A GENETIC PROGRAMMING APPROACH TO RECORD DEDUPLICATION

Software Requirement Specification

Version 1.0

Prepared by

Group No: 13
S6 CS B, B.Tech

11018819 Jils Joseph
11018829 Joseph Vithayathil
11018838 Marbin B
11018876 Suraj Kamal

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
VISWAJYOTHI COLLEGE OF ENGINEERING & TECHNOLOGY, VAZHAKULAM- MUQATTUPUZHA

MAHATMA GANDHI UNIVERSITY, KOTTAYAM
JANUARY 2014
Table of Contents

TABLE OF CONTENTS ....................................................................................................................... iii

1. INTRODUCTION .......................................................................................................................... 1
   1.1 PURPOSE ...................................................................................................................................... 1
   1.2 DOCUMENT CONVENTIONS ........................................................................................................ 1
   1.3 INTENDED AUDIENCE AND READING SUGGESTIONS ............................................................. 1
   1.4 PRODUCT SCOPE ......................................................................................................................... 2
   1.5 REFERENCES ............................................................................................................................. 2

2. OVERALL DESCRIPTION ............................................................................................................... 2
   2.1 PRODUCT PERSPECTIVE ............................................................................................................. 2
   2.2 PRODUCT FUNCTIONS ............................................................................................................... 2
   2.3 USER CLASSES AND CHARACTERISTICS ................................................................................. 5
   2.4 OPERATING ENVIRONMENT ....................................................................................................... 5
   2.5 DESIGN AND IMPLEMENTATION CONSTRAINTS ..................................................................... 5
   2.6 USER DOCUMENTATION ........................................................................................................... 6

3. EXTERNAL INTERFACE REQUIREMENTS ................................................................................... 6
   3.1 USER INTERFACES ...................................................................................................................... 6
   3.2 HARDWARE INTERFACES ........................................................................................................... 7
   3.3 SOFTWARE INTERFACES ............................................................................................................ 7
   3.4 COMMUNICATIONS INTERFACES ............................................................................................. 7

4. SYSTEM FEATURES ...................................................................................................................... 7
   4.1 GATHER EVIDENCE .................................................................................................................... 8
   4.2 ASSIGN FITNESS VALUE ............................................................................................................ 9
   4.3 PERFORM GENETIC OPERATION ............................................................................................... 10

5. OTHER NONFUNCTIONAL REQUIREMENTS ............................................................................... 11
   5.1 PERFORMANCE REQUIREMENTS ............................................................................................... 11
   5.2 SAFETY REQUIREMENTS ........................................................................................................... 11
   5.3 SECURITY REQUIREMENTS ....................................................................................................... 12
   5.4 SOFTWARE QUALITY ATTRIBUTES ......................................................................................... 12
   5.5 BUSINESS RULES ..................................................................................................................... 12

6. OTHER REQUIREMENTS .............................................................................................................. 13
<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>Reason For Changes</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jils Joseph</td>
<td>21/02/14</td>
<td>Initial Draft</td>
<td>1.0</td>
</tr>
<tr>
<td>Joseph Vithayathil</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marbin B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suraj Kamal</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Introduction

1.1 Purpose

This software requirement specification (SRS) document describes the functional and nonfunctional requirements of A Genetic Programming Approach to Record De Duplication version 1.0. The working and objectives is briefly summarized followed by detailed description of the system’s scope, vision, use case, features and other related requirement issues. In the project’s later phases, such as system design, database design, implementation and testing, this document should be referred as functional model of the system for release 1.0.

1.2 Document Conventions

All system development activities should follow the final version of this document. Any discrepancy that found during in later phases should be modified subject to SRS. However, this document may be subjected to change depending on the decision of the group members.

The typographical conventions used in writing this SRS are:

- SRS main headings: Font=Times New Roman, Bold, Size=18, 14.
- SRS sub headings: Font=Times New Roman, Bold, and Size=11.
- Header & Footer – Font Size: 10, Bold & Italics, Times New Roman. The document contains header on all pages. The header is the name of the project on top left end and page number on the top right end of the page.
- Bullets are used to denote main points in the section.
1.3 Intended Audience and Reading Suggestions

The document is intended for different types of readers such as developers, project managers, marketing staff, users, testers, and documentation writers. The rest of this SRS contains an overall description, external interface requirements, system features and other non-functional requirements. Developers and testers can go through the details mentioned from topic 2 to 5. Tester can rely on the document section 4, where each system feature is listed. Database designers will be interested on sections 2.5 and 3.

1.4 Product Scope

The project “A Genetic Programming Approach to Record De Duplication (GPARD)” is mainly focused to avoid the problems due to the duplication of the values in the database. It can avoid the loose of clarity performance and reduce the cost for the memory.

