An Arabic Natural Language Interface System for a Database of the Holy Quran

Khaled Nasser ElSayed
Computer Science Department, Umm AlQura University

***Abstract*—In the time being, the need for searching in the words, objects, subjects, and statistics of words and parts of the Holy Quran has grown rapidly concurrently with the grow of number of Moslems and the huge usage of smart mobiles, tablets and lab tops. Because, databases are used almost in all activities of our life, some DBs have been built to store information about words and surah of Quran. The need for accessing Quran DBs became very important and wide uses, which could be done through database applications or using SQL commands, directly from database site or indirectly by a special format through LAN or even through the WEB. Most of peoples are not experienced in SQL language, but they need to build SQL commands for their retrievals. The proposed system will translate their natural Arabic requests such as questions or imperative sentences into SQL commands to retrieve answers from a Quran DB. It will perform parsing and little morphological processes according to a sub set of Arabic context-free grammar rules to work as an interface layer between users and Database.**

Keywords—Natural Language Processing (NLP); Arabic Question Answering System; Morphology; Arabic Grammar; Database; SQL

I. INTRODUCTION

Language obeys regularities and exhibits useful properties at a number of somewhat separable "levels". Suppose that a database user has some requests that he wishes to convey to database. His requests impose linearity on the signal. All you can play with is the properties of a sequence of tokens. A meaning gets encoded as a sequence of tokens, each of which has some set of distinguishable properties, and is then interpreted by figuring out what meaning corresponds to those tokens in that order.

The properties of the tokens and their sequence somehow "elicit" an understanding of the meaning. Language is a set of resources to enable us to share meanings, but isn't best thought of as a means for *encoding* meanings. This is a sort of philosophical issue perhaps, but if this point of view is true, it makes much of the AI approach to NLP somewhat suspect, as it is really based on the "encoded meanings" view of language.

The expression "natural" language refers to the spoken languages, such as English, Arabic, and French as opposed to artificial languages like languages of programming. NLP systems are programs perform some processes on natural language in some way or another .NLP is considered as one of the most important subfields of AI. It draws on techniques of logical and probabilistic knowledge representation and reasoning, as well as on ideas from philosophy and linguistics. It requires an empirical investigation of actual human behavior, so it is complex and interesting [1].

The main function of NLP is to extract information from the natural input sentences with no care of method of inputting sentences to the computer. It could be used in many applications like: User interfaces (just tell the computer what to do in a textual interface), Knowledge-Acquisition (programs that could read books and manuals or the newspaper, with no need to explicitly encode all of the knowledge), Information Retrieval (find articles about a given topic and to determine whether the articles match a given query), and Translation (machines could automatically translate from one language to another) [2].

Because most of persons have no knowledge of database language, they find it difficult to access database. Recently, there is a rising demand for non-expert users to query relational database in a more natural language. Therefore the idea of using natural language instead of SQL triggered the development of new method of processing named: Natural Language Interface to Database (NLIDB) [3]. The advantages of NLIDB over formal query language and form based interfaces are ; No need to know the physical data structure, No need to learn AI, and Easy to use. In the other side, the disadvantages of NLIDB are; Difficult to decide success or failure of a query, Limited dealing with natural language, and Wrong assumption by users [4].

Mobile applications of the Arabic language are going to grow in the time being and near future. There is an increasing of the need of enriching the Arabic digital content. Almost, there is no study have had its focus on identifying the challenging aspects of developing mobile applications in mobile applications in Arabic. Many studies emphasized that there is a need of considering the identified challenges by interaction designers, developers, and other stakeholders in the early stages of the software life cycle [5]. Arabic Language understanding is an important field of AI. This field can be used to build an intelligent system for translating the natural Arabic request to SQL commands.

The proposed system performs parsing and interpreting of the natural Arabic input such as a question or an imperative sentence. It applied morphology and context-free parsing techniques on context-free grammar of Arabic Language. Then, the system produces an SQL command, which could retrieve the suitable answer from the database of Quran statistics. It uses an approach that lets the computer accepts natural language sentences, but extract only the essential

information from that command. Also, it enables users to learn how to build their SQL commands.

II. Related Work

Daoud introduced in [6] a SMS system named CATS, for posting and searching through free Arabic text using a technology of information extraction. This system can handle structured data stored in relational database and unstructured free Arabic SMS text. He used Arabic interaction language between sellers and buyers through SMS in a classified domain.

