

Automating The Normalization Process For Relational Database Model

P.B.Alappanavar¹, Radhika Grover², Srishti Hunjan³, Dhiraj Patil⁴ Yuvraj Girnar⁵

¹(Assistant Professor, Department of IT, STES's Sinhgad Academy of Engineering Pune-48, Maharashtra, India)
^{2,3,4,5}(Department of IT, STES's Sinhgad Academy of Engineering Pune-48, Maharashtra, India)

ABSTRACT

It has been estimated that more than 80 percent of all the computer programming is database related. Studies have shown that the vast majority of the content in the WWW resides in the deep web sources which store their content in backend databases which have been growing by leaps and bounds. Due to its great importance for database applications database schema design has attracted substantial research. Database normalization is a theoretical approach for structuring a database schema and it is very well developed but unfortunately, the theory is not yet understood well by practitioners. It has been difficult to motivate students to learn database normalization because students think the subject to be dry and purely theoretical.

In this paper, a tool called Web Based Relational Database Design and Normalization Tool is proposed, which handles normalization for Relational Databases. The tool is suitable for relational data modeling in systems analysis and design and data management process. This includes creating tables and establishing relationships between those tables according to rules designed both to protect the data and to make the database more flexible by eliminating redundancy and inconsistent dependency. It aims to provide relations normalized up to 3NF and provide an edge over the other tools already present to handle database normalization.

Keywords: Database, Database Normalization, Functional Dependency, Redundancy, Relational Data Model.

1. INTRODUCTION

A database can be described as a computer based record keeping system which provides users a mechanism to store information in an organized manner. The data is stored in a manner that is independent of the programs which use it.

A common and controlled approach is used in adding new data and in modifying and retrieving existing data within the database. A database is useful in automating as much work as possible to enhance

manual processes by minimizing or eliminating the amount of paper shuffling. Hence, a database is often conceived as the repository of information needed for running certain functions in a corporation or organization. Database users exist for just about any organization that you can imagine [1]. Think of individuals such as bankers, lawyers, accountants, customer service representatives, and data entry clerks [1].

Many different types of databases exist, some simple, others very complex. There are different database models but Relational Database model is the most popular database model used today. It has been around for many years and will be around for many years to come because of its capability to manage large amounts of data, its performance, reliability and integrity.

2. STRIVING FOR A GOOD DATABASE DESIGN

A good database design is mandatory for the long run success of any database being used by any organization. If much care is not taken, the quality of end product will suffer [1]. Therefore, it demands for a thought process to be involved in converting an organization's data storage needs into a relational database.

Designing a database is a complex task and the normalization theory is a useful aid in the design process.

3. DATABASE NORMALIZATION

Normalization is the process of organizing data in a database. It is concerned with the transformation of the conceptual schema (logical data structures) into a computer representable form.

This includes creating tables and establishing relationships between those tables according to rules designed both to protect the data and to make the database more flexible by eliminating redundancy and inconsistent dependency [2].

Normalization theory is built around the concept of normal forms. 1NF is critical in creating relations and all others are optional but in order to avoid update

anomalies at least three normal forms (3NF) is recommended [3].

Thus, normalization helps to attain the following advantages-

1. It reduces data redundancies.
2. It helps eliminate data anomalies.
3. It produces controlled redundancies to link tables.

4. NEED FOR AUTOMATING THE NORMALIZATION PROCESS

As the time passes, almost all the organizations experience the need to expand their databases by adding new attributes and new relations. In such situations, the performance of a database is entirely dependent upon its design. If the database is in a normalized form, the data can be restructured and the database can grow without forcing the rewriting of application programs, which is of utmost importance because of the excessive and growing costs of maintaining an organization's application programs and its data from the disrupting effects of database growth.

Normalization is carried out manually in many of the organizations, which demand for skilled personnel with expertise in Normalization. As the database grows, it becomes difficult to manually handle the normalization process.

Thus in this paper, a tool is proposed which aims to automate the most complex phase of the database design process-NORMALIZATION. It will help to achieve the trademarks of a good database design and eliminate the drawbacks of manual normalization process- More time which equates to higher costs, Greater possibility for errors, less structured design and development process and less consistent communication [1].

5. PREVIOUS WORKS

The literature survey done for the proposed system has immensely helped us to understand the purpose and functionality of the systems presently existing to perform database normalization.

Some of the systems presently existing to cater database normalization are as-

- 1) NORMIT [4].
- 2) A Web Based Relational Database Design Tool to Perform Normalization [5].
- 3) A Web-Based Tool to Enhance Teaching/Learning Database Normalization [6].
- 4) A Web-Based Environment for Learning Normalization of Relational Database Schemata [7].

