ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

The Amalgamation of NLP with Text Categorization

Amisha Shingala[1], Paresh Virparia[2], Anjali Jivani[3]

Assistant Professor, Dept. of MCA, SVIT, Gujarat Tecnological University, Vasad, Gujarat, India[1]

Professor, Dept. of CS, Sardar Patel University, V. V. Nagar, Gujarat, India[2]

Associate Professor, Dept. of CSE, The M. S. University of Baroda, Vadodara, Gujarat, India[3]

ABSTRACT: Text Mining (TM) and Natural Language Processing (NLP) are very closely related to each other. Though TM is not really associated with the semantic and the syntax of a language, it does try to link documents or unstructured data in such a way that at the end of the day we can say those documents are semantically closely associated. TM has a number of subtasks like Text Clustering, Text Summarizing, Text Categorizing, etc. In this paper we are discussing research related to converting a simple English language question to its equivalent Structured Query Language (SQL) statement for a novice to fetch information from a database. This tool that we have developed has been named the N-ELIDB – Natural English Language Query Interface to Database. This is a heuristic tool which after generating the SQL, tries to categorize the questions using the concept of Text Categorization. The previous part is for a normal user and the later part is for the Text Miner.

KEYWORDS: NLP; Text Mining; Text Categorization; Text Classification; Information Retrieval

I. Introduction

The usage of computer has increased manifold with the advent of Database Management System (DBMS) and with the dawn of the Internet. Everyday most of the people are interacting with information stored in database, usually, through the specialized and predefined programs written for such interaction. The limitation of such programs is that it restricts the interaction between user and database to predefined set of queries. This is against the basic objective/principle of DBMS system which is developed to support ad-hoc queries. Only few people who have knowledge of database structure and formal database language such SQL can retrieve the desired information from database.

A novice user having no knowledge of database structure and formal database query language cannot retrieve desired information if it is not supported by well thought application. Hence, it was a need of an hour to improve human computer interface that allows people to interact with the database in their natural language (such as English). The existing and effective Natural Language Interface to Database (NLIDB) system has high potential to mimic computer systems as conversational system thereby making it easy to use by large mass of people enhancing its utility value. At the end, several systems were designed and developed by researchers that support NLIDB. Some of the well-known NLIDB systems are: LIFFER, LUNAR, BANKS, ELF, NLBEAN, CINDI, SQL-TUTOR, SQ-HAL, Step, etc.

The objective of the research was, therefore, to provide more efficient and user friendly Natural-English Language Query Interface to Database (N-ELIDB) system that allows non-technical users, who are not aware of formal database language (such as SQL), to interact with database using their own English language. The system proposes the general framework for efficient processing of natural language query to extract intended information from the database. These statements that are converted to SQL queries are then being used as the training set for the next part of the research related to Text Categorization.

Text Categorization is a subtask of Information Retrieval that allows users to browse more easily the set of texts of their own interests, by navigating in category hierarchies. Generally statistical methods or the Machine Learning methods are used to categorize textual data. The most popular way being using the bag-of-words approach and thereafter the Vector Space Model. Typical natural language structures, e.g., morphology, syntax and semantic are completely neglected in the development of the classification functions.

10.15680/ijircce.2015.0304123

In this paper we discuss how a NLP based output can become the input to design classifiers which could be effectively be used in Text Mining Applications.

II. RELATED WORK

In [14], Conceptual Query Language/Natural Language (CQL/NL) is filtered for search predicates derived from conceptual schema constructs. Based on the identified search predicates, CQL/NL uses a set of predefined natural language templates to compose a natural language explanation of the query. In [11] Lifer/Ladder, has designed a natural language interface to a database of information about US Navy ships. This system used a semantic grammar to parse questions and query a distributed database. D L Waltz [10] developed Programmed Language-based Enquiry System- PLANES- at the University of Illinois Coordinated Science Laboratory. It carries out clarifying dialogues with the user as well as answer vague or poorly defined questions. Nalix [12] developed a generative interactive natural language query interface to an XML database. The system can accept an arbitrary English language sentence as query input, which can include aggregation, nesting, and value joins, among other things. SQ-HAL[13], a system which provides natural language interface to database. The major drawback of this system is if appropriate database driver not installed, program will not work. The database table names and column names have to be valid English words. It is not capable of determining relationship between tables. MASQUE [4], a system created by Androutsopoulos et, which is a powerful and portable natural language front end for Prolog databases. It answers written English questions by generating Prolog queries that are evaluated against the Prolog database.

