Information Retrieval System Cantered on RDBMS: A Survey

Ms. Sangeeta Vishwakarma[1] Mr. Prabhakar Sharma[2]
[1]M. Tech. Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Raipur Institute of Technology, Raipur C.G., India

Abstract— Relational Database management systems (RDBMS) are extensively used software products in numerous types of systems. As we all are familiar with natural language which is the main communication means for humans, but this causes it not easy to handle damaged information. Earlier, research work while querying data from relational databases frequently goes through one of two ways: the keyword-based approach and the structured query approach. Both the ways have their recompense and disadvantages. The structured query approach, while expressive and powerful, is not easy for naive users. The keyword-based approach is very friendly to use, but cannot express complex query intent accurately. This paper emphasis on Natural Language based query processor and we have discussed different approach of information retrieval system. Natural Language Processing is becoming more important in the field of Human Computer Interaction.

Key words: NLP, SQL, IR, SPARQL

I. Introduction

IR (Information retrieval) systems are worn for discovery, within a hefty text database, containing details desirable by a user. The intricate and weak understood semantics of documents and user queries has prepared feedback and alteration important distinctiveness of any IR systems. Hence natural language based IR system will be much favorable. Natural language processing based IR systems are enormously capable to symbolize and manipulate the intricate query as complex and uncertain relationship presented among them. The conventional query in relational database management system is not capable of satisfying the needs for dealing with queries which are in natural language.

What does 'Processing' Natural Language means? It Mean to 'Process' Human Language in computer language. NLP actually is used to analyze human expression or information by transforming an input expression into a source code (computer language) representation of that expression for use in further processing. Parsing creates a representation of the structure, or syntax, of an expression. Parsing can also refer to the creation of a representation of the meaning of an expression. Parsing actually reconstructs the input into are presentation more suitable for a specific application. The conflicting feature of 'processing' language 'generates' natural language from code representing information that is to be relayed to a human. Processing natural language combines 'parsing' and 'generation' with additional manipulations to provide a computer with the services necessary to allow simplistic communication with human’s communication. One of the vital issues of the NLP (Natural language Processing) fundamental is of natural language understanding. The procedure of building computer based programs which can understand natural language and its techniques involves three major issues. The first one relates to the assumed process, the second one is to represent and understand meaning of the linguistic input, and the third one to the world knowledge. Hence, an NLP system would start at the word level – to determine the morphological structure- it is the study of how things are put together, nature -such as part-of-speech of the word and then may switch on to the sentence level – to decide the word order, its grammar, meaning of the whole sentence and then to the whole environment or domain. A particular word or a sentence may have a particular meaning in a given domain, and may be associated to many other words and sentences in the given domain.

In natural language processing, user query is passed to spell correction module then query string is divided into tokens, afterwards query mapping has been done and corresponding SQL (Structured Query Language) query will be generated. Earlier, research work while querying data from relational databases frequently goes through one of two ways: the keyword-based approach and the structured query approach. Both the ways have their recompense and disadvantages. The structured query approach, while expressive and powerful, is not easy for naive users. The keyword-based approach is very friendly to use, but cannot express complex query intent accurately. In contrast, natural language has both advantages to a large extent: even naive users are able to express complex query intent in natural language [Fei Li et. al. 2014].

In this paper we have discussed different literature based on NLP in context of Relation data base management system. Further in section II we have why we have motivated toward study, section III elaborates several literatures, in section IV we provide comparative table among different approach.

II. Motivation

The different type search engine like Google, binge, AltaVista is used to fetch the information from the database by easy language. The non-technical employee they don’t understand the database and query cannot access the database. The proposed system is performing work as a search engine where users can fetch the information from the database by natural human sounding language. The previous existing system doesn’t able to solve queries in one easy statement. Also in previous existing system, common people can’t able to fetch information from the database of any organization so these motivated to propose a new system of query optimization using NLP (natural language processing). Hence proposed system is useful to solve queries in one simple statement using NLP (natural language processing). This proposed study is very fruitful to handle queries in user friendly approach. It gives hope to fetch information in easy way in various fields of government, education, and medical. And different government employee who doesn’t know about the computer and the database can easily access the database by the simple sentences or by the simple human sounding language.