- 5) A Computer Aided Learning Tool: Case of Normalization of Relational Schemata [8].
- 6) Web Based E-Learning System for Data Normalization [3].

5.1 NORMIT

NORMIT is a Web-enabled data normalization tutor which considers self-explanation to be an effective learning strategy. It provides students with a problem-solving environment where they are asked to explain their actions while solving problems.

The student needs to log on to NORMIT first, and the first-time user gets a brief description of the system and data normalization in general [4]. After logging in, the student needs to select the problem to work on. NORMIT lists all the pre-defined problems, so that the student may select one that looks interesting [4]. In addition, the student may enter his/her own problem to work on [4].

NORMIT requires the student to complete the following steps while solving a problem-

1. Determine candidate keys
2. Determine the closure of a set of attributes
3. Determine prime attributes
4. Simplify functional dependencies
5. Determine the normal form the table is in.
6. If necessary, decompose the table so that all the final tables are in Boyce-Codd normal form.

5.1.1 Advantages

1. NORMIT requires explanations in cases of erroneous solutions and for actions that are performed for the first time [4]. The student is asked to specify the reason for the action, and, if the reason is incorrect, to define the domain concept that is related to the current task [4]. If the student is not able to identify the correct definition from a menu, the system provides the definition of the concept [4].
2. Therefore, NORMIT guides the students through the entire process of normalization, making it easier for students to learn.

5.1.2 Disadvantages

1. Due to small number of participants, the tool did not compare the performances of the students who did not self-explained to the performance of their peers who self-explained and this made it difficult to use NORMIT reliably in all the scenarios [4].
2. It focuses more on providing information to the students on normalization which can be found on the web and textbooks as well.

5.2 A Web Based Relational Database Design Tool to Perform Normalization

It is a web based application which has been developed using JDK 1.6.0 and Apache Tomcat server 5.0 and uses MS- Access 2007 database to perform the normalization process.

The tool takes as the input:

1. Name of the relation say R
2. Number of attributes in R
3. Set of functional dependencies of R.

A provision is made to find the set of all the possible candidate keys of R, Primary key and Super key. The tool computes Prime and Non-Prime attributes of R based on the CKs generated [5].

5.2.1 Advantages

1. It generates all possible solution sets of Normalized relations because there is a chance that more than one solution may exist [5].
2. The step by step procedure to evaluate the candidate key can also be seen with the help of this Tool which makes the concept of finding keys very clear and pleasing [5].
3. A provision is made to check if the decomposition is lossless or dependency preserving [5].

5.2.2 Disadvantages

This tool uses MS-Access 2007 which is not only slow as compared to other relational databases but also primarily providing database solutions to small businesses which might not cater to the needs of any leading organization dealing with excessive amount of data for varied purposes.

5.3 A Web-Based Tool to Enhance Teaching/Learning Database Normalization

This tool has been evaluated in surveys on the basis of its effectiveness in teaching relational data model [6]. It has been developed using Java applet based on their modified normalization technique.

The following features are present in it:

1. Main Window- Used to accept one functional dependency at a time and display the same. Buttons are provided to view either the normalized result or step-by-step working of it.
2. Result Window- User can see the final result of normalization
3. Step-by-step Window- Each step of normalization is shown from initial data to 2NF and then in 3NF.

5.3.1 Advantages

1. The tool has been developed after a rigorous thought process and considering the actual needs of students.
2. It shows the step-by-step working involved in normalization.
3. The interface is simple and easy to understand.

5.3.2 Disadvantages

1. The system can work with only 10 functional dependencies.
2. There is no provision to save the output or print it.
3. The interface is far too unattractive to be used because being a web application, it has to look more appealing.

5.4 A Web-Based Environment for Learning Normalization of Relational Database Schemata

The implementation of the web-based learning environment is called as LDBN (Learn DataBase Normalization). The client side of LDBN is written in JavaScript following the AJAX techniques [7]. It is assignment driven such that, students have to first choose an assignment from a list with assignments, submitted by other users (lecturers) [7]. An assignment can also be created but only by a registered user.

Once the assignment has been loaded, the students need to go through the following steps:

1. Determine a minimal cover of the given FDs, also known as a canonical cover.
2. Decompose the relational schema which is in URF into 2NF, 3NF and BCNF.
3. Determine a primary key for each new relation/table.

After that the system analyzes the solution by performing the following checks:

1. Correctness of the minimal cover of the given FDs.
2. Correctness of the FDs associated with each relation R.
3. Losses-join properly for every schema in the decomposition.
4. Dependency preservation for each decomposition.
5. Correctness of the key of each relation. Correctness of the decomposition, i.e., if the decomposition is really in 2NF, 3NF and BCNF.

