AUTOMATIC SQL QUESTION GENERATION USING REVERSE APPROACH

Abhilasha A. Dwivedi[1], Dr. Dinesh D. Patil[2]

[1]PG Scholar, Department of Computer Science and Engineering, SSGBCOET, Bhusawal, India
[2]Associate Professor and Head, Department of Computer Science and Engineering, SSGBCOET, Bhusawal, India

---***---

Abstract - *Structured query language (SQL) is a language intended for managing data and basic SQL statements are a key basis when studying or learning practical database courses either in college or from any other educational institute. With the aim of making students learn SQL appropriately, the instructor spends a lot of time and effort to prepare SQL exercises for teaching SQL queries to them. However, with time limitation, the instructor had to reuse the old exercises having less diversity in SQL commands. Due to this, the students might not have sufficient questions to meet their needs. This paper discusses the reverse approach of creating SQL exercises by creating answers first instead of questions which not only saves the time of instructors but also generates variety of SQL questions, and also introduces the framework which is developed by using reverse approach. This paper also provides discussion on advantages and future scope of using reverse approach over traditional approach for generating SQL exercises.*

***Keywords*- reverse question generation, SQL, automatic question generation, database, SQL learning**

1. INTRODUCTION

Most of the applications in Information Technology involves storing and retrieving information from databases. The process of data storage, manipulation and retrieval needs the knowledge of database languages like SQL. Structured query language (SQL) is language designed for managing data held in relational database management system (RDBMS) and also a fundamental knowledge to all of computer science students. It is a standard language for forming queries to access relational database systems, e.g. create table, read, update, and delete data from the database. Thus, it is commonly taught in computer science classes and most engineering colleges includes SQL in their course.

In order to teach and making the students to learn and practice SQL commands appropriately, the SQL instructor or teacher prepares the SQL exercises by creating the SQL questions and their corresponding SQL queries. There are mainly two methods to prepare SQL exercises for teaching SQL. The first method is to prepare the exercises by taking database schema and here the benefit is for instructors because they do not have to validate the answer which is returned. On the other hand, students have to imagine their results by themselves. The second method is to create exercises with reference to both schema and data set which makes students learn and try the query with real database which results in more effective learning. However this method creates problems for instructors since it consumes more time and effort where input is the database [14]. These exercises which are created from schema and dataset involves SQL question, SQL query and answer validation for which instructor spends a lot of time and effort. But due to limitation of time, the instructor reuses the old exercises which might have lesser quantity of exercises having little varieties of questions. Also, in some systems queries are generated automatically [9]. Also validation of the queries might be a challenging task since it includes occupying a test database which may contain several tables and then examining the results of the query execution on the test database [10]. Fig. 1 represents the process of generating the SQL exercises manually by the instructor or teacher which consumes lot of time. Existing systems have also some restrictions on the amount of supported SQL statements [11].

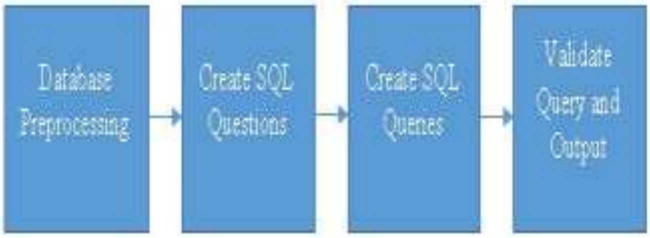


Fig-1: Generating SQL exercises manually

This paper presents the framework which uses reverse approach of generating SQL exercises by creating SQL queries first and from that queries corresponding SQL questions are generated. It takes the database as input having reference to both database schema and data and then generates the SQL queries from metadata and SQL questions from the SQL queries and then executes the queries at runtime thereby displaying the output of the executed query. This framework can be used to solve the above mentioned problems of teachers and instructors in generating SQL exercises.

2. LITERATURE REVIEW

Automatic question generation

Developing Automatic Question Generation systems became one of the important research issues because it requires understandings from a range of disciplines, such as Artificial Intelligence, Natural Language Understanding, and Natural Language Generation. In order to develop an automatic question generator, a number of researchers have presented their effort and many algorithms or methods are proposed to develop the automatic question from specified or given text.

