A Flexible and Efficient Natural Language Query interface to databases

B.Sujatha[1,] Dr.S. Viswanadha Raju[2]

[1]*Research Scholar,*
Dept. of CSE
JNTU Hyderabad

[2]*Professor & Head*
Dept. of CSE
JNTUCEJ
Karimnagar

Abstract: **Emergence of NDLIB (Natural language Interfaces to Databases) is the need of the hour. Crux of all Information technology (IT) applications is storing and retrieving information from databases. To retrieve information from database requires knowledge of database languages such as SQL. Naive users do not possess knowledge of SQL which is the forte of database expert who are computer proficient. Access to database systems by naïve users require a logically database independent interface. NDLIB suffices this need. NDLIB overcomes the need to write SQL queries as naïve users may not be aware of the structure of the database. NDLIB allows access to database through natural language queries making them logically independent from data. This led to the development of new type of processing called Natural language Interface to Database. Research indicates existing NDLIB systems lack flexibility of forming queries in user's own format. They also provide an incomplete specification to the requested data. Other areas of challenge are ambiguities in various forms: structural, word sense, referential and literal. We are proposing EFLEX system (A Flexible and Efficient Natural Language Query interface to databases) to tackle above challenges. The proposed system consists of three major components. They are a) an analyzer b) a mapper and c) a translator. The function of analyzer is to interpret the queries entered by the next user. Mapper is used to correspond natural language query to SQL query. Finally the Translator component performs actual translation of a query. The system is designed to be efficient in the way a user's query is translated into SQL query. The efficiency is achieved by using KMP Algorithm to translate a Natural language query into SQL query. The evaluation was performed on the first version of the software. The empirical validation of the prototype shows improved performance.**

Keywords: **NDLIB, EFLEX, SQL, KMP, Analyzer, Mapper, Translator**

I. INTRODUCTION

Brief background of Emergence of NLIDB systems:

NLIDB allows access to databases through natural language requests. Database management system (DBMS) is a specially designed software applications that interact with the user, other applications, and the database itself to capture and analyze data. A general-purpose DBMS is a software system designed to allow the definition, creation, querying, update, and administration of databases. Well-known DBMS include Oracle, MySQL, Microsoft SQL Server, IBM DB2, MariaDB, PostgreSQL, SQLite, SAP, dBASE, FoxPro, LibreOffice Base and FileMaker Pro.

A database is not generally portable across different DBMS, but different DBMSs can interoperate by using standards such as Structured Query Language (SQL) and Open Database Connectivity (ODBC) or Java Database Connectivity (JDBC) to allow a single application to work with multiple databases.

However the fundamental drawback is to managing it. It needs to grasp the information from connected Database by using a language known as Structured Query Language. SQL is a standard language for accessing and manipulating databases. It is used to execute, retrieve, insert, update, delete and creating tables, stored procedures, views in a database. This language essentially obtains the information from those records in the database which keeps the information with it.

But it needs to know about how to frame the query with the user requirement. As per the user mistreatment or framing the query for the information will depends upon their demands. It's impossible for everyone to be learn these languages. This challenge can be tackled through state of the art NLIDBs and EFLEX system which we are proposing. This type of serving to systems area unit developed by more individuals, with many totally different dimensions. Our focus is on these issues to develop an user interactive system to overcome these problem to serve as NLIDB systems (Natural language Interface to Database). It's some way to create up smart interface to the databases that edges the one who uses it.

NLIDB system permits the user right to use information replenish within the information once he/she enters the need, with the assistance of some natural language. The crux of this work is it states that NLIDB system expresses user entries. The aim of our work is to associate in nursing introduction of analysis within the space of linguistic communication interfaces to databases. It is by no means that an entire discussion of all the problems that are unique and relevant to linguistic communication interfaces databases.

The region of NLIDB system investigation is at trial level. And additionally this method is restricted up to a little level, however not at an outsized level. Managing higher level the current systems don't cope-up. There's a region wherever NLIDB system area unit commanding the question Methodology of information.

More investigation is currently additionally developing on in NLIDB system interfaces. The Regions for example error news, once a sentence could not be evaluated fruitfully, latest move towards to NLIDB systems, and automatic generation of data communication systems square measure still necessary unresolved issues.

However, the standard approach to data linguistic communication systems is well established. This approach creates a semantic grammar for each data and uses this to analyze English question. The linguistics creates an illustration of the linguistics, or meaning, of the sentence.