A dialog with the result is shown to the user [7]. In case of an error the system offers feedback in form of small textual hints, indicating where the error might

be. Different normalization algorithms which are divided into Decomposition algorithms and Testing algorithms have been used to provide the desired functionalities to the user [7].

5.4.1 Advantages

1. User friendly, fast and robust user interface.
2. Registered users can even leave comments on assignments. This provides an easy way for sharing ideas, reducing workload and efficient communication.
3. Performance enhancements in LDBN are as-
 - o Implementation of advanced and fast algorithms such as - ReductionByResolution, SLFD-Closure and Equivalence.
 - o Use of efficient data structures.
 - o Caching some algorithms' output for further use.
 - o Decentralizing the system architecture by moving most of the program logic to the client.
 - o Use of advanced tools for code optimizations such as the GWT's Java-to-JavaScript compiler.

5.4.2 Disadvantages

1. New data structures and algorithms will be needed to incorporate higher normal forms like 4NF.
2. The result is depicted only in a dialog box with the help of various checks performed. A more detailed result i.e. the normalized relations are more of importance and is expected by any user.

5.5 A Computer Aided Learning Tool: Case of Normalization of Relational Schemata

Bottom-up normalization approach has been used in this tool with the help of an enhanced version of the Bernestin algorithm. It has been developed using core java [8]. Data can be normalized to 1NF, 2NF, 3NF and BCNF (Boyce Codd Normal Form).

5.5.1 Advantages

1. A lot of focus is given to the GUI to make it more user-friendly.
2. The system will operate on all common platforms namely Windows, UNIX, Macintosh and Solaris.
3. The system will be able to adjust to the size of the users' Visual Display Unit.

5.5.2 Disadvantages

1. Double buffering in java for example caused flickering when unloading.
2. Swing components appear on top of the other which renders the GUI unusable.

5.6 Web Based E-Learning System for Data Normalization

Rational Unified Process modeling technique has been used to develop this tool where the web pages have been made using ASP.NET, VB.NET and AJAX.

The main features are:

1. User has to register to be able to use the tool.
2. Static pages describing functional dependencies, update Anomalies, Normalization terminology and rules (i.e. steps of Normalization and simple example of each Normal form (1NF, 2NF, 3NF)), is displayed.
3. The highlight of the tool is the "Case Study" page, which is an exercise for the learner, where he is presented with a document and he has to attempt to normalize it till a particular normal form. Feedback is provided after every step to aid and teach the user effectively.

5.6.1 Advantages

1. It has a very interactive page for the user to learn about the normalization forms.
2. The case study page provides feedback at each step of normalization and thus helps to clarify the user's concepts.
3. The static pages provide the users with a lot of information about normalization forms and techniques.

5.6.2 Disadvantages

1. There is no provision to add new exercises.
2. The user cannot add his own table data or constraints.
3. Normalization is available only till 3NF (i.e. 0NF-3NF)

6. PROPOSED SYSTEM

Web Based Relational Database Design and Normalization Tool aims to provide an interactive environment to give its users a hands-on experience in database normalization process. The environment is suitable for relational databases like Oracle and MySQL.

The new users need to register with the system and the existing users login into the system in order to access the system features.

Then the user needs to specify basic information associated with the database- Table Name, No of attributes, Attribute name along with its data type and constraints which have been imposed on it and the functional dependencies associated with the table.

Once all the input is accepted from the user, the relations are normalized up to 3NF according to a predefined algorithm. The user is provided with the normalized tables in different formats in a ready to use form.

6.1 Methodology employed for the proposed work

A methodology is the documented collection of policies, processes and procedures and used by a project to practice software engineering [3]. The proposed system uses Iterative Model as its methodology. The basic idea of Iterative model is that the software should be developed in increments, each increment adding some functional capability to the system until the full system is implemented.

The iterative approach is becoming extremely popular, despite some difficulties in using it in this context. There are a few key reasons for its increasing popularity.

- First and foremost, in today’s world clients do not want to invest too much without seeing returns. In the current business scenario, it is preferable to see returns continuously of the investment made [9]. The iterative model permits this—after each iteration some working software is delivered, and the risk to the client is therefore limited [9].
- Second, as businesses are changing rapidly today, they never really know the “complete” requirements for the software, and there is a need to constantly add new capabilities to the software to adapt the business to changing situations. Iterative process allows this [9].
- Third, each iteration provides a working system for feedback, which helps in developing stable requirements for the next iteration [9].

