

Automatic SQL Query Formation from Natural Language Query

Pranali Nagare¹, Smita Indhe², Dhanashri Sabale³

Guided by- Prof. D. Y. Thorat , Prof. G.K. Chaturvedi

¹²³UG STUDENT, Department of Information Technology (IT) Matoshri Asarabai Polytechnic Eklahre Nashik, M.S.(India)

ABSTRACT - Natural language processing is a field of computer science concerned with the interaction between computer and human (natural) languages. It is attractive areas in the interaction between human and computer. These include spoken language that incorporate speech and natural language. It is an interdisciplinary research area at the border between semantics and artificial intelligence, aiming at developing computer programs capable of human-like activities like understanding and producing text or speech in a natural language, such as English or conversion of natural language in text or speech form to languages like SQL. Information retrieval and information organization, machine translation is the most important application of natural language processing. The goal of Natural Language Processing is for an English sentence to be get interpreted by the computer and appropriate action taken respectively. In this paper we are going to implement the system which is currently capable of handling simple queries along with some complex queries. Because all the forms of SQL queries are not supported, further development would be required. In this system the objective is to parse the queries and with the help of the dictionary, carry out different phases like morphological analysis, syntactic analysis, semantic analysis etc. and finally the SQL query is generated.

Key Words - Natural Language Query, Machine

1. INTRODUCTION

While natural language may be the easiest system for people to learn and use, it has proved to be the hardest for a computer to understand. The goal of NLP is to allow communication between people and computers without resorting to memorization of complex commands and procedures. In other words, NLP is a technique, which can make the computer understand the languages naturally used by humans. In this project, we are translating English query into a SQL query using semantic grammar. The system will accept user's query in natural language as an input. The program will check whether the query is valid or not. Then we will generate tokens by performing the division of the question clause. Each token represents a single word in the user's query. The tokens from the query clause are compared with clauses already stored in the dictionary. The dictionary needs to be constantly updated. Then the algorithm scans

the tokens and tries to find attributes present in the query. Then we find all the tables in the database which contain the attributes by comparing syntax and semantics. Then we build the final SQL query and execute it on the database and return the result dataset to the user.

2. Proposed System Concept

2.1 Existing System

In the existing system we need to fire the SQL Queries for which we must have knowledge about DDL (Data Definition Language), DML (Data Manipulation Language), DCL (Data Control Language) queries of SQL. If the queries are error prone in syntactical manner its complex to resolve the error as it requires in depth knowledge about the query details.

2.2 Proposed System

From natural language query we can fire our query. It is the interdisciplinary research area at the border between linguistics and artificial intelligence, aiming at developing computer programs capable of human like activity like understanding and producing text or speech in natural language such as English or conversion of natural language in text or speech from to language like SQL. This is done to ensure that no authorized person should be able to retrieve data or modify the data in the database. The system design of proposed system is as shown below:

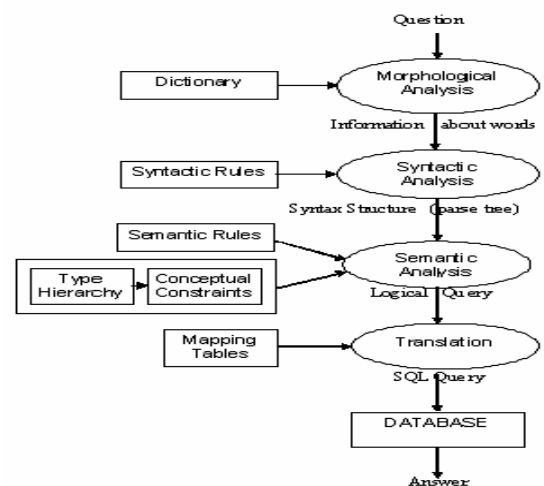


Fig 2.2: System design

3. UML Diagrams

3.1.1 Use case diagram

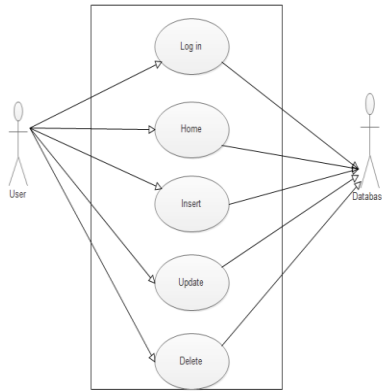


Fig 3.1.1: Use case

3.1.2 Class Diagram

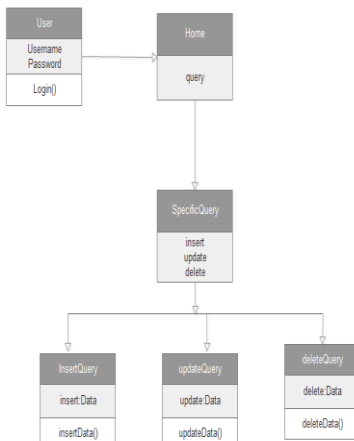


Fig 3.1.2: Class diagram

3.1.3 Sequence Diagram

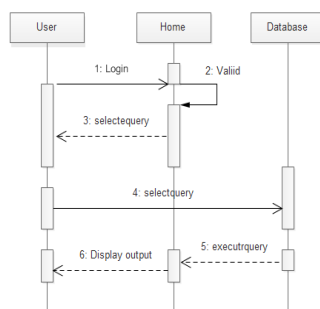


Fig 3.1.3: Sequence diagram

4. Methodology

4.1.1 Algorithm

[1] The first step in query formation is to process the input query, i.e. divide the user's query into tokens, which are individual words in the query

