning library to facility us applying statistical learning methods. And we adopt the library function to auto select the best model with highest accuracy to process the data. In essence, we separate the analysis part as a black box, so that we can focus on the server building and event handling.

As for deep learning part, we adopt another python library to test our assumptions to use ANN to make some difficult prediction such as predict final grade at the beginning of the course.

5) Other functionality based on the prediction model:

1) Can filter out top student study habit model
2) Define risk factor $\mapsto$ can find the student with high risk and make alert
3) Comparison different student study model to give corresponding predefined result
4) Can figure out some suspect students with unmatched study style and current result

VIII. Conclusion

Appendix A
Proof of the First Zonklar Equation

Some text for the appendix.

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

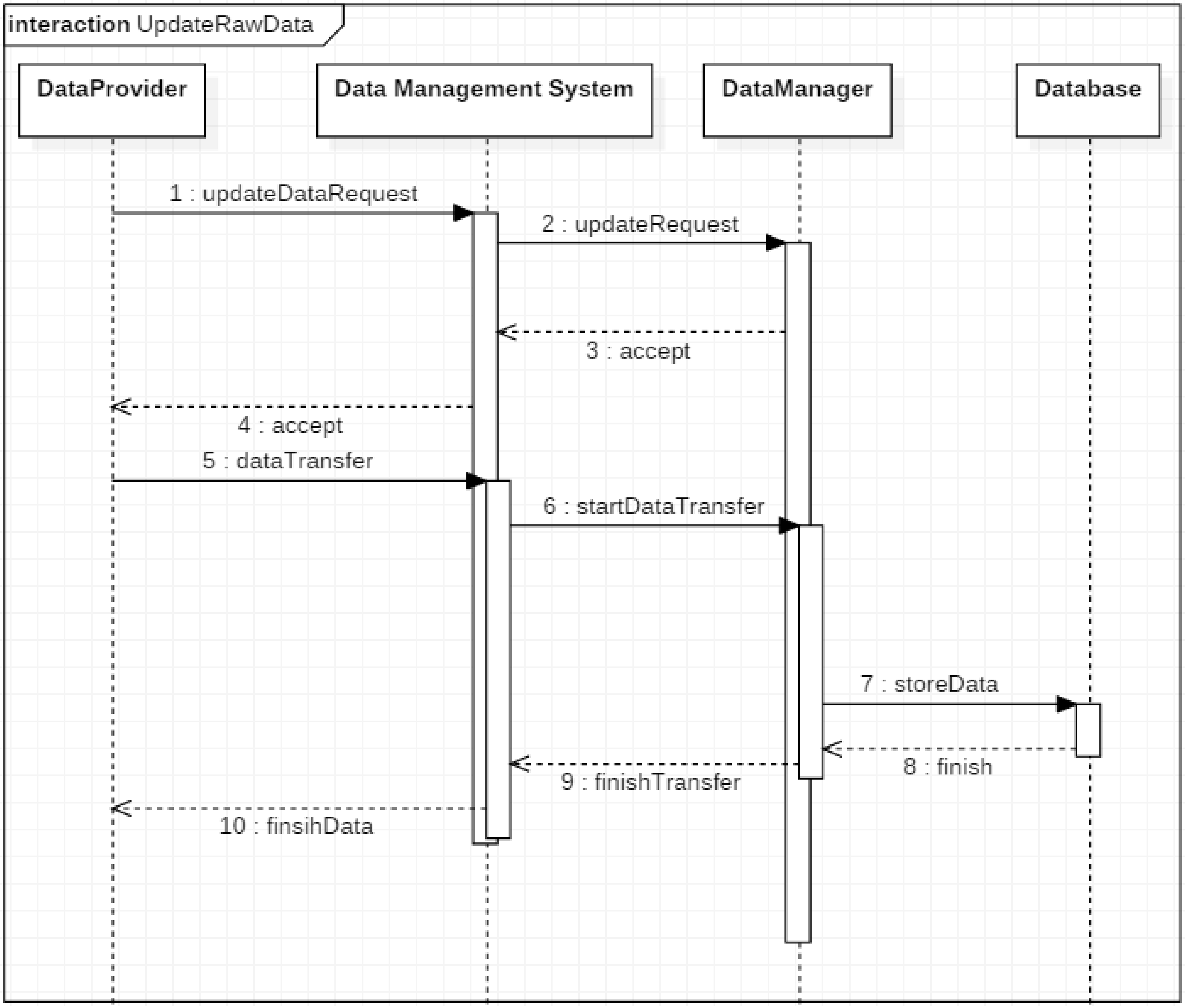

Fig. 3. Sub-system relation

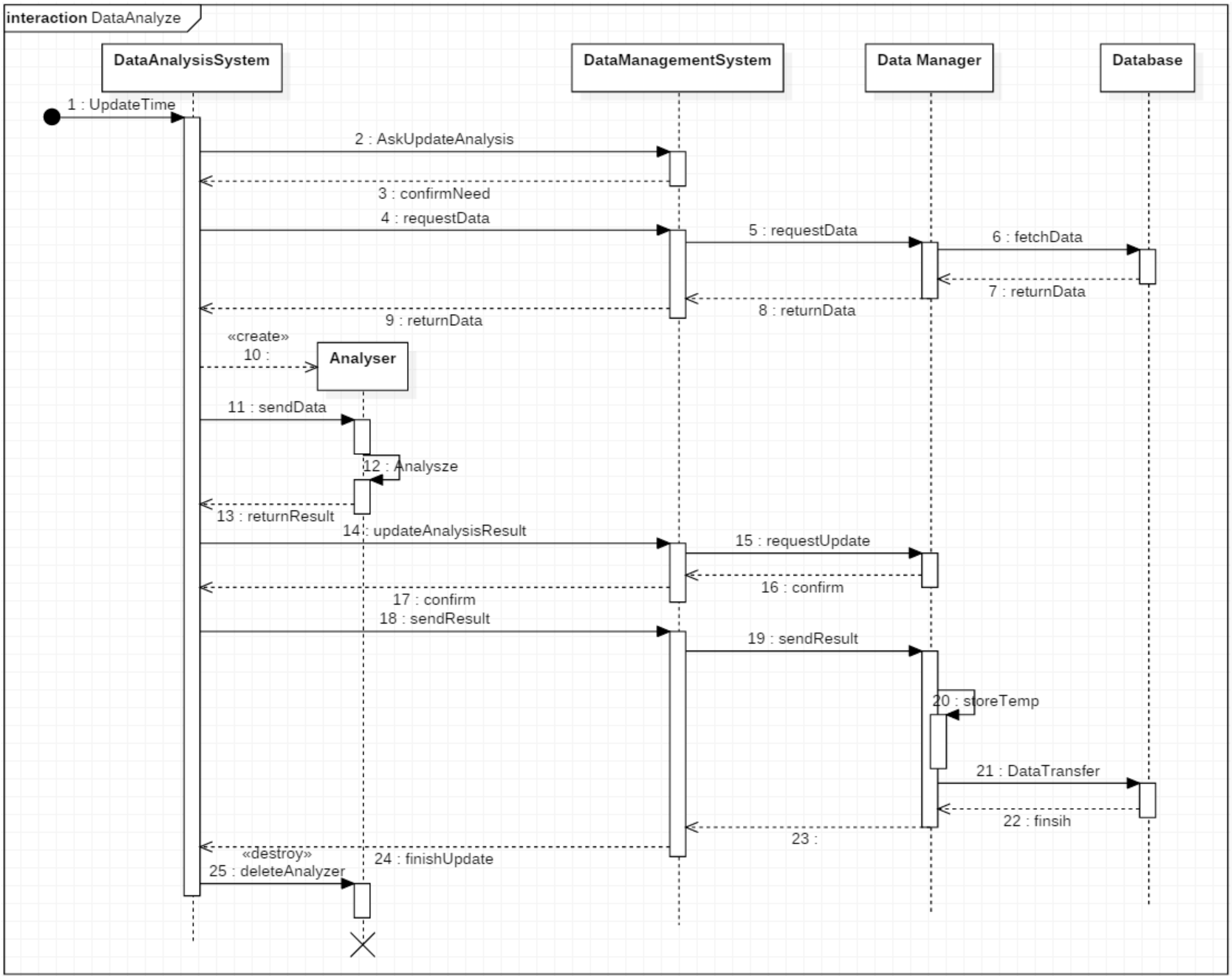

Fig. 4. Sub-system relation

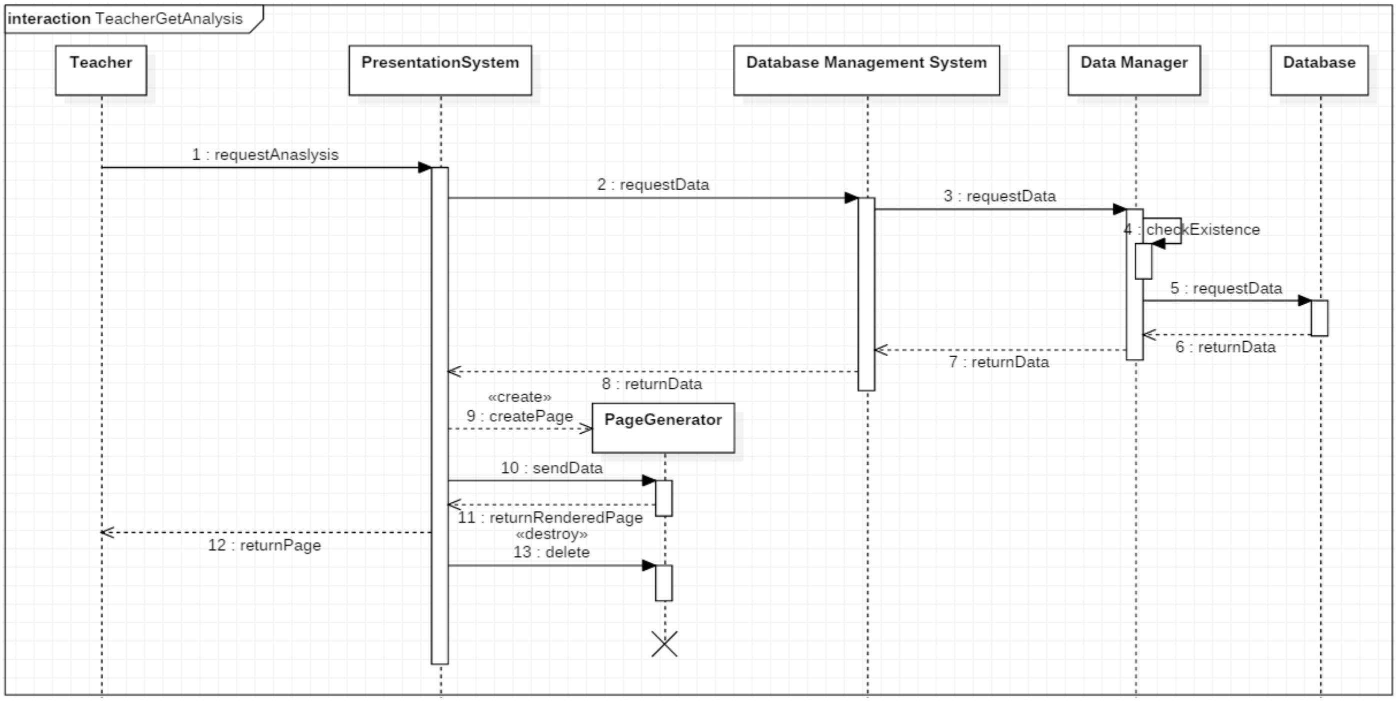

Fig. 5. Sub-system relation