International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 4 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195438

A Novel IR for Relational Database using Optimize Query Building

Sangeeta Vishwakarma[1], Avinash Dhole[2]

[1]M. Tech. Scholar, Department of Computer Science and Engineering, Raipur Institute of Technology, Raipur C.G. India

[2]Associate Professor, Department of Computer Science and Engineering, Raipur Institute of Technology, Raipur C.G. India

ABSTRACT

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. 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. We have proposed the use of query optimization approach to convert complex NLP query to SQL query, SPAM word removal, POS tagger applied over NL query and concluded that execution time lesser when query size increases.

Keywords : Fuzzy, IR,NLP, Precision

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. 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. Unsatisfactory information can be incoherent, inexact, unclear, uncertain or vague. However, the Complexity is limited in specific data processing and is not capable of directly expressing fuzzy concepts of natural language. Earlier system work based on the architecture as fig. 1 in which user get the database information in his/her language.

Motivation towards this study is 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.

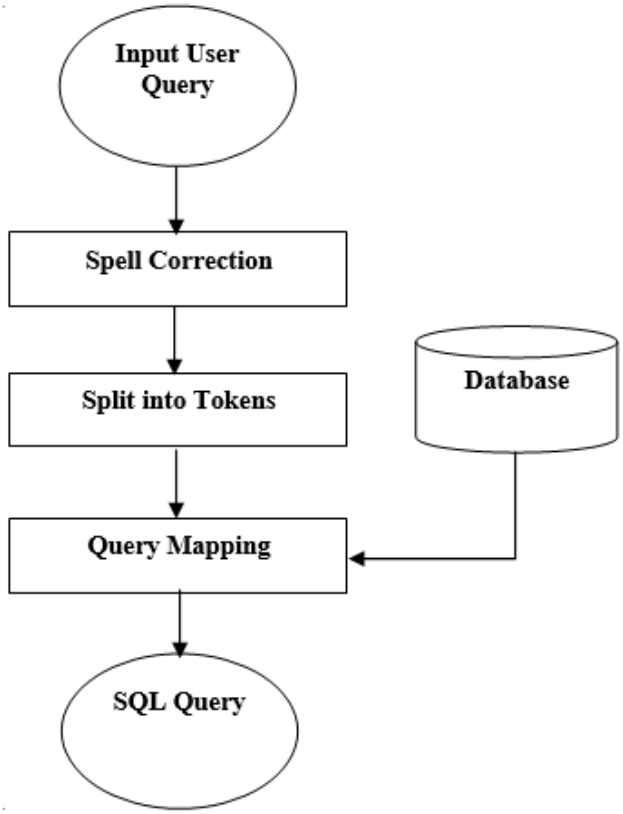

Fig 1. Earlier System

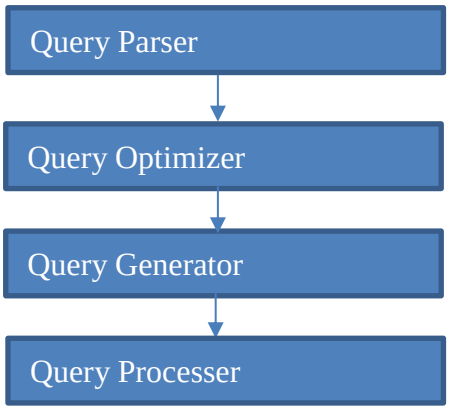

Fig 2. NLP Parser

The area of query optimization is very large within the database field. It has been studied in a great variety of contexts and from many different angles, giving rise to several diverse solutions in each case. The purpose of this chapter is to primarily discuss the core problems in query optimization and their solutions. Actually the main purpose of NLP parsing relates to query optimization where it performs the work of separating the grammatical meaning of each word in term of noun, pronoun, verb and etc.The path that a query traverses through a DBMS until its answer is generated. The system modules through which it moves have the following functionality:

- The Query Parser checks the validity of the query and then translates it into an internal form, usually a relational calculus expression or something equivalent.
- The Query Optimizer examines all algebraic expressions that are equivalent to the given query and chooses the one that is estimated to be the cheapest.
- The Code Generator or the Interpreter transforms the access plan generated by the optimizer into calls to the query processor.

The Query Processor actually executes the query.
Further in this paper in section II we have gone through background study and given tabular comparison among different literature, in section III we have formulized the problem in existing system, in section IV proposed methodology discussed, in section V we have given performance comparison at last section conclusion and some future direction given.