To automatically generate question by creating answer first is a better approach to lighten the instructor's workload and enhance learning practices for students. By reviewing the reverse approach for question generation, basically four approaches are found to be used.

1. “questions by mutation” mutates the specified text and generates questions from it.

Lee *et al.* in “Generating Grammar Questions Using Corpus Data in L2 Learning” has used collection of errors which were found by analyzing existing corpora and then generated questions having common errors which the students have to identify [2] and explored automatic generation of English grammar questions based on statistical machine learning techniques.

In the same way, Funabaki *et al.* in “Toward Personalized Learning in JPLAS: Generating and Scoring Functions for Debugging Questions” has created debugging questions for programming courses [3], [4].

2.“questions by keyword” use keywords from the input and then creates a new question from that extracted keyword.

Liu *et al.* in “Using Wikipedia and conceptual graph structures to generate questions for academic writing support” presents an intelligent question generator tool that supports writing literature reviews. It helped students improve their literature reviews by parsing sentences from their reviews, extracting key phrases and finding appropriate sentences from Wikipedia pages, then forming questions about those sentences [5].

3. “questions from input”, generates the questions from the input such as images.

Jain *et al.* in “Creativity: Generating diverse questions using variational auto encoders” generated questions from the input image [6]. Their research algorithm can be applied in a number of fields, e.g. educational study, driving assistance, entertainment, etc.

4.“questions from metadata”, generates the questions from the metadata.

Abdul Khalek and Khurshid’s “Automated SQL query generation for systematic testing of database engines” generates new SQL queries from the database schema for testing database performance [7] which is capable of automatically generating syntactically and semantically appropriate SQL queries for testing, and input data to populate test databases, and also expected result of executing the given query on that generated data. However, it did not consider data in tables or generate a text explanation of the query which was to be used as an SQL question in exercises.

All of the above research papers fundamentally converts a source constituent into a new constituent i.e. question. This source constituent can be either the solution of the question, such as a grammatically correct sentence or bug free source code, or something more complex, such as an input image or literature review of a student. Some research papers use SQL queries as source constituent but without considering data or some may consider data but only with some selected SQL commands [13].

Reverse SQL Question Generation (RSQLG) Algorithm

Thanakrit Julavanich, Srinual Nalintippayawong and Kanokwan Atchariyachanvanich, in “RSQLG: The Reverse SQL Question Generation Algorithm” proposed an algorithm that uses the reverse method of generating SQL exercises by generating the SQL queries first instead of SQL questions [1]. RSQLG algorithm uses different method by starting with the answer, and then generating the question.

The RSQLG algorithm reverses the process by creating an SQL query first and then that query from first step is taken as input to generate SQL question and then it is validated simultaneously. The RSQLG algorithm supports only DML commands such as SELECT, UPDATE OR DELETE [12]. It does not support many unsupported commands like DDL commands and also needs some more work to be done on complex grammars and queries.

The Reverse SQL Question Generation (RSQLG) Algorithm requires some improvement in complex exercises, language support which has more complex grammar and more unsupported commands like Data Definition Language (DDL). Thus for increasing learning efficiency, the proposed framework introduced in this paper considers data in tables and provides SQL commands which are unsupported in existing systems.

3. PROPOSED METHODOLOGY

This section presents the proposed framework which reverses the traditional approach of manually writing the SQL questions and then creating its corresponding SQL queries. The proposed framework uses the reverse approach which starts creating the SQL query (answer) first and then create the SQL query explanation (question). The process of generating SQL exercises using reverse approach is represented in Fig. 2. This proposed framework uses a database as input to generate the SQL query and then from the query it generates the text description as a question. It basically reverses the traditional process by creating an SQL query (answer) first. The generated query is an input to generate query explanation (question) and is validated simultaneously. Hence, the proposed work uses two approaches i.e. ''questions from keywords'' for SQL question generating process and ''questions from metadata'' for SQL answer generating process. [8]. Fig. 3 represents the block diagram of the proposed framework.