III. THE N-ELIDB ARCHITECTURE

The operational methodology of N-ELIDB system has two major components as shown in Figure 1. The components are: (a) Linguistic Component and (b) Database Component. The Linguistic component discusses (i) Morphological Analysis like query pre-processing & context resolution, word-based n-gram processing, stop words removal, spelling check, domain mapping and knowledge base management (ii) Lexical Analysis like identifying token type and checking for attribute token (iii) Syntactic Analysis using Stanford Parser and Multi-Liaison algorithm, (iv) Semantic Analysis using WordNet for semantic representation of lexicon and proper noun resolution. The Database component discusses (i) SQL Query Generator which processes intermediate query representation and generates SQL query for DDL, DML & SELECT statements, and (ii) SQL Query Execution using database adaptor.

In short, the Linguistic Component translates the natural language input to an expression of Intermediate Query Representation (IQR), which is subsequently passed to Database Component for generation of Structured Query Language (SQL) statement. The resulting SQL statement is then executed by the database management system. The Linguistic Component consists of morphological analysis, query pre-processing & context resolution, lexical analysis, syntactical analysis and semantic analysis; and Database Component consists of SQL query generation and SQL query execution.

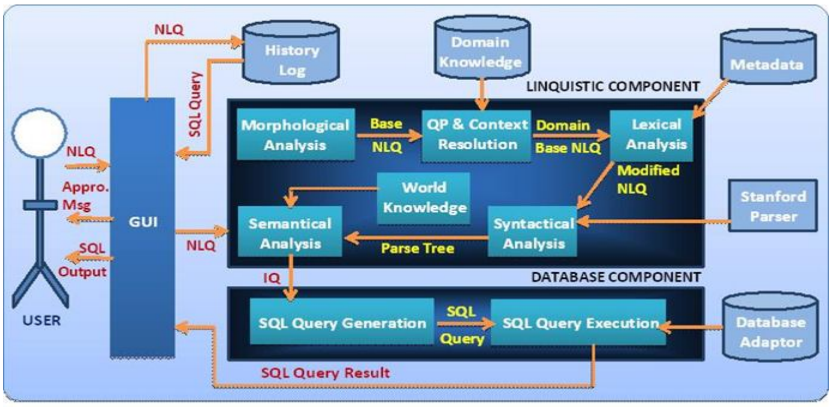


Figure 1: The N-ELIDB Architecture

A. *The Linguistic Component*

The word Linguistic means a study of language. It has mainly three aspects to study which includes language form (syntax), language meaning and language context. The computational linguistic is an interdisciplinary field known and used by computer scientists for processing of natural language. The linguistic component deals with various analyses such as morphological analysis, query pre-processing & context resolution, lexical analysis, syntactic analysis and semantic analysis.

Morphology is the study of the way words are built up from smaller meaning-bearing units called morphemes [1]. It is also known as Lemmatization, which is a process of analyzing the token morphologically in order to find their basic forms. For example: a student has two stems of a morpheme: one is student and other is students. It takes two kinds of knowledge to correctly search for singular and plural of these forms. It allows obtaining information about words that form the sentence. For this purpose, it uses a stem dictionary with root form of words.

The natural language input first undergoes a pre-processing phase in which it identifies the domain that pertains to the input query. For this, it tokenizes the input, performs morphological analyses of the words and looks them in lexicon dictionary to retrieve their syntactic and semantic properties. The pre-processing of the input query includes: (a) stopwords removal(b) word based n-gram generation and its conversion into base words (c) spelling check (d) identifying domain, and (e) knowledge reuse. The stopwords removal, n-gram generation and conversion to base words as well as spell check are the most simple and general steps that are followed by any NLP or TM algorithms. The next step is the lexical analysis and then the syntactic analysis. The Stanford Parser has been used to provide a representation of grammatical relations between words in the input sentence. [6]. The Multi-Liaison Algorithm [7] has been used to find the subjects, objects and verbs in the sentence.

B. *The Database Component*

Identifying domains is a very important step. The degree of correct interpretation and processing of natural language queries relies on existence of the exhaustive domain specific lexicon [5]. The context can be resolved by identifying the domain class in form of <database name><domain name>< key terms><attribute terms> as shown in Table 1.