1.5 References

Hospital management software is one of the widely used software and hence its documentation is available in the below mentioned sites and publications.

[1] [IEEE Standard 181-1998]: The standard followed by the current SRS.

[2] Software Engineering: A Practitioner’s Approach, Roger S. Pressman- Is a complete reference for the entire development of a software from the requirement collection to the final release of the software.

2. Overall Description

2.1 Product Perspective

The Genetic Programming Approach to Record De Duplication is a self-contained system that manages data duplication in a data set. It automatically suggests DE duplication functions based on evidence present in the data repositories. Combine the best evidence available in order to identify whether two or more distinct record entries are replicas or not. Its main policy is to avoid the problems due to the duplication of the values in the database.
2.2 Product Functions

In this project collection of data is from different data sets. Results will be updated in the database. Duplicate values are taken from the data set and are transferred to a relation table. The pure data set without the duplicate values are stored back to the data set replacing the original.

- Browse data set
- display the duplicate data
- displaying the data set without the duplicate value

LEVEL 0-DATA FLOW DIAGRAM

(level 0 representation of data flow diagram)
2.3 User Classes and Characteristics

1. Administrator

Admin is responsible for the maintenance of the software and he will see for the security measures for the system. He should be given the authority to add, delete.

2. Data entry operator

Data entry operator is one who enters the data. He is responsible for the updating the databases. The Admin is given higher privilege for using the software.

2.4 Operating Environment

2.4.1 Hardware Requirements:

- GPARD will run successfully in Pentium 1v or above.
- GPARD need a 64 MB or higher RAM capacity to work.
- GPARD require all basic input output devices attached
- GPARD compulsorily need any hardware to provide internet connection.

2.4.2 Software Requirements:

- To support GPARD, one of the Operating system environments like Windows XP, Windows7.
- The back end of GPARD is MySQL.
- To access GPARD online, one require an upgraded web browser.

2.5 Design and Implementation Constraints

2.5.1 Output Design: - Output design involves:

- Duplicate values are decided to display instead of print option.
- Data set without the duplicate values is to be displayed.
2.5.2 Input Design: - This involves:

- Input to the front end of the system is usually designed to be the text data browsed from system.
- Message boxes are used to notify the users while wrong positions are given.

2.5.3 Control Design: - Controls provide ways to:

- Input validations are done.
- Determine whether any necessary data have been omitted.

2.6 User Documentation

GPARD system implementation should focus on user friendliness. Thus providing a user documentation is given the same importance as that of the release of GPARD. The user manual should be provided along with the system, to give guidance to the users. Additionally normal users can go through the ‘About Us’ link provided in the developed system site.

- **User Manual:** The document that helps the users to work the software.
- ‘**About Us**’: Information regarding the system is provided to the users along with the website itself, using a link ‘About Us’

2.7 Assumptions and Dependencies

GPARD being a system focused mainly on the data set in a system, there are many dependencies and assumptions associated with them. They are

- must be a valid data set.
- contains at least two attributes.
- Details in the data set is assumed to be accurate and precise

3. External Interface Requirements

3.1 User Interfaces

The interface between user and the system include many provisions from where they can access the whole system. It contains the option list to move one form to another as well as searches form that is as follows:

1. Home page
2. Data set selection page

3. Page showing the de duplicated values

3.2 Hardware Interfaces

The entire software requires a completely equipped computer system including monitor, keyboard, and other input output devices. The data’s are manually entered to the system and the output are produced automatically.

3.3 Software Interfaces

The system can use Microsoft, Linux as the operating system platform. System also makes use of certain GUI tools.

3.4 Communications Interfaces

The system makes use of internet services and hence any web browser and a proper networking should be enabled within the entire premise.

4. System Features

The genetic program to record DE duplication uses the user inter interfaces to specify the database and shows the fitness values calculated and assigned to the evidences. This also shows the function generated for the duplication and removes the duplicate values. This project has GUI based software that will help in storing and retrieving the information through various user-friendly menu-driven modules.
USE CASE DIAGRAM

4.1 Gather Evidence

4.1.1 Description and Priority

The user can gather all the evidence required automatically to generate the DE duplication function.

Jero algorithm

\[
d_j = \begin{cases} 
  0 & \text{if } m = 0 \\
  \frac{1}{3} \left( \frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & \text{otherwise}
\end{cases}
\]
4.1.2 Functional Requirements
FREQ-1: System should provide a registration form.
FREQ-2: System should provide necessary validation for input data. FREQ-3: System should respond with appropriate messages.

4.1.3 Non-Functional Requirements
NFREQ-1: Collected evidence is real.

4.2 Assign Fitness value

4.2.1 Description and Priority
It is taken so that to find out the two values are duplicate and whether the duplicate values in the data set are deleted.