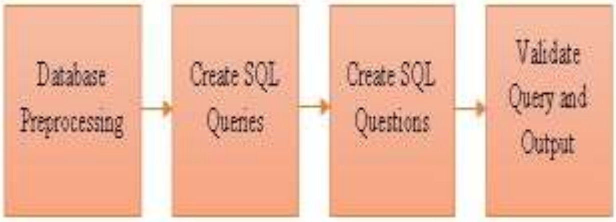


Fig-2: Generating SQL exercises using Reverse approach

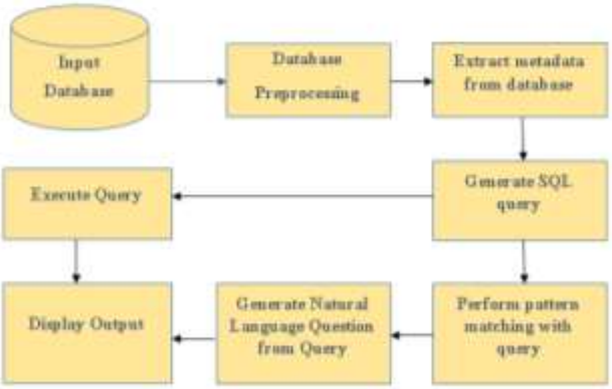
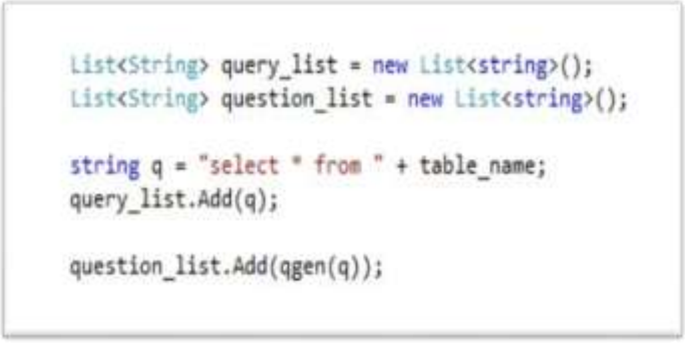

Input Database
Database Preprocessing
Extract metadata from database
Generate SQL query
Execute Query
Display Output
Generate Natural Language Question from Query
Perform pattern matching with query

Fig-3: The Proposed Framework

The reverse approach used in proposed framework firstly performs database pre-processing which extracts structure of database needed for generating SQL questions. Then, other important step is query metadata generation which is necessary for SQL query generation. After generating queries, next important step is to convert the generated SQL query into SQL question.

Database Preprocessing: Database pre-processing extracts the data and metadata necessary for SQL query creation, e.g. database schema, table schema, attribute data type, etc. This reduces the time to generate queries.

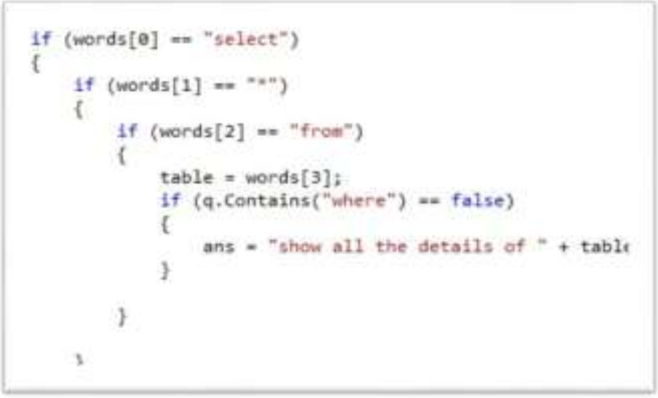
SQL Query generation: After query metadata has been generated, the metadata extracted from the database will be processed to generate the SQL query. An example of query generated by using reverse approach in our proposed framework is shown in Fig. 4.



```
List<String> query_list = new List<string>();
List<String> question_list = new List<string>();

string q = "select * from " + table_name;
query_list.Add(q);

question_list.Add(qgen(q));
```

Fig-4: Example of SQL query generation

SQL Question generation: The SQL question generation expands the generated query with its corresponding text explanation which is a natural language description of the query by separating out each part of command and replacing it with natural language. Fig. 5 represents an example of generating SQL questions from the SQL queries.