Table 1: Domain Identification Mapping

Database name	Domain name	Key terms	Attribute terms
Student	Student	roll no, name, villa, residence, phone no, mobile no, department, etc.	stud_id, contact_no, address, birthdate, studnm, city, sem_no, branchid, gender, etc.
Exam	Student	exam, percentage, semester, branch	examid, examnm, sem, mks, etc.
Library	Student, Faculty	books, return, issue, library, author, edition, title, publisher, etc.	bookid, booknm, isbnno, authornm, publisherid, etc.
Branch	Student	department, course	branchid, branchnm

Syntax driven semantic analysis can be thought as the meaning of the sentence composed by the meaning of its parts [2]. Here, the input is passed through a parser to derive its syntactic structure. The syntactic structure is, then, passed to a semantic analyzer to determine its contextual meaning. For example, syntactic constitution means *sentence* -> *noun* | *verb* and the semantic attachment leads to the semantic property of noun (proper noun, common noun, etc.) or a verb. The output of the semantic interpretation module gives a logical expression of the words in the lexicon, to generate the logical query. The approach used here is an Intermediate Query (IQ) representation which can express the meaning of user input in terms of high level concepts, independent of database structure. For example, if the users input in the natural language query as:

"Display all students who stay in Baroda city", which can be translated into logical form such as:

acomp(give-1, stud-2)

```
dep (location-4, who-3)
dep(stud-2, location-4)
nn(city-7, Vadodara-6)
prep_in(location-4, city-7)
```

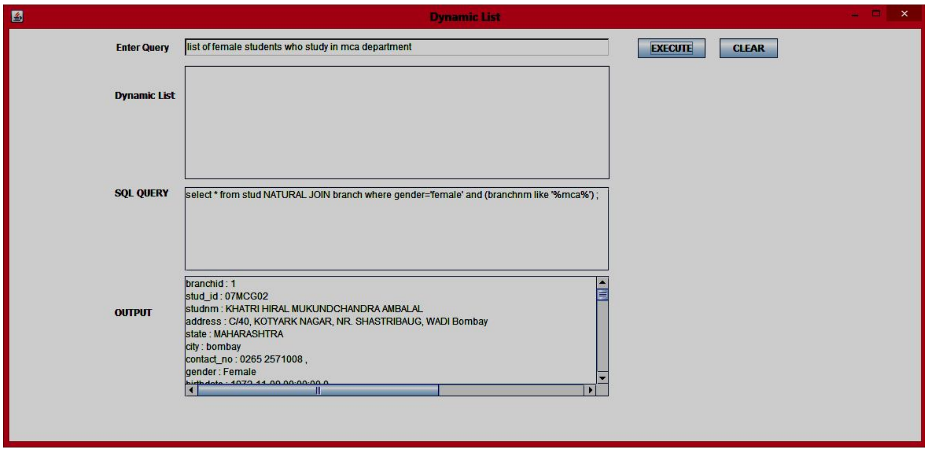
SQL Query: select * from stud where (city like '%vadodara%')

The logical query as stated above represents the meaning of user’s question to find the different pairs like [city, Vadodara] such that a city is a City and Vadodara is a name of a city. The logical query generated by the parsing and semantic interpretation module, expresses the meaning of the user’s question in terms of a logical concept. The logic query does not refer directly to the database object such as tables or columns; it does not specify how to search the database to retrieve the necessary information. In order to retrieve the information requested by the user, the logic query has to be transformed into query expressed in some structured query language supported by the underlying database management system.

We have used WordNet to identify the synonymy, hypernymy, hyponymy, etc. in the development of this system. The tokens which WordNet cannot identify are defined using another two methods (a) Lexicon Semantic representation and (b) Rules for identification of person name, location and date. The N-ELIDB GUI screen shot is given in Figure 2.

