**Proceedings Paper:**
Hamoud, B and Atwell, ES orcid.org/0000-0001-9395-3764 (2016) Quran question and answer corpus for data mining with WEKA. In: 2016 Conference of Basic Sciences and Engineering Studies (SGCAC). 2016 Conference of Basic Sciences and Engineering Studies (SGCAC), 20-23 Feb 2016, Khartoum, Sudan. IEEE , pp. 211-216. ISBN 978-1-5090-1811-6

https://doi.org/10.1109/SGCAC.2016.7458032

# Quran question and answer corpus for data mining with WEKA

Bothaina Hamoud
[1]Preparatory Year
Umm Al-Qura University
Makkah, Saudi Arabia
[2]College of Computer Science and Information Technology
Sudan University of Science and Technology
Khartoum, Sudan
Boothy2007@yahoo.com

Eric Atwell
School of Computing, Faculty of Engineering
University of Leeds
Leeds LS2.9JT, England
e.s.atwell@leeds.ac.uk

***Abstract*—This paper presents the compilation of a holy Quran question and answer dataset corpus, created for data mining with Waikato Environment for Knowledge Analysis (WEKA). Questions and answers from the Quran were collected from multiple data sources, and then a representative sample of the question and answers were selected to be used in our model. Then the data was cleaned to improve data quality to the level required by the WEKA tool, and then converted to a comma separated value (CSV) file format to provide a suitable corpus dataset that can be loaded into WEKA. Then StringToWordVector filter was used to process each string into a bag or vector of word frequencies for further analysis with different data mining techniques. After that we applied a clustering algorithm to the processed attributes, and show the WEKA cluster visualizer.**

***Index Terms*—Data mining, WEKA, dataset, Corpus, Quran**

## I. INTRODUCTION

There are tremendous amounts of data on the internet, in many different formats and various templates, written using different applications. Data collection is an important phase in many research projects; data is collected for many purposes to perform some required tasks such as to aid in research on statistical methods, or to train classifiers for machine learning experiments. Collecting questions and answers in any domain can be an important resource to serve that domain, for example it can be used in a question answering system to answer online user questions. There are question and answer datasets available in many domains, for research and practical use; however there is no existing public resource specifically designed for Quran questions and answers. There are existing sources which do include questions and answers from the Quran, but these are scattered between different webpages and not integrated in one corpus. Each of these corpora has its own format and style. There is a requirement to create a unified dataset corpus for Quran questions and answers. Collecting Quran questions and answers in a unified corpus and converting it to text file format (.CSV) that can be loaded into WEKA to do further analysis is our challenge. This paper aims to create a Quran question and answer corpus dataset; this corpus should be useful in the application where Quran questions and answers have to be searched. Our main contribution can be summarized in the following points:

- Collecting a corpus dataset of questions and answers from the Quran harvested from a range of sources, and merging it manually to form one corpus
- Cleaning the merged corpus and normalizing the range of source formats into a standard CSV file format
- Loading the corpus dataset into WEKA, and applying some preprocessing techniques to demonstrate data mining experiments with WEKA

### *A. Data mining*

Data mining is defined as the process of discovering useful patterns in data. The process should be automatic or semiautomatic [1]. Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, useful, and understandable patterns in data. Data mining is sometimes called knowledge discovery from databases. In knowledge discovery, what is retrieved is not explicit in the database. Rather, it is implicit patterns. Data mining finds these patterns and relationships using data analysis tools and techniques to build models. There are two main types of models in data mining. One is predictive models, which use data with known results to develop a model that can be used to explicitly predict values. Another is descriptive models, which describe patterns in existing data. All the models are abstract representations of reality, and can be guides to understanding business and suggest actions. The two high-level primary goals of data mining, in practice, are prediction and description. The main tasks for data mining are: (1) classification: learning a function that maps (classifies) a data item (record or instance) into one of several predefined classes, (2) estimation: given some input data, coming up with a value for some unknown continuous variable, (3) prediction: same as classification and estimation except that the records are classified according to some future behavior or estimated future value, (4) association rules: determining which attribute

or feature values typically go together in a data record or instance, also called dependency modeling, (5) clustering: segmenting a population of data records or instances into a number of subgroups or clusters, (6) Description and visualization: representing the data using visualization techniques, for human inspection of patterns.

Learning from data is categorized in two types: directed (supervised) and undirected (unsupervised) learning. Classification, estimation and prediction are examples of supervised learning tasks, while association rules, clustering, and description and visualization are examples of unsupervised learning tasks. In unsupervised learning, no variable is singled out as the target; the goal is to establish some relationship among all variables. Unsupervised learning attempts to find patterns without the use of a particular target field.