II. LITERATURE SURVEY

Soeken M. ei al (2013), has proposed The starting point is a set of NLAs given in terms of English sentences (indicated in the figure by zigzag lines). These assertions are automatically partitioned into subsets of high abstraction level and low abstraction level assertions in the first step. The idea is that high level assertions contain implicit and imprecise information which impedes automatic translation and therefore need to be translated manually as in the conventional flow. Consequently, high level assertions are not further considered in the remainder of the flow. In the second step the determined low level assertions are partitioned into clusters of similar sentences.

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.

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.

III. PROBLEM IDENTIFICATION

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

- Problem in query optimization to large distributed databases: Query optimization in large distributed databases explicitly needed in many aspects of the optimization process, often making it imperative for the optimizer to consult underlying data sources while doing cost based optimization. This is not only increases the cost of optimization, but also changes the trade-offs involved in the optimization process significantly. The leading cost in this optimization process is the "cost of costing" that traditionally has been considered insignificant. In a large scale distributed system, both data access and computation can be carried out at various Sites. In a Distributed database cost must be divided into multiple currencies, but typically there are other costs that are valuable to expose orthogonally, including response time, data freshness, and accuracy of computations.
- Spelling correction for mistakes made by the user while firing query map the natural language query into database query language.
- POS tagger not applied to find out phrase in input query.
- Natural language query had to be in double quotes ("").
- In earlier system works onWho/How type question.

- Earlier system did not supported single word query.
- Difficulty of translating user-specified query structure to the actual schema structure in the database.
- Stop/spam word removal may not present in earlier system.

IV. PROPOSED METHODOLOGY AND RESULT AND DISCUSSION

Firstly we take the input user query in the database. Our Next step is to work on spell correction in sentence then next steps in split into the tokens then next steps in semantic analyzer in the sentence. Next steps in query builder in the database and solved the query and finally find the result.

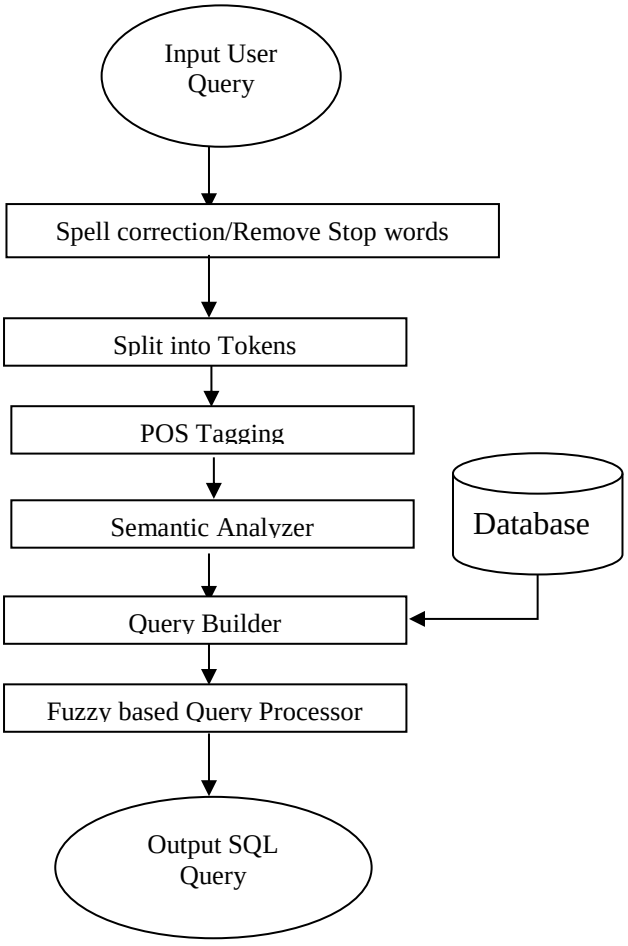

Fig 3. Proposed System Layout

For instance, in patients' relational database, to deal with a query statement like "young age and good heartbeat or low blood sugar ", it is difficult to construct SQL because the query words are fuzzy expressions. In order to obtain query results, there are two basic methods of research in the use of SQL Combined fuzzy theory in DBMS. The first is still to build a classic relation database, only to modify or extend SQL query by transforming query conditions to a fuzzy scope. After that, change it to precise SQL clause. This procedure is easy and consistent with ordinary query, but lacks flexibility. Sometime it is apt to produce query errors. The second approach is to assume that the database is fuzzy sets and fuzzy logic is used to make it easier and more human consistent. This is mainly done by constructing a database model based on fuzzy logic. When designing this database and modify its data structure, many tables including fuzzy fields' values may be added. These tables could be transformed from row tables. For instance, in the patients' relational database "Age", "Blood pressure", and "heart beat" etc. are fuzzy fields. When data is being inputted, the precise data may be transformed to fuzzy data and stored in database.