Al-Johar and McGregor proposed in [7] developing an Arabic natural language interface to database systems in prolog. They used the approach of intermediate meaning representation in building LMRA notation as a representative for this approach for the Arabic language. This notation divides common nouns into two classes: A mammal common noun (more than one possible gender), and a non-mammal common noun (one possible gender). It has logical formulas to represent a number of Arabic words and phrases.

Mohammad, Nasser, and Harb produced in [8] a Knowledge Based Arabic Question Answering System (AQAS) in prolog. Their system has a knowledge base of a radiation diseases domain. It divided the Arabic query into two parts: the required part (the information requested) and the known part (the thing asked about). Its parser converts the input query into internal meaning representation (IMR), and then it is processed to locate and retrieve the answer for the user. Its IMR is looking for certain words in the query to specify the required information about certain thing.

El-Mouadib, Zubi, Almagrous, and El-Feghi introduced in [9] the design and implementation of an English natural language interface to a database system. Its name is Generic Interactive Natural Language Interface to Databases (GINLIDB). It has two types of semantic grammars parser to supports a wide range of natural language statements: The first is a single lexicon semantic grammar which consists of individual words and some of their synonyms that are used in the English language grammar. While the second is a composite lexicon semantic grammar which is a combination of terminal words (terminals that exist only in the lexicon) that form phrases or sentences in a specific structure. It is designed using of UML and developed using Visual Basic.

Kanaan, Hammouri, Al-Shalabi, and Swalha presented in [10] the architecture of a question answering system. Their system depends on data redundancy rather than complicated linguistic analyses of either questions or contender answers. So, it is different from the other similar system, because a wrong answer is often worse than no answer. It receives Arabic natural language questions, and then it attempts to generate short answers. They used an existing tagger to identify proper names and other crucial lexical items and build lexical entries. They provided an analysis of Arabic question forms and attempted to formulate better kinds of more appropriated answers.

Abu Shawar introduced in [11] a method to access Arabic Web Question Answering (QA) corpus using a chatbot. This method was used properly with English and other European languages. With this method, there is no need for sophisticated NLP or logical inference. Any NLP interface to QA system is constrained to reply with the given answers, so there is no need for NLP generation to recreate well-formed answers, or for deep analysis or logical inference to map user input questions onto this logical ontology. There is simple (but large) set of pattern-template matching rules. This paper used the same chatbot to react in terms of Arabic Web QA corpus.

III. System Structure

The system receives simple requests in natural Arabic language questions as inputs from users. It is responsible of generating a final SQL command and executing it to retrieve the available answer from the database of the holy Quran. To perform that, the input sentence passes through multiple processing operations. Figure 1 presents the structure of the system.

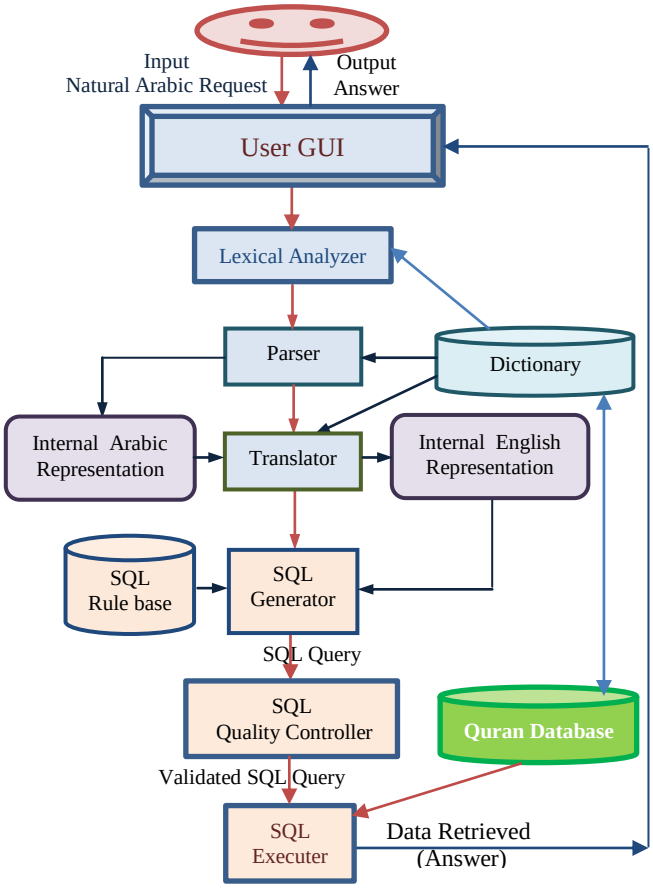

Fig. 1. Structure of the system

User GUI: the system provides an easy Graphical User Interface for easy **interaction** with its users. It is user friendly for none experts of computer or database and even for children. This is done through a menu driven and easy visual forms.