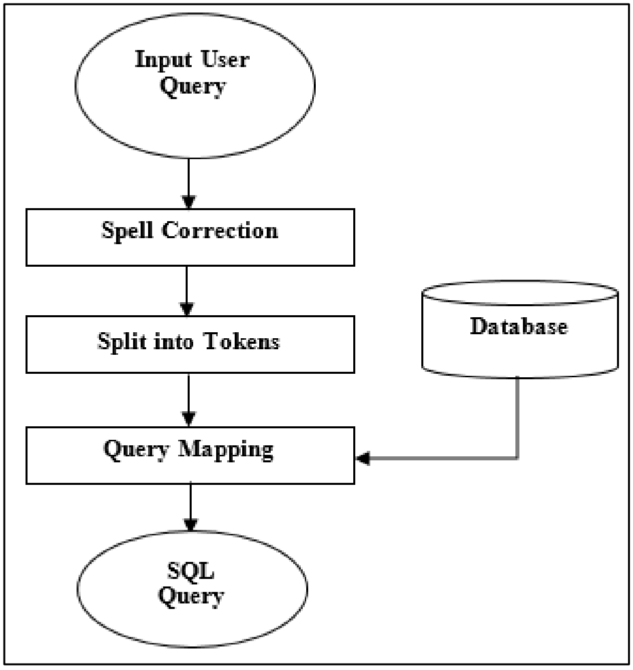

Fig. 1: Earlier System

III. Literature Survey

The literature review is very significant in any research project as it clearly establishes the need of the work and the background development. It generates related queries regarding improvements in the study already done and allows unsolved problems to emerge and thus clearly define all boundaries regarding the development of the research project. It also gives a clear picture about the research topic that how much work has been done. Also what were the positive outcomes in the research work? What can be the future aspects in the research field? By studying the various papers, it is found that all data retrieving concepts have a need of query to fetch data. But the major problem is that there is no easiest scope of accessing data based on the sentences based query. The important thing of accessing information is, the way of processing the query to understand the data and retrieving the data in a simple manner. So the proposed system will provide Query Optimization using NLP (Natural Language Processing) on the basis of simple query based sentence.

Bihani R. et al (2014) has proposed converts the natural language query into SQL (Structured Query Language) which is a database programming language. We will perform the following steps for the transformation of query from natural language to database query (SQL) sequentially as listed in the following points: First we will accept the string in natural language. After accepting the query we will check the query for misspelled words (if any) using word pair mining. After that we will split query into tokens. After getting tokens we perform the SQL mapping for transformation. This paper conclusion is the search interface that will be applicable for the online applications, provides ease to the user by reducing their part of recalling complex database language syntax. Our system also corrects the spelling mistakes did by the users, automatically and also takes care of grammatical errors. Proposed system generates output query irrespective of the database.

Cambria E. et all (2012) has proposed Natural Language Processing Research. NLP research according to three different paradigms, namely: the bag-of words, bag-of-concepts, and bag-of-narratives models. Borrowing the concept of ‘jumping curves’ from the field of business management, this survey article explains how and why NLP research has been gradually shifting from lexical semantics to compositional semantic and offers insights on next-generation narrative-based NLP technology. This paper conclusion is Web where user-generated content has already hit critical mass, the need for sensible computation and information aggregation is increasing exponentially, as demonstrated by the ‘mad rush’ in the industry for ‘big data experts’ and the growth of a new ‘Data Science’ discipline. The democratization of online content creation has led to the increase of Web debris, which is inevitably and negatively affecting information retrieval and extraction. To analyze this negative trend and propose possible solutions, this review paper focused on the evolution of NLP research according to three different paradigms, namely: the bag-of words, bag-of-concepts, and bag-of-narrative models. Borrowing the concept of ‘jumping curves’ from the field of business management, this survey article explained how and why NLP research is gradually shifting from lexical semantics to compositional semantics and offered insights on next-generation narrative based NLP technology.

Akshay G. Satav,et al (2013) [8] has proposed a system which gives interface between user and computer, in the form of database query language. Also the spelling corrections of misspelled words in query are getting correct. Search session consist of set of user queries fired by a single user within a short time period. Many of the search session consisted of misspelled word and there corrected spelling. We segment the query stream from user into sessions. If the time between the two queries exceeds a certain period of time then we put session boundary between the two queries. The time period between the two queries is kept short, the reason behind this that it is observed that user firing query is corrects the misspelled word immediately After recognizing the mistake. Finally we make pairs misspelled and corrected word across the whole sessions and give it a frequency to each word pair and later on discard the word pair that has lower frequency. The following table shows some example of correct and misspelled word. Query Mapping In this the natural language query is taken in English Language, any form of statement (WH type questions, word, any type of statement etc.). Then the query in English language mapped according to syntax of SQL query that provides user the accurate data from database after execution of mapped SQL query. The accuracy in mapped Query is focused here. This paper conclusion is search interface that will be applicable for the online applications, provides ease to the user by reducing their part of recalling complex database language syntax. Our system also corrects the spelling mistakes did by the users, automatically and also takes care of grammatical errors. Proposed system generates output query irrespective of the database. Future scope of the work is proposed system converts the Natural Language query in SQL query, also it corrects the spelling automatically. In future we will be focusing towards reformulation of query and transforming the Natural Language Query in the SPARQL database language required for the semantic search to provide the more accurate result as the user require instead of checking the multiple link as we do at present. Today's generation also use lots of abbreviation so keeping it in mind we will be reformulating the query (for e.g. LA by Los Angeles and TOI by Times Of INDIA) so in the future we will try to implement the above discussed goals.