Natural Language User Interfaces (LUI or NLUI) are a type of computer human interface where linguistic phenomena such as verbs, phrases and clauses act as UI controls for creating, selecting and modifying data in software applications.

In interface design natural language interfaces are sought after for their speed and ease of use, but most suffer the challenges to understanding wide varieties of ambiguous input.[1] Natural language interfaces are an active area of study in the field of natural language processing and computational linguistics. An intuitive general Natural language interface is one of the active goals of the Semantic Web.

Text interfaces are 'natural' to varying degrees. Many formal (un-natural) programming languages incorporate idioms of natural human language. Likewise, a traditional keyword search engine could be described as a 'shallow' Natural language user interface.

2 History Of Nlidb

The really initial makes and effort at natural language processing data interface are merely a sold because the different natural language processing analysis. Asking inquiries to databases in language is extraordinarily convenient and simple technique of information access, notably for casual users World Health Organization do not understand sophisticated data query language like SQL. Here are some samples of the language Interface to data systems

1. Lunar System[1, 2]:

As per [1, 2]: “The system satellite may well be a system that answers queries on samples of rocks brought back from the moon. The system was informally introduced in 1971. To accomplish it’s perform the satellite system uses a pair of databases; one for the analysis and so the various for literature references. The satellite system uses Associate in nursing increased Transition Network (ATN) Trojan horse and Woods' procedural linguistics. The satellite system performance was quite impressive; it managed to handle seventy eight of requests with none errors and this magnitude relation rose to ninetieth once reference work errors were corrected. But these figures are additionally dishonourable as a result of the system wasn't subject to intensive use because of the limitation of its linguistic capabilities.”

2. Ladder:

As per [1, 3]: “The LADDER system was designed as a communication interface to a data of data regarding US Navy ships. The LADDER system uses linguistics synchronic linguistics to interrupt down inquiries to question a distributed data. The system uses linguistics grammars technique that interleaves grammar and linguistics method. The question respondent is completed via parsing the input and mapping the break down tree to a data question. The system LADDER depends on a three bedded style. The first a part of the system is for casual communication Access to Navy data (INLAND), that accepts queries throughout a communication and produces an issue to the data. The queries from the midland unit of measurement directed to the intelligent data Access (IDA), that's that the second a part of LADDER. The midland half builds fragment of a question to IDA for each lower level grammar unit at intervals land input question and these fragments unit of measurement then combined to higher level grammar units to be recognized. At the sentence level, the combined fragments unit of measurement sent as a command to IDA. IDA would compose an answer that is relevant to the user’s original question in addition to coming up with the proper sequence of file queries. The third a part of the LADDER system is for File Access Manager (FAM).The task of FAM is to go looking out the position of the generic files and manage the access to them at intervals the distributed data. The system LADDER was enforced in LISP. At the time of the creation of the LADDER system was ready to technique a data that is like a database with fourteen tables and 100 attributes.”

3. Rendezvous System:

As per [1, 4]: “The Rendezvous system appeared in late seventies. In this, users may access databases via relatively unrestricted language. Throughout this Cods’ system, special stress is placed on question paraphrasing and in collaborating users in clarification dialogs once there is issue in parsing user input.”

4. Planes

As per [1, 5]: “Planes system was developed in late seventies for (Programmed Language-based Enquiry System) at the University of Illinois Coordinated geographic point. PLANES embrace academic degree West Germanic language facet with the flexibleness to grasp and expressly answer user requests. It carries out informative dialogues with the user likewise as answer obscure or poorly made public queries. This work is being assigned exploitation data based totally upon information of the U.S. Navy 3-M (Maintenance and Material Management), it is a data of craft maintenance and flight data, although the ideas area unit typically directly applied to totally different ungraded record-based databases.”

5. Philiqa

As per [1, 6]: “The Philiqa system was developed in 1977 and was observed as Philips Question respondent System, uses a grammar program which runs as a separate pass from the linguistics understanding passes. This technique is mainly committed problems with linguistics and has three separate layers of linguistics understanding. The layers square measure referred to as “English Formal Language”, “World Model Language”, and “Data Base Language” and appears to correspond roughly to the "external", "conceptual", and "internal" views of data."