Lexical Analyzer : this stage includes performing four functions; splitting step (scan the user input character by character until recognizing a word to divide natural input into the lowest level of lexemes or words), spelling (checks the spelling of words using the dictionary, if word is not found

correction is done or a new word is added to it), tokenizing (produces a token -the category or meaning – for each word), and abstracting (removes non important words that has no effect in the meaning of request and could be considered as noisy or excessive words).

Parser: The parser performs parsing according to context-free parsing techniques. The syntax of input sentences is represented in context-free grammar rules. It produces the internal Arabic representation of the input sentence according to the suitable grammar rule, using the dictionary. Its syntactic analysis is based on Augmented Transition Network (ATN), which checks if the structure of input tokens is allowed according to grammar rules.

Translator: Early research speculated that computers could be used for Machine Translation "Translation from one language to another". The translator in the proposed system uses the internal Arabic representation of the input sentence produced by the parser and uses the information stored in the dictionary to produce the corresponding internal English representation. It brings the English word corresponding to name of column or table in the applied database.

SQL Generator: The generator uses internal English representation produced by the translator and the SQL rule base to produce the SQL command. It uses the format of SQL rules stored in the SQL rule base as format or a frame and fill slots from the internal English representation.

SQL Quality Controller: The task of the SQL quality controller is to verify the generated SQL query. The query should be verified for valid names of tables, columns and format before applying to the Quran database.

SQL executor: The task of the SQL executor is to retrieve the suitable answer from the database of the holy Quran. Then, the system will consult the answer (the retrieved data) as output to the user.

IV. Database of the Holy Quran

The database used by the system was prepared to hold data about the Quran, to enable executing SQL queries retrieving its statistics data. Mainly, it keeps data about word(s), ayah(s), surah(s), and 30 Jozaa (chapter). Each Jozaa has with two Hezb (section), while each section has four quarters. Figure 2 present the Entity-Relationship diagram for the database of this system.

The entity DICTIONARY has the attributes: WordCode, WordNum, Word_Text, WordNumOfChar, Word_Meaning, Word_Root. While the entity AYAH has the attributes: AyahCode, AyahNum, Ayah_Text, AyaPageNo, and Ayah_meaning. The entity SORAH has the attributes: SurahCode, SurahNum, SurahName, Surah_Area, RevelationOrder, SurahPageNum, Quarter#, Hezb#, Jozaa#. Also there some attributes in the relationship between entities WORD and AYAH, like Word#InAyah.

Most of statistics of the words, ayah, and surah of the Holy Quran for the presented system is transferred from the database of TANZIL resources [12].

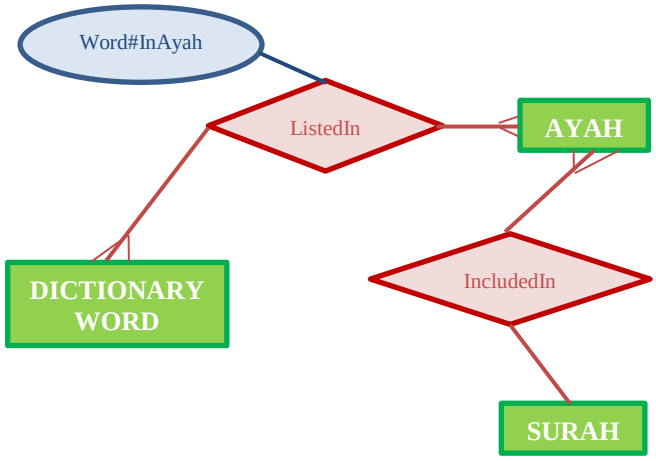

Fig. 2. Entity-Relationship Diagram of Quran Database

V. Parsing and Morphology

A. Parsing Process

Parsing (syntactic analysis) is the core of the proposed system where the input utterance is being checked to ensure that its syntax is correct and structured representations of the possible parses are generated. In parsing, a grammar is used to determine what sentences are legal [2].