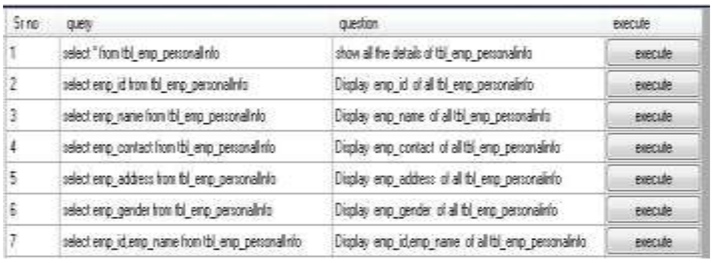
```
if (words[0] == "select")
{
    if (words[1] == "*")
    {
        if (words[2] == "from")
        {
            table = words[3];
            if (q.Contains("where") == false)
            {
                ans = "show all the details of " + tabl
            }

        }

    }
```

Fig-5: Example of SQL question generation

Query execution: The proposed application works with data stored in the database and will generate the output of the SQL command when the query is executed.

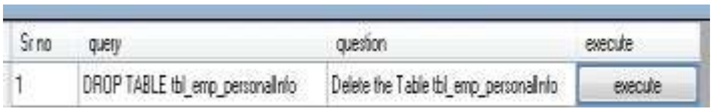
4. RESULTS

We have used one database containing two tables in the proposed framework for generating questions by taking database as input and output is SQL questions and SQL answers. The number or quantity of questions generated depends upon the database schema, table schema and data in the tables. More the number of tables in the database and number of attributes and data in the table, more will be the questions generated. Fig. 6, Fig. 7 and Fig. 8 represents the example of output generated, example of DDL command and query execution output respectively by the proposed framework.



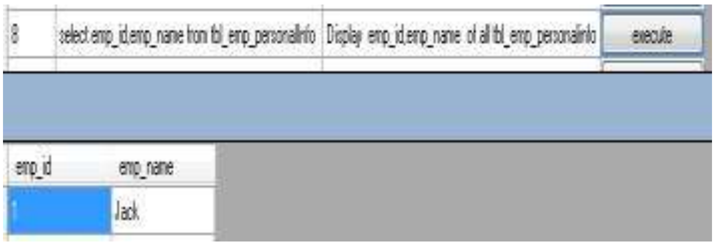
Sr no	query	question	execute
1	select * from tbl_emp_personalinfo	show all the details of tbl_emp_personalinfo	execute
2	select emp_id from tbl_emp_personalinfo	Display emp_id of all tbl_emp_personalinfo	execute
3	select emp_name from tbl_emp_personalinfo	Display emp_name of all tbl_emp_personalinfo	execute
4	select emp_contact from tbl_emp_personalinfo	Display emp_contact of all tbl_emp_personalinfo	execute
5	select emp_address from tbl_emp_personalinfo	Display emp_address of all tbl_emp_personalinfo	execute
6	select emp_gender from tbl_emp_personalinfo	Display emp_gender of all tbl_emp_personalinfo	execute
7	select emp_id,emp_name from tbl_emp_personalinfo	Display emp_id,emp_name of all tbl_emp_personalinfo	execute

Fig-6: Example of generated output



Sr no	query	question	execute
1	DROP TABLE tbl_emp_personalinfo	Delete the Table tbl_emp_personalinfo	execute

Fig-7: Example of DDL command and its equivalent generated Question



8	select emp_id,emp_name from tbl_emp_personalinfo	Display emp_id,emp_name of all tbl_emp_personalinfo	execute

emp_id	emp_name
1	Jack

Fig-8: Example of Query Execution Output

5. CONCLUSIONS

Creating variety of questions for teaching SQL appropriately to students or learners requires large amount of time and efforts which can increase the teachers or instructor's workload .Using reverse approach for creating SQL exercises is a better approach than the traditional method. This reverse approach used in this paper can reverse the manual question creation process which starts creating query first instead of creating questions. The reverse approach has ability of generating bulk of SQL questions requiring less time and effort. The proposed framework provides query execution results since it works with both database schema and datasets. The teachers or instructors can set the language, format and question explanation by themselves.