Mr. Klein Dan and Christopher D. Manning [9] says that two databases were used, the chemical analyses and the literature references. The program used an Augmented Transition Network (ATN) parser and Woods' Procedural Semantics. The system was informally demonstrated at the Second Annual Lunar Science Conference in 1971 suppose considers a database say ORACLE within this oracle database we have placed certain tables, which are properly normalized. Now if the user wishes to access the data from the table, he/she has to be technically proficient in the SQL language to make a query in the ORACLE database. The project eliminates this part and enables the end user to access the tables in his/her language.

Douali N. et al (2012) has proposed focus on the representation of such information in an appropriate language so as to facilitate the execution of guidelines in CDSS. Bayesian network, belief network or directed acyclic graphical model is a probabilistic graphical model that represents set of variables (nodes) and their conditional interdependencies. Nodes can represent medical observables, such as medical goals, therapies, examinations and clinical signs while their directed interconnected arches can bear measures of quantitative or qualitative origin to describe the given relationship. The nodes are known with certainty or even uncertainty described by a subjective probability. Subjective probabilities express the degree of a person's belief, given a certain knowledge background of him. This notion of probability differs from the most used classical probability. Thus objective or Bayesian probabilities can describe a value of belief to unique events that are not repeatable.

Mr. Elworthy D. [6] Proposed the unprecedented large volume of semi-structured data has exacerbated the need for an easy-to-use query interface for semi-structured data sources. Natural language interfaces and keyword search techniques that take advantage of the data set structure make it very easy for ordinary users to access the data. In this paper, it is introduced that important challenges that lie in the way of building an elective and efficient keyword and/or natural language query interface for semi-structured databases. It shows that the current approaches to this problem rely heavily on heuristics that are intuitively appealing but ultimately ad hoc. Their assumptions are valid for some domains, database designs, and/or schema structures but they are not correct in general. Thus, these often retrieves false. Positive answers, overlook correct answers, and cannot rank answers appropriately.

IV. COMPARISON

S. No.	Author/Paper title/Year	Name of Algorithm /Method	Description
1.	Mathias Soeken et. al. Automating the Translation of Assertions Using Natural Language Processing Techniques FDL Proceedings ECSI 2014 [1]	High abstraction level and low abstraction level assertions.	Author presented an algorithm that automates the translation of natural language assertions into System Verilog Assertions using natural language processing techniques. Instead of manually translating each assertion separately.
2.	Ryuichiro Higashinaka et. al. Towards an open-domain conversational system fully based on natural language processing Proceedings of COLING 2014 [2]	Rule-based system.	This paper proposes an architecture for an open-domain conversational system and evaluates an implemented system. The proposed architecture is fully composed of modules based on natural language processing techniques.
3.	Anupriya et. al. Fuzzy Querying Based on Relational Database IOSR-JCE Jan-2014 [3]	Fuzzy Logic	This paper mainly discusses the realization of fuzzy query through fuzzy theory and SQL combined C#. Also, a real life application of fuzzy query based on relational database (the Patient Information database) is provided.
4.	Lei Zou et. al. Natural Language Question Answering over RDF — A Graph Data Driven Approach SIGMOD 2014 [5]	Graph Data Driven Approach	Author proposes a semantic query graph to model the query intention in the natural language question in a structural way, based on which, RDF Q/A is reduced to subgraph matching problem. More importantly, author resolve the ambiguity of natural language questions at the time when matches of query are found. The cost of disambiguation is saved if there are no matching found
5.	Joao P. Carvalho et. al. A Critical Survey on the use of Fuzzy Sets in Speech and	Fuzzy Set	This paper shows how the use and applications of Fuzzy Sets (FS) in Speech and Natural Language Processing (SNLP) have seen a steady decline to a point where FS are virtually unknown or unappealing for most of the

	Natural Language Processing IEEE 2012 [7]		researchers currently working in the SNLP field, tries to find the reasons behind this decline, and proposes some guidelines on what could be done to reverse it and make FS assume a relevant role in SNLP.
6.	Tareq Abed Mohammed et. al./ Intelligent Database Interface Techniques using Semantic Coordination/ IEEE 2018 [4]	Semantic Coordination	The intelligent interface utilizes semantic coordinating procedure to change natural language query to Structured Query Language (SQL) by depending lexicon and set of creation guidelines. The lexicon comprises semantics sets for tables and sections.