Grammar is being applied using a parsing algorithm to produce a structure representation, or parse tree. Parser reads every input sentence, character by character, to decide what is what. It can recognize the underlying structure of a source text and checks that a token is a part of a legal pattern specified by the language grammar. It also gets some attributes of tokens from the dictionary [8].

Context-free grammar is used in parsing the Arabic request inputted to the proposed system, as will as in parsing of most of programming language because it has several advantages. It can deal with the word level and the phrase level. Also, it knows where it is in the sentence at all times. Its main disadvantage is that, it can't handle the numerous valid ways that a language can construct, due to the limitations of size and speed [8].

Context-free grammars are simply grammars consisting entirely of rules with a single symbol on the left-hand side of the BNF rules. The obvious advantage of BNF is that it is simple to define. Many of the grammars used for NLP systems are BNF, as such they have been widely studied and understood and hence highly efficient parsing mechanisms have been developed to apply them to their input.

The system uses context-free parsing technique. It begins by looking at the rules for the sentence, then it looks up the rules for the constituents of the sentence and progresses until a complete sentence structure is build up. If a sentence rule match the input sentence then the parsing process is ended, otherwise, the parser restarts again at the top level with the next rule. It performs syntax analysis recursively until firing certain rule structure or fail.

The parse tree breaks down the sentence into structured parts so that the computer can easily understand and process it. For the parsing algorithm to construct this parse tree, a set of BNF rules, which describes what tree structures are legal, must be available. These rules say that a certain symbol may be expanded in the tree by a sequence of other symbols.

Also, the system used noise disposal parsing for our application, because it is suitable for those application that concern only with a few keywords that a sentence contains, not with all associative words that make up a language. In essence, these types of applications are interested only in the information included in the sentence. This task is done by considering all unknown and un-required words as noise and discarding them. Simply, all sentences must follow a rigid format that resembles natural language.

Its main advantages are the easier implementation of extracting information from sentences, while it is not useful outside restricted situations such as the database queries. This is because it is based on two assumptions: the first is that the sentence follows a strict format, the second is that, only a few keywords or symbols are important. While in normal conversation, most words are important in some way or another.

B. Morphology and Dictionary

Morphology: NLP system doesn't always include morphological analysis. The alternative is to put all possible forms of each word into the dictionary, however storing all possible variants is inefficient and unnecessary. The terms: noun, verb, etc. are morphological but the nominal and verbal which are defined by the distribution of the forms in the sentences are syntactic analysis. Morphology process depends mainly on the dictionary entries and language grammar inflection [13]. This system will discard the prefix and postfix additional characters from a word. It is always one of those listed in Table 1.

Dictionary: a dictionary for NLP system contains the vocabularies known by that system. Its main function is to assist the parser in translating the input sentence into an internal meaning representation (IMR) to be processed.

Any word in the input sentence must be located in the dictionary, taking in consideration the necessary morphological process done by the system. It determines the capabilities of the system. The problem of the format and structure of the dictionary are closely related to the problems of the text storing. If the text is compressed to optimize storage size, the processing time is increased to compress and expand the data.

The format of each dictionary entry depends on the information stored in that entry. The most important data item within each entry is the morpheme itself called the head. Each entry has its appropriate information. The morphological algorithm is responsible of isolating the heads of the dictionary entries from the stream of the input words. Each entry contains the corresponding English meaning of each Arabic word stored in it. Actually, it has the English meaning of imperative verbs or interrogatives and the names of columns and tables of the applied database.

TABLE I. A LIST OF ADDITIONAL PREFIXES AND POSTFIXES

Addition Type		Examples and Meaning	
Pronoun	ضمير	هو, هي, هما, هم, هـ, هن	Its, his, her, their, he, she, they
Preposition	حرف جر	لـ, بـ, كـ,	for, with, as
Preposition and Pronoun	ضمير مع حرف جر	له, لها, لهما, لهم, لهن, به, ..	for him, with him, as him, ….
Definition character	أداة تعريف	ال	the
Preposition and Definition character	حرف جر مع أداة التعريف	لل, بال, كال	for the, with the, as the

VI. SQL QUERIES, INPUTS AND OUTPUTS

First of all, we should take in consideration the already exist SQL queries and their format. Then we should find out how to map between the expected inputs and the generated SQL queries, to finally generate the suitable answer.

