**ITCS 6155 KNOWLEDGE BASE SYSTEMS**

**TERM PROJECT SUMMER 2013**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**FLEXIBLE QUERY ANSWERIING SYSTEM**

**Project Group**

**Raja Narasimha Karthik Tangirala**

**Santhosh Reddy Chintala**

**Srikanth Bachati**

**Vinay Polisetti**

------------------------------------------------------------------------------------------------------------------------------------------

**Introduction**:

Flexible Query answering system tries to suffice the problem of failing query problem using Muslea’s algorithm. In the project we used the numerical data set from UCI archive on breast cancer which has 11 attributes and we chose a query with 4 attributes which does not return any data from the given dataset.

Link for the data set: <http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/>.

**WEKA(Brief Description):**

The Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. The original non-Java version of Weka was a TCL/TK front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains but the more recent fully Java-based version (Weka 3), is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

**Implementation of the Muslea’s algorithm (Project implementation):**

Initially we are reading the query given by the user and then splitting them in to different parts(attributes specified, operator symbol, and the values used i.e. the threshold specified for each attribute in the query) then putting them in to the corresponding arrays.

**Extracting the domain knowledge:**

In this step we look at the constraints given in the query and create duplicate data set for each constraint in which the constraint attribute value is given a binary value depending on whether the constraint is satisfied or not in the original dataset. Then generate rules by using J48 class of the WEKA library and generate a set of decision rules.

**Finding the most useful rule:**

From the decision rules learnt from the step 1 the most useful rules is derived using the nearest neighbor technique. Here we used the distance function for continuous data set, maximum and minimum vales of each attribute to find the refiner statement. In this we are generating the arff files corresponding to each attribute which are specified ion the query along with ‘Yes / NO ’ for each attribute by comparing the rest of the query attributes.

**Relaxing the failing query:**

The implementation creates a relaxed query that contains only constraints on attributes that appear both in the user input query and the duplicate datasets generated in the previous steps, for each of these constraints the relaxed query uses the less constraining value of those in the user input query and the data set generated. If the attributes does not have any values in common(or does not have any intersecting values) then those attributes are ignored while finding the relaxed query.

**Steps to execute the project:**

* Download the zip file.
* Import the project files into Eclipse.
* Make sure to add WEKA .jar files located in the lib folder. (For the current project we have already added the required jar files so need to add them again.)
* Copy the files Dataset.csv, Keywords.txt, Minvalues.txt, Maxvalues.txt which are present in the project resources folder to the path specified in the Class ‘FilePath.java’ (by default it is set to C:\\FQAS Project).
* As we have developed a web application using servlets we need to add servlet – api.jar to this project. This can be done by
* Right click the project
* Select Properties and in that select Java Build Path and select Libraries tab
* Click on the Add External JARs button and select the ‘servlet-api.jar’ file present in the lib folder of the tomcat server. For convenience this jar file is attached along with the project documents.
* Configure Apache tomcat server and add the project to the server and run on server.

**Project input and output**

**Input Query:**

Sample code number > 1056784 ^   
Clump Thickness > 6 ^   
Uniformity of Cell Size > 5 ^   
Uniformity of Cell Shape > 7

**Output:**

* Relaxed Query**:**

Uniformity of Cell Shape > 7 ^ Uniformity of Cell Size > 5

* Rules:

[Uniformity of Cell Size <= 2 AND

Uniformity of Cell Size <= 1 ^ Clump Thickness > 6, Uniformity of Cell Shape <= 2 ^ Clump Thickness > 6, Uniformity of Cell Size <= 3 AND

Uniformity of Cell Shape <= 3 ^ Clump Thickness > 6, Uniformity of Cell Shape <= 5 AND

Uniformity of Cell Shape <= 2 ^ Uniformity of Cell Size > 5, Uniformity of Cell Shape > 6 ^ Uniformity of Cell Size > 5, Uniformity of Cell Shape <= 3 AND

Clump Thickness <= 6 ^ Uniformity of Cell Size > 5, Uniformity of Cell Size <= 6 AND

Uniformity of Cell Size <= 4 ^ Uniformity of Cell Shape > 7, Uniformity of Cell Size <= 6 AND

Sample code number > 412300 AND

Clump Thickness > 2 ^ Uniformity of Cell Shape > 7, Uniformity of Cell Size > 9 AND

Clump Thickness > 4 ^ Uniformity of Cell Shape > 7]