The proposed framework can be combined with some other methods or can be modified so that it can be used in colleges to teach students or in computer institutes for practicing SQL or can be used in e-learning systems. It has several advantages such as it reduces time and effort of instructor in creating exercises, decreases workload of instructor of creating questions manually, which saves their time. Also students will get variety of questions for practicing that may improve SQL skills of students and learning results of students will be more efficient and so on.

ACKNOWLEDGEMENT

It is my pleasure in submitting this paper on "Automatic SQL Question Generation using Reverse Approach". I would like to express true sense of gratitude towards my Principal Dr. R. P. Singh and special thanks to my guide and H.O.D., Dr. Dinesh D. Patil who at every discrete step in preparation of this paper contributed his valuable guidance, support, encouragement and help to solve every problem that arose. Also I thank all the researchers and authors of reference papers referred by me for this paper. Also, most likely I would like to express my sincere gratitude towards my parents for always being there and for their continuous support and encouragement.

REFERENCES

[1] Thanakrit Julavanich, Srinual Nalintippayawong and Kanokwan Atchariyachanvanich, "RSQLG: The Reverse SQL Question Generation Algorithm," in IEEE 6th International Conference on Industrial Engineering and Applications, 2019.

[2] K. Lee, S.-O. Kweon, H. Seo, and G. G. Lee, "Generating grammar questions using corpus data in L2 learning," in Proc. IEEE Spoken Lang.Technol. Workshop (SLT), Dec. 2012, pp. 443_448.

[3] N. Funabiki, T. Mohri, and S. Yamaguchi, "Toward personalized learning in JPLAS: Generating and scoring functions for debugging questions," in Proc. IEEE 5th Global Conf. Consum. Electron., Oct. 2016 pp. 1_4.

[4] S. Yamaguchi, T. Mohri, and N. Funabiki, "A function for generating debugging questions in a Java programming learning assistant system," in Proc. IEEE 4th Global Conf. Consum. Electron. (GCCE), Oct. 2015, pp. 350_353.

[5] M. Liu, R. A. Calvo, A. Aditomo, and L. A. Pizzato, "Using Wikipedia and conceptual graph structures to generate questions for academic writing support," IEEE Trans. Learn. Technol., vol. 5, no. 3, pp. 251_263, Jul. /Sep. 2012.

[6] U. Jain, Z. Zhang, and A. Schwing, "Creativity: Generating diverse questions using variational autoencoders," in Proc. IEEE Conf. Comput. Vis.Pattern Recognit. (CVPR), Jul. 2017, pp. 6485_6494.

[7] S. A. Khalek and S. Khurshid, "Automated SQL query generation for systematic testing of database engines," in Proc. IEEE/ACM Int. Conf. Automated Softw. Eng., Sep. 2010, pp. 329_332.

[8] Bikash Chandra, BhupeshChawda, and BiplabKar, "Data Generation for Testing and Grading SQL Queries", may 2017.

[9] Q. Do, R. Agrawal, D. Rao, and V. Gudivada. "Automatic Generation of SQL Queries". In: Proceedings of the 121st

ASEE Annual Conference & Exposition. ASEE, June 2014, pp. 1 –11.

[10] Claudio de la Riva, María José Suárez-Cabal, Javier Tuya, ”Constraint-based Test Database Generation for SQL Queries”, in Proceedings - International Conference on Software Engineering, May 2010, pp. 266-276.

[11] Fernando Almeida, "Practical SQL Guide for Relational Databases ", January 2016.

[12] B.-G. Itzik, L. Kollár, D. Sarka, and S. Kass, Inside Microsoft SQL Server 2008-T-SQL Querying: Microsoft Pr., 2009.

[13] K. Atchariyachanvanich, S. Nalintippayawong, and T. Permpool, “Development of a MySQL sandbox for processing SQL statements: Case of DML and DDL statements,” in Proc. 14th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE), Jul. 2017, pp. 1–6.

[14] C. Domínguez and A. Jaime, ‘‘Database design learning: A project based approach organized through a course management system,’’ Comput. Educ., vol. 55, no. 3, pp. 1312–1320, Nov. 2010.