A. SQL Queries

This system is run over MySQL database. So, it should generate complete SQL Query as usual in MySQL. Any SQL Query consists of SQL command beside names of attributes (columns) from certain tables and tables themselves and given values of some attributes as conditions if there is.

SQL commands are classified into two categories: Data Definition Language (DDL) commands like: CREATE, ALTER, DROP, DESCRIBE, etc. Data Manipulation Language (DML) commands like: SELECT UPDATE, DELETE, INSERT, etc. [14]. This system will generate SQL queries with the command SELECT only. Dependent on the natural input request, it translates the predicted request to this command. No way to generate another command.

B. Inputs and SQL Queries

The proposed system is expected to process the input sentence in the some of the following two modes [15]:

1) Imperative Mode: This sentence starts with imperative verb like:

استخرج الآيات التي بها كلمة الجنة

Retrieve the Ayah that include the word Paradise.

Table 2 shows list of some examples of imperative verbs beside their meanings and objects. As example, the first five imperative verbs (green color) are allowed from the user. But the last four imperative verbs (red color) are disallowed.

2) Question Mode: This sentence starts with interrogative question like:

ما اسم السورة التي تتضمن أحكام الصوم؟

Retrieve name of Sora include fasting rules?

Table 3 shows some examples of Questions beside their meanings and goals.

TABLE II. LIST OF SOME IMPERATIVE REQUESTS AND SQL QUERIES

Imperative Verb فعل أمر باللغة العربية	English meaning	Corresponding SQL Command
استخرج	Retrieve	SELECT
اعرض	Show	SELECT
أعطني	give me	SELECT
أذكر	List	SELECT
وضّح	illustrate	SELECT
عدّل	change	UPDATE
احذف	Erase	DELETE
صف	describe	DESCRIBE
أدخل أو خزّن	add/store	INSERT

3) Imperative and Question Mode: This sentence starts with imperative verb followed by interrogative question like:

وضح لي أين توجد آيات الحج؟

Tell me, Where is the Ayah of Pilgrim?

Similarly, the user can mix any of the allowed imperative verbs, like shown in table 2 with question like those listed in table 3.

TABLE III. LIST OF SOME QUESTIONS AND CORRESPONDING SQL

Interrogative	English meaning	Corresponding SQL Command
ما	Which	SELECT
ماذا	What	SELECT
من	Who	SELECT
كم	how many/much	SELECT
أين	Where	SELECT

C. The Output Answer

It is predicted the output retrieved from the database of the holy Quran, to be statistics about word (s), Sora (s), and subject(s).

VII. PROCESSING ARABIC REQUEST

All Arabic requests received from users have a common part and parcel. This common part contains all necessary information needed to build an output SQL command, and was given the name REQUEST. The REQUEST part has several forms. Each form represents certain one of the three expected input requests listed above, and were represented by some BNF grammar rules.

The REQUEST part might consists mainly from four components beside some noise data. The first part is the mode itself which included explicitly with the imperative or interrogative or declarative sentence. The second part is the TABLES component, which consists of SQL table names. The third part is the REQUIRED component, which has the names of the items of information actually needed to be retrieved (needed values of attributes of database table). The fourth part is the CONDITION component, which implies the condition to be applied on certain SQL query.

The main task of the SQL generator is to map the elements of the natural query to the elements of the SQL commands of the used databases. There is a general SQL command generated for all queries which is SELECT. So, the SQL generator should finds out the columns (attributes) to be in front of the SELECT command, table(s) to be in front of FROM clause, conditions for WHERE clause and a the method or displaying resulted data in certain order if needed.

VIII. CONCLUSIONS & FUTURE WORK

The presented system satisfies the need for accessing Quran DBs through LAN or WEB for all users, especially with no knowledge of database. It could accept natural Arabic requests such as imperative statements or questions. Then, it generated the suitable SQL command to be verified and executed. Finally it presents the answer from a database of Quran data to the user in an easy manner.

It performed parsing and little morphological processes according to a sub set of Arabic context-free grammar rules to work as an interface layer between users and Database.

In future, the database will be extended to include more tables and attributes. Also, the system will be extended to accept more complex search request and link answer with explanation of meaning of surah and ayah of the holy Quran.