**Files Generated(In the same folder where we have placed the Input files):**

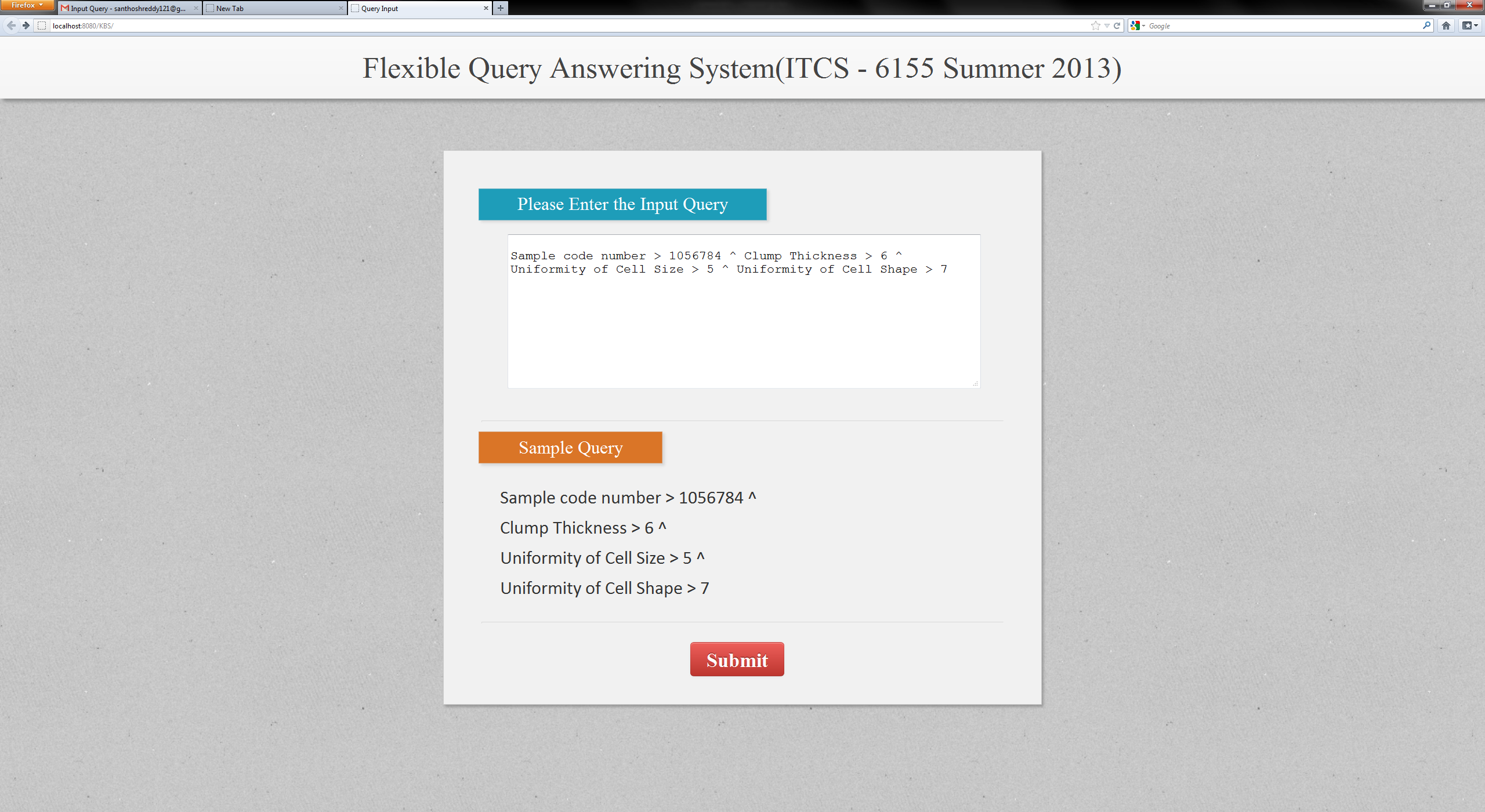
* Uniformity of Cell Size.arff and Uniformity of Cell Size excel file
* Sample code number.arff and Sample code number excel file.
* Clump Thickness.arff and Clump Thickness.arff excel file.
* Uniformity of Cell Shape.arff and Uniformity of Cell Shape

Sample data from *uniformity of Cell Size* excel file.