The N-ELIDB is capable of handling different types of queries like those with/without conditions, grouping, joins, etc. The different paraphrases of queries are - *list of students whose city is Anand, list names of students who live in Anand city, list address of students who stay at Baroda city, list names of students who study in semester 4, names of students who are born in the month of May, etc.*

The very fact that it can handle different types of queries leads to the next level of using this for creating classifiers for Text Categorization. It is a subtask of Information Retrieval that allows users to browse more easily the set of texts of their own interests, by navigating in category hierarchies [9]. The combination of Text Categorization with NLP gives a new vision for research.



list of female students who study in mca department

select * from stud NATURAL JOIN branch where gender='female' and (branchnm like '%mca%') ;

```
branchid : 1
stud_id : 07MCG02
studnm : KHATRI HIRAL MUKUNDCHANDRA AMBALAL
address : C/40, KOTYARK NAGAR, NR. SHASTRIBAUG, WADI Bombay
state : MAHARASHTRA
city : bombay
contact_no : 0265 2571008 ,
gender : Female
```

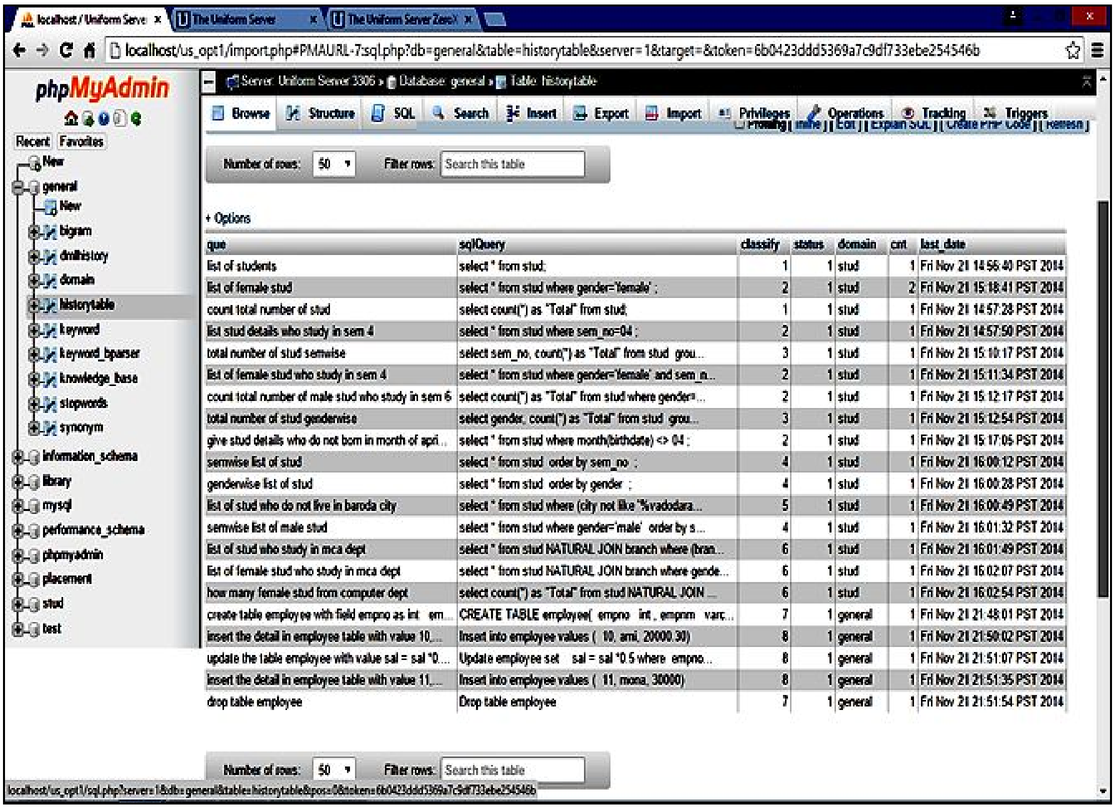
Figure 2: The N-ELIDB Screen Shot

IV. RESULT OF TEXT CATEGORIZATION WITH NLP

Normally statistical methods or the Machine Learning models are popular for Text Categorization. They use the bag-of-words representation to train the target classification function. The Vector Space Model also is very popular in Text Mining. Only the single words with their tf-idf scores and distance measures like the Cosine Distance are generally used. Typical natural language structures like morphology, syntax and semantic are completely neglected in the developing of the classification function [8]. In turn, the semantic information generated by the Text Categorization models is not used yet for the most important natural language applications. Information Extraction, Question/Answering and Text Summarization should take advantage from category information as it helps to select the domain knowledge that language applications usually use in their processing.

We have tried to generate classifiers based on the kind of the query that is asked by a user which is converted to SQL by N-ELIDB. The classifiers have been predefined as per the nature of the query and every new question asked by a user gets classified according to the class it is closest to. It is a heuristic approach which allows the classifiers to grow as per every new query asked and equivalent SQL generated.