REFERENCES

[1] S. J. Russell and P. Norvig, "Artificial Intelligence – A Modern Approach", 3th edition, Pearson, 2010.

[2] E. V. Provider, "Talking with Computer in Natural Language", Springer-Verlog, NY, 1986.

[3] M. Tyagi, " Natural Language Interface to Databases: A Survey", International Journal of Science and Research (IJSR), Volume 3 Issue 5, pp. 1443-1445, May, 2014. http://www.ijsr.net/archive/v3i5/MDIwMTMyMTU3.pdf

[4] A. Popescu, O. Etzioni, H. Kautz, "Towards a theory of natural language interfaces to databases", In: 8th Intl. Conf. on Intelligent User Interfaces, Miami, FL, pp. 149–157, USA, 2003. http://dl.acm.org/citation.cfm?id=604070

[5] S. N. Zawati and M. A. Muhanna, "Arabic mobile applications: Challenges of interaction design and development", 2014 International Conference (IWCMC) of Wireless Communications and Mobile Computing, Nicosia, pp 134-139, 2-8 Aug 2014. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6906345&url

[6] D. Daoud, "Building an Arabic Application employing information extraction technology", In Proceedings of the Second International Conference on Information Technology (ICIT05), Amman, Jordan, pp. 1-9., 2005. http://icit.zuj.edu.jo/icit05/2005/Information%20Systems/193.pdf

[7] B. Al-Johar and J. McGregor, "A Logical Meaning Representation for Arabic Representation (LMRA)", . In: Proceedings of the 15th National Computer Conference, Riyadh, Saudi Arabia, pp 31-40, 1997. http://www.ccse.kfupm.edu.sa/~sadiq/proceedings/NCC1997/Pap24.doc

[8] F. Mohammad, Kh .Nasser and H. Harb "A Knowledge Based Arabic Question Answering System (AQAS)." In SIGART Bulletin, A Quarterly Publication of the ACM, Special Interest Group on Artificial Intelligence, Vol. 4, No. 4, October, 1993. dl.acm.org/citation.cfm?id=165488

[9] F. A. El-Mouadib, Z. S. Zubi, A. A. Almagrous, I. El-Feghi, "Interactive Natural Language Interface", Journal WSEAS TRANSACTIONS on COMPUTERS, Issue 4, Volume 8, pp 661-680, April 2009. http://dl.acm.org/citation.cfm?id=1558765

[10] G. Kanaan, A. Hammouri, R. Al-Shalabi, M. Swalha, "A New Question Answering System for the Arabic Language", American Journal of Applied Sciences 6: pp. 797-805, 2009. http://thescipub.com/PDF/ajassp.2009.797.805.pdf

[11] B. Abu Shawar, "A Chatbot as a Natural Web Interface to Arabic Web QA", iJET – Volume 6, Issue 1, pp. 37-43, March 2011. http://www.editlib.org/p/44956/article_44956.pdf

[12] Tanzil, "Tanzil Quran Navigator", 1/3/2015 http://tanzil.net/

[13] A. Walker, "Knowledge Systems and Prolog", IBM T.J. Watson Research Center, Addison-Wesley, 1997.

[14] R. Elmasri, and S. Navathe, "Fundamentals of Database Systems", 6th edition, Addison Wesley, 2010.

[15] F. Noama, "Summary of Arabic Grammar", Scientific Center for translation, Cairo, 1988.

AUTHOR PROFILE

The Author is Dr. Eng. Khaled Nasser. ElSayed. He was born in Cairo, Egypt 9 Oct. 1963. He have got his PhD of computers and systems from Faculty of Engineering, Ain Shams University, Cairo, Egypt, 1996.

He has worked as an associate professor of computer science, in Umm-AlQura Uni. in Makkah, Saudi Arabia since 2008. Artificial Intelligence is his major. His interest research is Distant Education, E-Learning, and Agent. Dr. Khaled Nasser ElSayed translated the 4th edition of "Fundamentals of Database Systems", Ramez Elmasei and Shamkant B. Navathe, Addison Wesley, fourth edition, 2004, published by King Saud University, Riyadh, Saudi Arabia, 2009. He is also the author several books in programming in C & C++, Data Structures in C& C++, Computer and E-Society, Database Design and Artificial Intelligence.