6. Chat-80

As per [1, 7]: "CHAT-80 system is one in all the foremost documented natural language processing systems within the eighties. The system was enforced in Prologue. The CHAT-80 was a formidable, economical and complex system. The information of CHAT-80 consists of facts (i. e. oceans, major seas, major rivers and major cities) regarding one hundred fifty of the countries world and a little set of West Germanic vocabulary that area unit enough for querying the information. The CHAT-80 system processes Associate in Nursing English language question in 3 states as delineate."

7. Team

As per [1, 8, 9, 10]: "Team system was developed in 1987. Associate in nursing oversized a region of the analysis of that time was dedicated to immovability issues. TEAM was designed to be merely configurable by data administrators with no information of NLIDBs."

8. Ask

As per [1, 11, 12]: "This system developed in 1983, allowed end-users to point out the system new words and ideas at any purpose throughout the interaction. Raise was extremely a complete data management system, providing its own built-in data and so the power to act with multiple external databases, email correspondence programs and different portable computer applications. All the applications connected to lift were accessible to the highest user through communication requests. The user declared his/her requests in English and lift transparently generated applicable requests to the appropriate underlying systems."

9. Janus

As per [1, 13, 14, 15, 16]: "Janus system had similar abilities to interface to multiple underlying systems (databases, knowledgeable systems, graphics devices, etc). All the underlying systems could participate at intervals the analysis of a tongue request, whereas not the user ever turning into conscious of the dissimilarity of the system. Deity is in addition one in every of the few systems to support temporal queries."

10. Eufid

As per [1, 17]: "The EUFID system consists of three major modules, not count the code. Initial is analyzer module, second is mapper module and third is translator module."

11. Datalog

As per [1, 18]: "It is academic degree English data question system supported Cascaded ATN linguistics. By providing separate illustration schemes for linguistic information, general world information, and application domain information, DATALOG achieves a high degree of movability and extendibility."

3 Roadmap Of Nlidb

Research has been really active in developing interfaces for accessing structured information, from faceted search, wherever information is sorted and represented through taxonomies, to menu guided and form-based interfaces like those oared by the information and data Management (KIM) platform. These interfaces still need that the user is aware of the queried arrangement. However, casual users need to be able to access the information despite their queries not matching exactly the queried information structures. To keep with the interface analysis systems developed to support language Interfaces unit of measurement perceived because the foremost acceptable by end-users. This conclusion is drawn from a usability study that compared four varieties of language interfaces to information bases and anxious fifty two users of general background. Net users unit of measurement accustomed typing primitive queries into the text box of a search engine. Search engines like Google unit of measurement capable of respondent easy queries like what is the capital of geographical region. However, the power of connected data is at intervals the potential to answer lots of complicated queries that the answer cannot be found through Google. Also, lots of data on the net is accessible through the use of applications supported relative databases.

4. Challenges

Natural language interfaces have in the past led users to anthropomorphize the computer, or at least to attribute more intelligence to machines than is warranted. On the part of the user, this has led to unrealistic expectations of the capabilities of the system. Such expectations will make it difficult to learn the restrictions of the system if users attribute too much capability to it, and will ultimately lead to disappointment when the system fails to perform as expected as was the case in the AI winter of the 1970s and 80s. Some of the major challenges of Natural Language Interface are

Modifier attachment

The request "List all employees in the company with a driving licence" is ambiguous unless you know companies can't have drivers licences.

Conjunction and disjunction

"List all applicants who live in California and Arizona" is ambiguous unless you know that a person can't live in two places at once.

Anaphora resolution

Resolve what a user means by 'he', 'she' or 'it', in a self-referential query.

Other goals to consider more generally are the speed and efficiency of the interface, in all algorithms these two points are the main point that will determine if some methods are better than others and therefore have greater success in the market.

Finally, regarding the methods used, the main problem to be solved is creating a general algorithm that can recognize the entire spectrum of different voices, while disregarding nationality, gender or age. The significant differences between the extracted features - even from speakers who says the same word or phrase - must be successfully overcome.