Algorithm

1. Select database.
2. Input NLP Query
3. Spell correction of Input query.
4. Stop/ spam word removal from input query.
5. Divided input sentence in to tokens.
6. Apply POS tagging.
7. Semantic analyzer will analyze NLP query and NL word with RDBMS syntax.
8. Query will generated by fuzzy approach.
9. Query passed to DBMS.

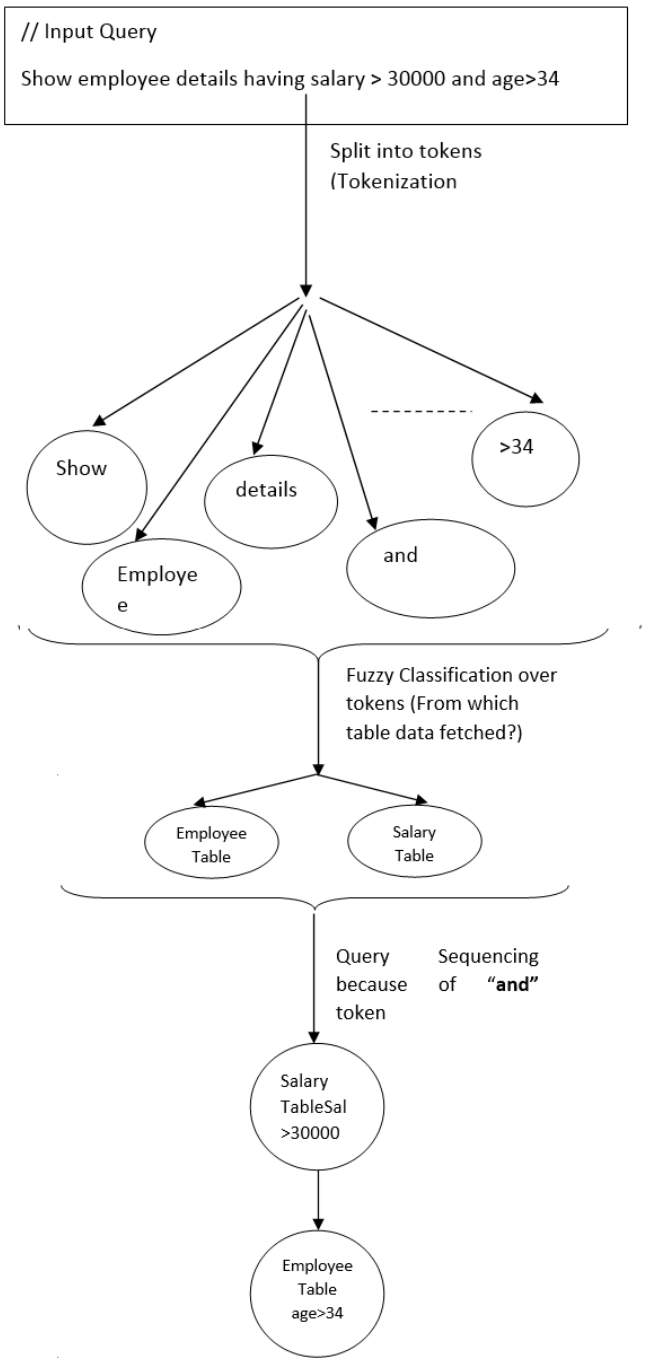

Fig 4. Fuzzy Approach for NLP

For implementation of novel IR we have used Mysql for RDBMS and JAVA 1.8 for IR. Fig.-4 shows the main user interface of our IR, in which user has to select database and table name, further it will describe the selected table i.e. it will show the column names present in selected table.