[2] Then replace the starting of the query with appropriate syntax. Identify the words in the user's query which are column names (attributes), the words which represent the name of the tables and values, if any, present in the user's query.

[3] Replace synonyms of column names or table names with the actual attributes or table names.

[4] Construct the query as follows:

- Identify the attributes which the user wants to retrieve. This will be appended to the select keyword.
- Identify the table to which these attributes belong. This will be appended to the from keyword.
- Identify the conditions or values, if any, specified by the user in his/her query.

[5] Generate the final query and fire it on the database to get the required result which will be displayed to the user.

4.1.2 Flowchart

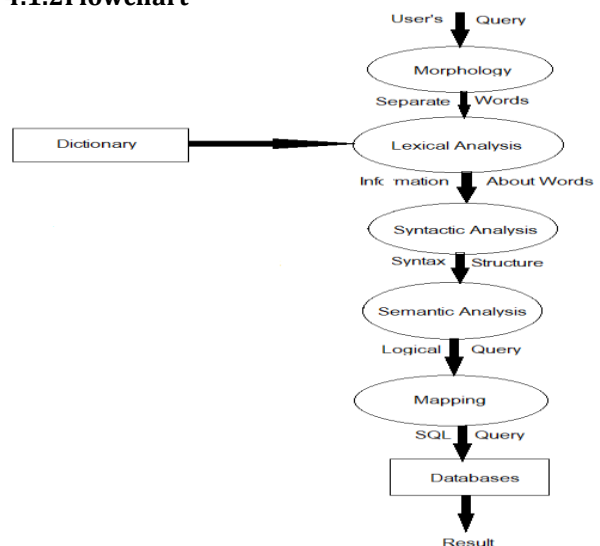


Fig 4.1.2: Flowchart

5. Advantages of Proposed System

- 1 This system will help user to get the results of the database as per his/her understandable English language.
- 2 This system will reduce the dependency of SQL queries for data fetching.

6. Proposed System Scope

The scope of the proposed system is as follows:

- 1 To work with RDBMS one should know the syntax of the commands of SQL.
- 2 The interface language is chosen to be English for accommodating wider users.
- 3 Input from the user is taken in the form of questions (like what, who, where).
- 4 A limited data dictionary is used where all possible worlds related to a particular system will be included. The data dictionary of the system must be regularly updated with words that are specific to the particular system.
- 5 Split the question string in to tokens and give order number to each token identified.
- 6 To remove excessive words from the user input statement. Escape words have been considered which must be regularly updated with words that are specific to the particular system.
- 7 To construct an RDBMS query using tokens, an algorithm has to be developed.
- 8 Ambiguity among the words will be taken care of while processing the natural language.
- 9 All text labels in each figure are legible. Here the natural language processing is done in English i.e. the input statements have to be in English

Information Technology Research (IJCSEITR) Vol. 3 Issue 2, June 2013, 161-166.

[3] "Automatic SQL Query Formation from Natural Language Query", International Journal of Computer Applications.

[4] Huangi, Guiang Zangi, Phillip C-Y Sheu – "A "Natural Language database Interface based on probabilistic context free grammar", IEEE International workshop on Semantic Computing and Systems 2008.

7. Software Requirement

1. Microsoft Windows Operating System: Windows 7
2. Programming Language: JAVA (jdk 1.8+)
3. Database: MySQL Server (MySQL 5.5)

8. Hardware Requirement

- 1 Processor : Pentium IV
- 2 RAM: 1GB
- 3 HDD: 40GB

9. Conclusion

Natural Language Processing is very powerful tool which can change the complete working of the computer program interface. The aim is to evaluate correct SQL queries by using NLP. This system is currently capable of handling simple queries along with some complex queries because not all forms on SQL queries are supported, further development would be required.

References

[1] "SQL Query Formation Using Natural Language Processing (NLP)", Prof. Debarati Ghosal¹, Tejas Waghmare², Vivek Satam³, Chinmay Hajirnis⁴, International Journal of Advanced Research in Computer and Communication Engineering(IJARCC) Vol. 5, Issue 3, March 2016

[2] "Natural Language to SQL Conversion System", International Journal of Computer Science Engineering and