Fitness value
To assign fitness value for all the evidence gathered is done by two elements;

Traditional precision (P)
Recall metrics (R)

Where

\[
P = \frac{\text{Number Of Correctly Identified Duplicated Pairs}}{\text{Number Of Identified Duplicated Pairs}}
\]

\[
R = \frac{\text{Number Of Correctly Identified Duplicated Pairs}}{\text{Number Of True Duplicated Pairs}}
\]

Where:
- \( m \) is the number of matching characters.
- \( t \) is half the number of transpositions.
Fitness value is set by F1 metrics
\[ F1 = \frac{2*P*Q}{P+Q} \]

4.2.2 Functional Requirements

FREQ-1: System should provide a form to search the patient.
FREQ-2: The system should provide means to request the admin.

4.2.3 Non-Functional Requirements

NFREQ-1: Data’s collected from the data set should be accurate.
NFREQ-2: Data entry should be done carefully.

4.3 Perform Genetic Operation

4.3.1 Description and Priority

The main part of the project in which all the duplicate values are found out using the methods as per the genetic programming methodology.

4.3.2 UC3: Genetic programming approaches

1. Reproduction:

Reproduction is the operation that copies individuals without modifying them. Usually, this operator is used to implement an elitist strategy [8], that is adopted to keep the genetic code of the fittest individuals across the changes in the generations. If a good individual is found in earlier generations, it will not be lost during the evolutionary process.
2. Cross Over:
The crossover operation allows genetic content exchange between two parents, in a process that can generate two or more children.

3. Mutation:
The mutation operation has the role of keeping a minimum diversity level of individuals in the population, thus avoiding premature convergence. Every solution tree resulting from the crossover operation has an equal chance of suffering a mutation process. In a GP tree representation, a random node is selected and the corresponding subtree is replaced by a new randomly created subtree.

4.3.3 Functional Requirements
FREQ-1: data should be having same parameters
FREQ-2: System should allow the duplicate data’s to remain until the user said so.

4.3.4 Non-Functional Requirements
NFREQ-1: Data collected from the data set must be verified before updating the database.
NFREQ-2: Data entry should be done carefully.

5. Other Nonfunctional Requirements

5.1 Performance Requirements
The performance of the system lies in the way it is handled. Every user must be given proper guidance regarding how to use the system. The other factor which affects the performance is the absence of any of the suggested requirements.
5.2 Safety Requirements

To ensure the safety of the system, perform regular monitoring of the system so as to trace the proper working of the system. An internal staff has to be trained to ensure the safety of the system. He has to be trained to handle extreme error cases.

5.3 Security Requirements

Any unauthorized user should be prevented from accessing the system. Password authentication can be introduced.

5.4 Software Quality Attributes

1. Planned approach towards working: - The working in the organization will be well planned and organized. The data will be stored properly in data stores, which will help in retrieval of information as well as its storage.

2. Accuracy: - The level of accuracy in the proposed system will be higher. All operation would be done correctly and it ensures that whatever information is coming from the center is accurate.

3. Reliability: - The reliability of the proposed system will be high due to the above stated reasons. The reason for the increased reliability of the system is that now there would be proper storage of information

4. No Redundancy: - The main objective of proposed system is to provide de duplicate values . This would assure economic use of storage space and consistency in the data stored.

5. Immediate storage of information: - In manual system there are many problems to store the largest amount of information.

6. Easy to Operate: - The system should be easy to operate and should be such that it can be developed within a short period of time and fit in the limited budget of the user.
5.5 Business Rules

Capacity of an organization to provide useful services to its users is proportional to the quality of the data handled by its systems. This is having a big scope and space in today’s society. GPARD frees the user from the burden of choosing the replica identification boundary value, thus user input is very limited. Only an administrator or a data enter is needed.

6. Other Requirements

The data set are collected from different resources having different parameters. User must collect the data if the data is not provided. The data he need can be:

- Restaurant data
- Library
- Data repositories

Data collected can be any type regardless of its kind and must be safe to ensure the data is correct, that is without errors.

Appendix A: Glossary

- SRS: Software Requirement Specification
- GPAPD: Genetical Programing Approach for Record DE duplication
- GUI: Graphical User Interface
- FREQ: Functional Requirement
- NFREQ: Non Functional Requirements
- SQL: Structured Query Language
Appendix B: Analysis Models

Hierarchical model

Flow Chart of Record DE-Duplication
LEVEL 1-DATA FLOW DIAGRAM

1.0 Receive and transfer data’s

2.0 Update duplicate values

3.0 Update non duplicate values

4.0 Produce admin status

User

Data table
Consistent data set

D1 Duplicate data set

D2 Non Duplicate data set

Duplicate set
Non duplicate set

Formatted duplicated data
Non formatted duplicate data

Specific duplicate value
Specific non duplicate value

Admin
LEVEL 2-DATA FLOW DIAGRAM