Objective Of Research

Natural Language Interface to information People via laptop all round the world, access, accumulate and manipulate large amount of knowledge each second of the day. These large amounts of knowledge are situated in private personal computers or remote locations. Mostly, knowledge is kept in some quite repository system like information. Knowledge in information is sometimes managed by database management system and access to information is expedited through a special interaction language referred to as SQL or some version of it. To override the quality of SQL for non-professionals, many researchers have proposed to use Natural Language (NL). The concept of mistreatment NL has prompted the event of latest kind of process methodology referred to as language Interface to information. A language Interface to a Database (NLIDB) may be a system that permits the user to access data keeps during a database by writing requests expressed

CONCLUSIONS

We are proposing EFLEX system (A Flexible and Efficient Natural Language Query interface to databases) to tackle challenges in NLIDBs. The empirical validation of the prototype shows improved performance. We have designed a new product by name EFLEX and hopefully will be accepted and used by the society for their ease in data retrieval through queries given in natural language. We are also planning to file for patent of EFLEX (around 18 claims) with Indian Patent office, Mumbai. We validated EFLEX model using the following:

Statistical analysis technique called ANOVA showed linearity and strong correlation with state of the art. The results validated the hypothesis. ANOVA results confirm difference between proposed EFLEX System and state-of-the-art cannot be considered.

We evaluated product EFLEX using SUMI questionnaire. SUMI showed summarized quantification of user experience. The results showed that the product PS had the following features: high learn ability, need of less keystrokes and an easier interface. The scope for improvement in the product was in the following areas: proper popup of error messages, consistency, documentation and better retrieval of data. The evaluation was performed on the first version of the software. The benchmark for evaluation being PRECISE our proposed EFLEX System reaches at par or beyond performance. Our goal was to evaluate correct sql translations for natural language queries. The accuracy on text input is 93.9. The results also showed improvement as parser showed improvement. Precision, recall and response error rates are on acceptable side when compared with state of the art. EFLEX showed that it performs favorably in terms of both preciseness and complexity in queries compared to existing learning methods requiring similar amount of supervision, and showed higher robustness to changes in task complexity and word order. EFLEX eliminated the need for a formal query-formatting syntax and the formulation of a request as logical document set manipulations of specific terms.

Parsing, semantic analysis, machine learning and pattern matching constructs from state of the art are exploited leading to better or at par results.

The algorithm not just relied on key matching but added some intelligence through learning sets so that complex queries can be resolved. Detailed parsing lead to acceptable response time.

FUTURE WORK

Better lexicon learning mechanism to be identified to improvise precision of the algorithm. To improvise response time in more acceptable norms. Emergence of large data sets and scale is a future challenge as it will lead to several ambiguities arising out of multiple interpretations. Representation post translation in a presentable and user understandable way. Personas, profiling will lead to better query resolution. Interface with auditory natural language query system. Auto language translation mechanism for auditory queries.

REFERENCES

Androutsopoulos, G.D. Ritchie, P. Thanisch, "Natural Language Interfaces to Databases – An Introduction",arXiv:cmp-lg/9503016v2 16 Mar 1995

W.A. Woods, R.M. Kaplan, and B.N. Webber", The Lunar Sciences Natural Language Information System: Final Report. BBN Report 2378", Bolt Beranek and Newman Inc.,Cambridge, Massachusetts, 1972.

G. Hendrix, E. Sacerdoti, D. Sagalowicz, and J. Slocum, "Developing a Natural Language Interface to Complex Data." ACM Transactions on Database Systems, 3(2):105–147,1978.

E.F. Codd, "Seven Steps to RENDEZVOUS with the Casual User.", In J. Kimbie and K. Koffeman, editors, Data Base Management. North-Holland Publishers, 1974

D.L. Waltz, "An English Language Question Answering System for a Large Relational Database.", Communications of the ACM, 21(7):526–539, July 1978.

R.J.H. Scha, "Philips Question Answering System PHILIQA1.", In SIGART Newsletter, no.61. ACM, New York, February 1977.

D. Warren and F. Pereira, "An Efficient Easily Adaptable System for Interpreting Natural Language Queries", Computational Linguistics, 8(3-4):110–122, July-December 1982.

B.J. Grosz, "TEAM: A Transportable Natural-Language Interface System.", In Proceedings of the 1st Conference on Applied Natural Language Processing, Santa Monica, California, pages 39–45, 1983.

B.J. Grosz, D.E. Appelt, P.A. Martin, and F.C.N. Pereira. "TEAM: An Experiment in the Design of Transportable Natural-Language Interfaces", Artificial Intelligence, 32:173–243, 1987.

P. Martin, D. Appelt, and F. Pereira, "Transportability and Generality in a Natural-Language Interface System.", In B.J. Grosz, K. Sparck Jones, and B.L. Webber, editors, Readings in Natural Language Processing, pages 585–593. Morgan Kaufmann Publishers, California, 1986.