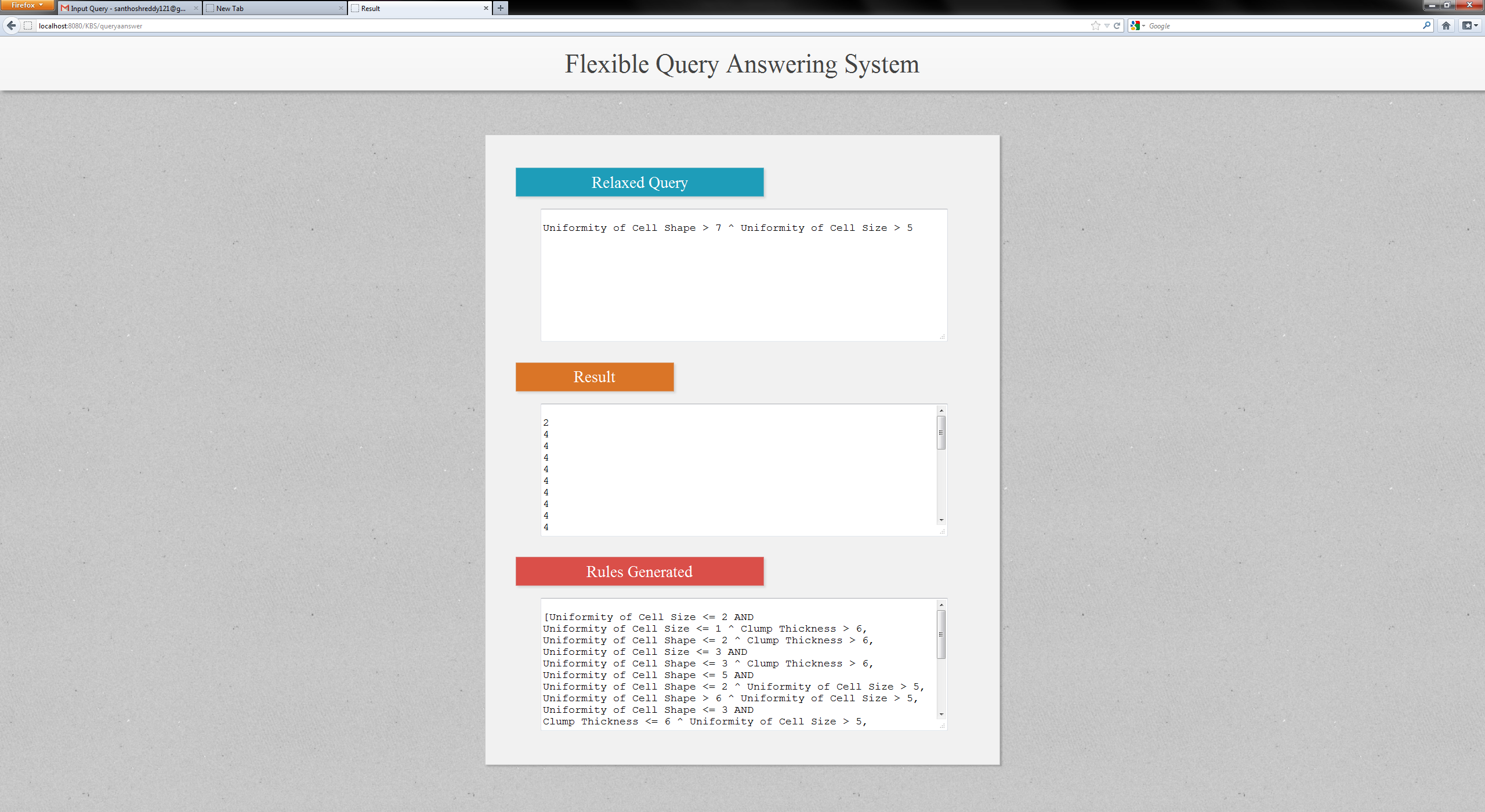
|  |  |  |  |
| --- | --- | --- | --- |
| Sample code number | Clump Thickness | Uniformity of Cell Size | Uniformity of Cell Shape |
| 1000025 | 5 | NO | 1 |
| 1072179 | 10 | YES | 3 |

**Screen shots:**

The UI of the project is shown below, enter the input query in the text box present take a look at sample query before entering your own. Click on the submit button to view the results.



The Output screen is shown below which shows the Relaxed query, the rules generated and the corresponding decision attribute values



**References:**

Machine Learning for Online Query Relaxation by Ion Muslea (SRI International) .

Action Rules Mining, A. Dardzinska

Data set- <http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/>

<http://en.wikipedia.org/wiki/Weka_%28machine_learning%29>