A screen shot of the classification is given in Figure 3. We have tried to implement this on paragraphs and small documents with the aim to improve Text Categorization keeping in mind the NLP context. The combination of Text Categorization with NLP can be used in the applications related to business intelligence, text analytics, MIS, etc. The amalgamation of these two concepts can result in better, enhanced and superior quality of outputs. We are continuing this research for documents now.


localhost / Uniform Serve
The Uniform Server
The Uniform Server ZeroX
localhost/us_opt1/import.php#PMAURL-7:sql.php?db=general&table=historytable&server=1&target=&token=6b0423ddd5369a7c9df733ebe254546b
phpMyAdmin
Recent
Favorites
New
general
New
bigram
dmlhistory
domain
historytable
keyword
keyword_bparser
knowledge_base
stopwords
synonym
information_schema
library
mysql
performance_schema
phpmyadmin
placement
stud
test
Server: Uniform Server 3306 » Database: general » Table: historytable
Browse
Structure
SQL
Search
Insert
Export
Import
Privileges
Operations
Tracking
Triggers
Profiling [Inline] [Edit] [Explain SQL] [Create PHP Code] [Refresh]
Number of rows: 50
Filter rows: Search this table
+ Options
localhost/us_opt1/sql.php?server=1&db=general&table=historytable&pos=0&token=6b0423ddd5369a7c9df733ebe254546b

que	sqlQuery	classify	status	domain	cnt	last_date
list of students	select * from stud;	1	1	stud	1	Fri Nov 21 14:56:40 PST 2014
list of female stud	select * from stud where gender='female' ;	2	1	stud	2	Fri Nov 21 15:18:41 PST 2014
count total number of stud	select count(*) as "Total" from stud;	1	1	stud	1	Fri Nov 21 14:57:28 PST 2014
list stud details who study in sem 4	select * from stud where sem_no=04 ;	2	1	stud	1	Fri Nov 21 14:57:50 PST 2014
total number of stud semwise	select sem_no, count(*) as "Total" from stud grou...	3	1	stud	1	Fri Nov 21 15:10:17 PST 2014
list of female stud who study in sem 4	select * from stud where gender='female' and sem_n...	2	1	stud	1	Fri Nov 21 15:11:34 PST 2014
count total number of male stud who study in sem 6	select count(*) as "Total" from stud where gender=...	2	1	stud	1	Fri Nov 21 15:12:17 PST 2014
total number of stud genderwise	select gender, count(*) as "Total" from stud grou...	3	1	stud	1	Fri Nov 21 15:12:54 PST 2014
give stud details who do not born in month of apri...	select * from stud where month(birthdate) <> 04 ;	2	1	stud	1	Fri Nov 21 15:17:06 PST 2014
semwise list of stud	select * from stud order by sem_no ;	4	1	stud	1	Fri Nov 21 16:00:12 PST 2014
genderwise list of stud	select * from stud order by gender ;	4	1	stud	1	Fri Nov 21 16:00:28 PST 2014
list of stud who do not live in baroda city	select * from stud where (city not like '%vadodara...	5	1	stud	1	Fri Nov 21 16:00:49 PST 2014
semwise list of male stud	select * from stud where gender='male' order by s...	4	1	stud	1	Fri Nov 21 16:01:32 PST 2014
list of stud who study in mca dept	select * from stud NATURAL JOIN branch where (bran...	6	1	stud	1	Fri Nov 21 16:01:49 PST 2014
list of female stud who study in mca dept	select * from stud NATURAL JOIN branch where gende...	6	1	stud	1	Fri Nov 21 16:02:07 PST 2014
how many female stud from computer dept	select count(*) as "Total" from stud NATURAL JOIN ...	6	1	stud	1	Fri Nov 21 16:02:54 PST 2014
create table employee with field empno as int em...	CREATE TABLE employee(empno int , empnm varc...	7	1	general	1	Fri Nov 21 21:48:01 PST 2014
insert the detail in employee table with value 10,...	Insert into employee values (10, ami, 20000.30)	8	1	general	1	Fri Nov 21 21:50:02 PST 2014
update the table employee with value sal = sal *0....	Update employee set sal = sal *0.5 where empno...	8	1	general	1	Fri Nov 21 21:51:07 PST 2014
insert the detail in employee table with value 11,...	Insert into employee values (11, mona, 30000)	8	1	general	1	Fri Nov 21 21:51:35 PST 2014
drop table employee	Drop table employee	7	1	general	1	Fri Nov 21 21:51:54 PST 2014

Figure 3: The Text Classifier with NLP

V. Conclusion and Future Work

There has been a lot of research work being done in Text Mining as well as NLP. The thin line between the two is still very difficult to differentiate. Text Mining and Text Categorization especially does not involve the meaning of the unstructured text data that is used to generate classifiers. Most of the time it is the bag-of-words approach involving the Vector Space Model. It is really a challenge to include the semantics as well. This project has been an approach towards creating the connection between the two. Creating more complex classifiers based on the semantics and NLP to classify the test data is what we plan to do next.