V. Conclusion

We have experienced different literary works and discovered some bottleneck in normal dialect inquiry handling which are as per the following:

1) Spelling correction for mistakes made by the user while firing query map the natural language query into database query language.
2) Semantic checking over user query.
3) In earlier system user has to fire queries in WH type question.
4) Earlier system did not supported single word query.
5) A limited data dictionary was used in which words were updated after regular period of time.
6) Natural language query had to be in double quotes (“”).

Question words	Meaning	Examples
who	person	Who's that? That's Nancy.
where	place	Where do you live? In Boston
why	reason	Why do you sleep early? Because I've got to get up early

Table 1: WH Type Question Example

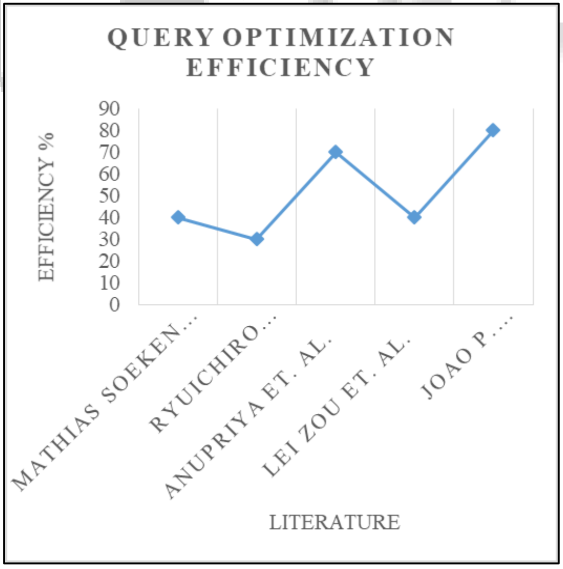

Fig. 2: Performance comparison of Efficiency

The conventional query in relational database management system is not capable of satisfying the needs for dealing with queries which are in natural language. From different literature we can conclude that fuzzy based query processor are efficient query processing but still there is chance of improvement i.e. we can optimize the query processing so that query processing time decreases.

References

[1] Mathias Soeken, Christopher B. Harris, Nabila Abdessaied, Ian G. Harris, Rolf Drechsler, Automating the Translation of Assertions Using Natural Language Processing Techniques FDL Proceedings | ECSI 2014.

[2] Ryuichiro Higashinaka1, Kenji Imamura1, Toyomi Meguro2, Chiaki Miyazaki Towards an open-domain conversational system fully based on natural language processing Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 928–939, Dublin, Ireland, August 23-29 2014.

[3] Anupriya, Prof. Rahul Rishi Fuzzy Querying Based on Relational Database OSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. I (Jan. 2014), PP 53-59.

[4] Tareq Abed Mohammed et. al./ Intelligent Database Interface Techniques using Semantic Coordination/ IEEE 2018.

[5] Lei Zou, Ruizhe Huang, Haixun Wang Natural Language Question Answering over RDF — A Graph Data Driven Approach SIGMOD’14, June 22–27, 2014.

[6] Arash Termehchy, Keyword and Natural Language Query Processing for Semi-Structured Data Sources, Proceedings of the Third SIGMOD PhD Workshop on Innovative Database Research (IDAR 2009) June 28, 2009.

[7] Joao P. Carvalho et. al. A Critical Survey on the use of Fuzzy Sets in Speech and Natural Language Processing IEEE 2012.

[8] Akshay G. Satav, Archana B. Ausekar, Radhika M. Bihani, Mr Abid Shaikh A Proposed Natural Language Query Processing System WARSE Volume 3, No.2, March - April 2014.

[9] Jasmeen Kaur *, Bhawna chauhan *, Jatinder Kaur Korepal, Implementation of Query Processor Using Automata and Natural Language Processing, International Journal of Scientific and Research Publications, Volume 3, Issue 5, May 2013