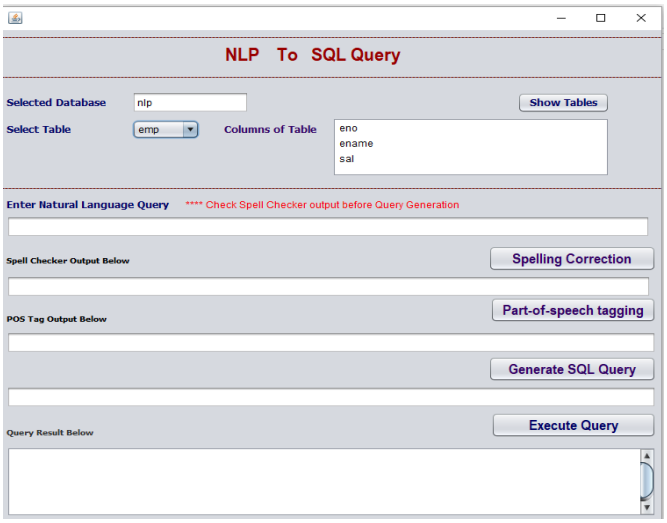

Fig.-6 Main UI

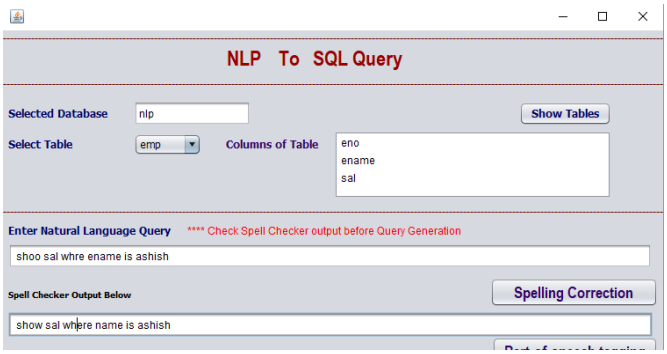

Fig.7 Spell correction

Fig. 5 depicts the spell correction. In which:

Input Query as:

Shoo sal wheee name is sangeeta

Output spell Correction

Show sal where name is sangeeta

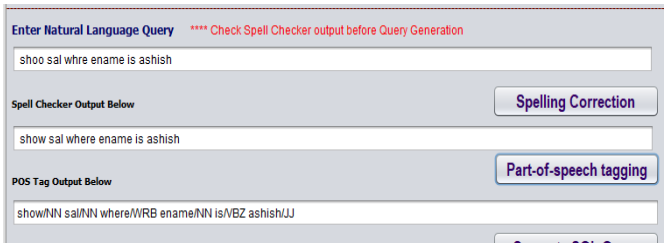

Fig.-8 POS Tagging

Input

showsal where name is sangeeta

Output

show/NNsal/NN where/WRB ename/NN is/VBZ sangeeta /JJ

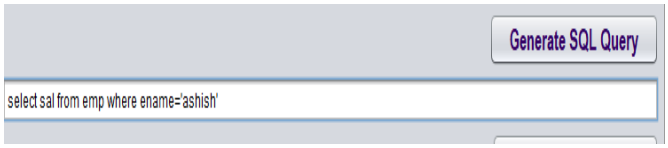

Fig.-9 Sql Query Generated

Input

showsal where name is ashish

Output

select sal from emp where ename='sangeeta'

Fig.-10 Query Execution

Table-1 Execution Time

S. No.	Query	Length	Execution Time (ms)
1.	show sal where ename is ashish	30	30 ms
2.	show details of name where name is starting from G	50	32 ms
3.	CITY,ADDRESS_ID where SUPPLIER_ID IS A032	41	30ms
4.	show sal,ename,eno where ename is sangeeta *** Query Optimized because NL query ask for all column it converted to *	42	30ms

Table-2 Accuracy

S. No.	No. of NL query Sample	Correct Query Conversion
1.	57 (Without change in NL Query)	55
2.	57 (With change in NL Query in column name listed in UI)	56

V. CONCLUSION

Natural Language Processing can bring powerful enhancements to virtually any computer program interface. This system is currently capable of handling simple queries with standard join conditions. Because not all forms of SQL queries are supported in optimization. Hence optimization of query from one single and simple statement is desirable to fetch more and more knowledge by understanding the language in easy way. Achieved almost 97% accuracy over number of sample stated in table-2.

VI. REFERENCES

[1]. Mathias Soeken, Christopher B. Harris, Nabila Abdessaied, Ian G. Harris, RolfDrechsler, 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: 2278-8727 Volume 16, Issue 1, Ver. I (Jan. 2014), PP 53-59.

[4]. Antonio Lieto, Andrea Minieri1, Alberto Piana1 and Daniele P. Radicioni A Knowledge-Based System for Prototypical Reasoning ISSN: 0954-0091 print/ISSN 1360-0494 oct 2014.

[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]. Andre Freitas, Edward Curry Natural Language Queries over Heterogeneous Linked Data Graphs: A Distributional- Compositional Semantics Approach ACM 978-1-4503-2184-6/14 Feb 2014.

[7]. Fei Li, H. V. Jagadish NaLIR: An Interactive Natural Language Interface for Querying Relational Databases SIGMOD'14, June 22–27, 2014.

[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]. 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..

[10]. Higashinaka Ryuichiro, Imamura Kenji , Meguro Toyomi, Miyazaki Chiaki Kobayashi Nozomi , Sugiyama Hiroaki , Hirano Toru, Makino Toshiro ,Matsuo Yoshihiro Towards an open-domain conversational system fully based on natural language processing

Cite this article as :

Sangeeta Vishwakarma, Avinash Dhole , "A Novel IR for Relational Database using Optimize Query Building", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 4, pp. 251-257, July-August 2019. Available at doi : https://doi.org/10.32628/CSEIT195438 Journal URL : http://ijsrcseit.com/CSEIT195438