REFERENCES

1. Ann Taylor," The Penn Treebank: an overview, Chapter 1", university of York, UK, http://www.cis.upenn.edu/treebank, 2000.
2. Daniel Jurafsky and James Martin, "Book on Speech and Language Processing: An Introduction to Natural language Processing, Computational Linguistics and Speech Recognition", Pearson education, second edition,2009.
3. Androutsopoulos, I. Ritchie & Thanisch P, " Database Interface, a handbook on natural language processing", 209-240,2000.
4. Anxerre P and Inder R. MASQUE "Modular Answering System for Queries in English" - User Manual. AI Applications Institute, University of Edinberg, 1986.
5. George A. Miller, "WordNet: A Lexical Database for English", Communications of the ACM Vol. 38, 1995 http://wordnet.princeton.edu.
6. Marie-Catherine de Marne and Christopher D. Manning (2008), Stanford typed dependencies manual, revised for Stanford parser v 1.6.2 in feb 2010.
7. Ms. Amisha H. Shingala, Ms. Anjali Jivani and Dr. Paresh V. Virparia, Research paper on Multi-Liaison Algorithm, published in International Journal of Advanced Computer Science and Applications (IJACSA), volume 2, issue 5, May 2011.
8. Radu Vlas and William N. Robinson, (2011) A Rule-Based Natural Language Technique for Requirements Discovery and Classification in Open-Source Software Development Projects, Proceedings of the 44th Hawaii International Conference on System Sciences – 1530-1605/11 @2011 IEEE.
9. Alessandro Moschitti (2003), Ph.D thesis on Natural Language Processing and Automated Text Categorization: A study on the reciprocal beneficial interactions, University of Rome, May 8, 2003.
10. D.L. Waltz., "An English Language Question Answering System for a Large Relational Database", Communications of the ACM, 21, pp 526–539,July 1978.
11. G. Hendrix, E. Sacrdoti, D. Sagalowicz, and J. Slocum, "Developing a natural language interface to complex data", ACM Transactions on Database Systems, Volume 3, No. 2, USA, pp. 105 – 147, 1978.
12. H.V. Jagadish Yunyao Li, Huahai Yang. 2005. Nalix: "An interactive natural language interface for querying XML". In Proceedings of the SIGMOD 2005.
13. Ruwanpura : "SQ-HAL Natural language to SQL translator" , Monash University May 2002.
14. Vesper Oweia, "Natural Language Query Filtration in the Conceptual Query Language", Proceedings of The Thirtieth Annual Hawwaii International Conference on System Sciences ISBN 0-8186-7862-3/97 © 1997 IEEE.

BIOGRAPHY

Dr. Amisha H. Shingala is an Assistant Professor in the MCA Department, Sardar Vallabhbhai Patel Institute of Technology, Gujarat Technological University, Vasad, Gujarat. She had done her research in NLP and has published many National and International level papers in conferences and journals. Her area of interest is Natural Language Processing (NLP), Database Management System, Data Mining and Artificial Intelligence.

Dr. Paresh V. Virparia is Director and Professor in G. H. Patel Department of Computer Science & Technology, Sardar Patel University, Vallabh Vidyanagar, Gujarat. He has authored many National and International level papers and has also guided many Ph.D. research scholars and M.Phil students. He is member in editorial board/review committee in many international/national journals. His area of interest is Simulation and Modeling, Computer Network, IT Enabled Services, Web Designing, Artificial Intelligence, Optimization Techniques.

Dr. Anjali G. Jivani is Head and Associate Professor in the Computer Science and Engineering Department, The M. S. University of Baroda, Vadodara, Gujarat. She has published many National and International papers in conferences and journals. She is member in editorial board/review committee in many international/national journals. Her area of research and interest is Text Mining, Data Mining and Database Management Systems.