TABLE 1. Proposed System Iterations

Iteration	Tasks to be Performed
1	1. Acceptance of input from the user- Table Name, Attributes with their data type and constraints and functional dependencies. 2. Ensuring that the relations are in 1NF. 3. Providing normalized relations to the user in text file format.
2	1. Converting the relations into 2NF. 2. Providing normalized relations to the user in text file and dump file format.
3	1. Converting the relations into 3NF. 2. Providing normalized relations to the user in text file and dump file format.
4	Providing normalized relations to the user in ERD format using animation in Silverlight.

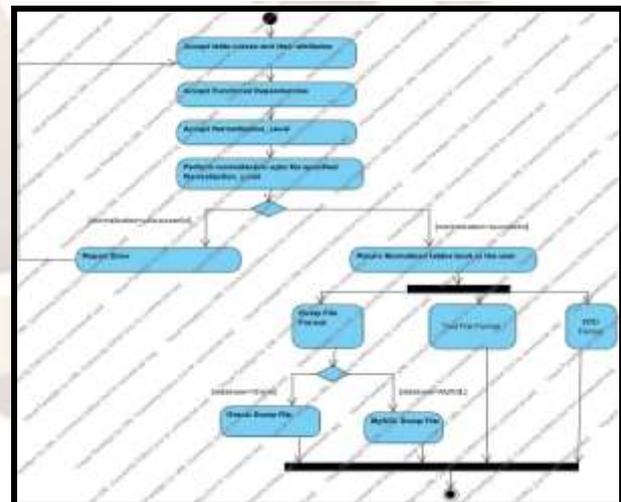


Fig. 1: Activity Diagram depicting the workflow for the proposed system.

6.2 Advantages of the proposed system

- 1) First and foremost, it has a broader scope as it aims to cater to both-leading organizations and students.
- 2) Our tool aims to overcome the disadvantage of existing tools using MS-Access 2007 as it

uses Oracle and MySQL as database which will ensure a more robust approach towards security, performance and disaster recovery.

- 3) It is useful in all the situations as it is not restricted to some predefined problem. It deals with real time scenario for a user defined problem.
- 4) It caters the needs of novice users as a basic input is taken from the users and they will be provided back with the normalized tables itself.

7. CONCLUSION

All the tools existing today for automating normalization primarily provide a learning environment for students by giving a hands-on experience on a predefined or a simple dummy problem statement. A normalized database is not provided back to the users which can help to save their valuable time.

The proposed approach aims to overcome the shortcomings of the existing systems which just aim at developing a learning environment for students as it will not only develop a learning environment in order to give students the ability to easily and efficiently test their knowledge of the different normal forms in practice and can be even used by organizations dealing with large amounts of data for easy, quick and efficient normalization of their tables. It will provide users access to the normalized database.

Our tool uses Oracle and MySQL as database which will ensure a more robust approach towards security, performance and disaster recovery as compared to other tools using MS-Access 2007.

8. FUTURE SCOPE

The tool can be further extended to include features like-

1. Normalize the tables till 5NF.

Thus the tool can be used to provide following

Normal Forms-

- 1) First Normal Form (1 NF)
- 2) Second Normal Form (2 NF)
- 3) Third Normal Form (3 NF)
- 4) Boyce Codd Normal Form(BCNF)
- 5) Fourth Normal Form (4 NF)
- 6) Fifth Normal Form (5 NF)

2. The application can be hosted on Cloud and be used as Software as a Service (SaaS) since SaaS is an "on-demand software", which is a software delivery model in which software and associated data are centrally hosted on the cloud.

REFERENCES

- [1] *Database Design* Ryan K. Stephens and Ronald R. Plew.
- [2] <http://support.microsoft.com/kb/283878>
- [3] Web Based E-Learning System for Data Normalization by Sara Javanbakht Yousefi (Dublin Institute of Technology).
- [4] Supporting Self-Explanation in a Data Normalization Tutor by Antonija MITROVIC- October 2002.
- [5] A Web Based Relational Database Design Tool to Perform Normalization- Radhakrishna Vangipuram (Associate Professor of CSE), Raju velpula (Department of CSE), V.Sravya (Department of CSE) - *International Journal of Wisdom Based Computing, Vol. 1(3), December 2011*
- [6] A web-based tool to enhance teaching/learning database normalization - Hsiang-Jui Kung (Georgia Southern University), Hui-Lien Tung (Troy University).
- [7] A Web-Based Environment for Learning Normalization of Relational Database Schemata by Nikolay Georgiev (Department of Computing Science, Umea University)- *September 2008 Master's Thesis in Computing Science, 30 ECTS credits.*
- [8] A Computer Aided Learning Tool: Case of Normalization of Relational Schemata- King'oo Samuel Mwanzia (BSc Eng, UoN).
- [9] A Concise Introduction to Software Engineering (Undergraduate Topics in Computer Science)- Pankaj Jalote ISSN: 1863-7310 ISBN: 978-1-8400-302-6