Data mining steps in the knowledge discovery process are as follows [2]:

- Data cleaning: To remove noise and inconsistent data.
- Data integration: To combine multiple sources of data.
- Data selection: The retrieval of relevant data from the database.
- Data transformation: The consolidation and transformation of data into forms suitable for mining.
- Data mining: Using intelligent methods to extract patterns from data.
- Pattern evaluation: To identify the interesting patterns

*B. Waikato Environment for Knowledge Analysis*

The Waikato Environment for Knowledge Analysis *(WEKA*) is computer software developed at the University of Waikato in New Zealand. *WEKA* is free Java software available under the *GNU* General Public License. It is a collection of machine learning algorithms to solve data mining problems. *WEKA* has become one of the most widely used data mining systems while it offers many powerful features [1]. *WEKA* support many different data mining tasks such as data preprocessing and visualization, attribute selection, classification, prediction, model evaluation, clustering, and association rule mining. It is written in Java and can be run with almost any computing platform. *WEKA* can be used to preprocess without writing any program code, and it comes with a graphical user interface to provide easily used tools for beginner users to identify hidden information from database and file systems in a simple way by using options and visual interfaces. There is a specific default .ARFF data file format that WEKA accepts. The data should be as a single flat file or relation; the data can be imported from a Comma Separated Value (*.CSV*) file, a database, a *URL* etc. where each data point is described by a fixed number of attributes. *WEKA* supports numeric, nominal, date and string attributes types. *WEKA* can be used to learn more about the data by applying a learning method to a dataset and analyze its output, and it is also used to generate predictions on new instances by using learned models, as well as to apply several different learners and compare their performance in order to choose one for prediction. The desired learning method is selected from a menu. A common evaluation module is used to measure the performance of all classifiers.

The most valuable resource that *WEKA* provides is the implementations of a wide range of data filtering tools, machine learning schemes, evaluation methods, and visualization tools. Filters are used to preprocess the data; we can select filters from a menu and then adjust their parameters according to our requirements. *WEKA* also includes implementations of algorithms for learning classifiers, association rules, clustering data for which no class value is specified, and selecting relevant attributes in the data. Also, there are many tools developed by third parties as add-ons. For example, *WEKA* was not designed for 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 not covered by the algorithms included in *WEKA* is sequence modeling [3], but third-party add-ons are being developed.

The *WEKA* Graphical user Interface *(GUI)* Chooser provides a starting point for launching *WEKA's* applications, this GUI Chooser consists of four buttons, to start applications [4]. The main application is the Explorer, which explores data with *WEKA*. The simple command line interface (SimpleCLI) allows direct execution of *WEKA* commands for operating systems that do not provide their own command line interface. KnowledgeFlow supports the same functions as the Explorer but with a drag-and-drop interface, and it provides a framework for incremental experiments in machine learning. The Experimenter is used to carry out experiments and perform statistical tests between several learning schemes.

The Explorer is the most used tool, and is composed of several panels to allow access to the main components of the workbench : (1) the Preprocess panel which is used to import data , and preprocess this data by using filters to transform the data and prepare it according to specific criteria, (2). the Classify panel which allow applying classification and regression algorithms to a dataset, (3) the Cluster panel enables access to the clustering techniques in WEKA, (4) the Associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data, (5) the Select Attributes panel gives algorithms for identifying the most predictive attributes in a dataset, (6) the Visualize panel shows picture representations of data and results, such as a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

The first four buttons at the top of the Preprocess section enable us to load data into *WEKA*, by importing it from any file in the default ARFF format supported by WEKA, or any other format accepted by *WEKA* for which a filter is implemented, such as Excel Comma Separated Value *(.CSV)* file, a SQL database, a URL, etc., for preprocessing.

### C. CSV Comma Separated Value

Comma-Separated Values (.*CSV*) files are a common data exchange format that stores tabular data in plain text format [5], which can be read using any standard text editor. *CSV* is supported by many applications and therefore a large amount of tabular data can be transferred between these applications. Each line of the *CSV* file is a record composes of one or more fields, separated by commas. Records are separated with end of line character, and this is used by the preprocess system. The tab-separated values and space-separated values are commonly used field delimiters in *CSV* files. *CSV* files are given the extension .csv. While there are various specifications and implementations for the CSV format, there is no official standard format and there are a variety of interpretations of *CSV* files. There is variation in the handling of fields which contain long strings, quote and double quote marks, and/or line-breaks; these are commonly found in Text Analytics datasets, where each data item may be a string representing a document, such as a news story, or a chapter or verse from a book. The format that is applied by most implementations is summarized as follows: [6] [7]

- Each record is placed on a separate line, delimited by a line break (*CRLF*), but a record may span to more than one line when fields contain line-breaks (and field that containing line-breaks must be surrounded by double-quotes – see below)
- The first record in the file may be a header record containing fields (columns) names, with the same format, and the same number of fields as the records in the rest of the file
- All records must contain the same number of fields throughout the file
- The last record in the file may or may not end with an ending line break
- Fields are separated with commas.
- The last field in the record must not be followed by a comma.
- Each field may or may not be enclosed in double quotes (some programs, such as Microsoft Excel, do not use double quotes). If fields are not enclosed with double quotes, then double quotes may not appear inside the fields
- A field which contains commas inside it must be enclosed in double-quote characters
- A field that containing line-breaks must be surrounded by double-quotes
- A field which contains double quote characters must be enclosed in double-quotes, and each of the embedded double-quotes must be also enclosed in double quotes

### D. Cross Industry Standard Process for Data Mining

It will be useful to use the Cross Industry Standard Process for Data Mining (*CRISP-DM*) to prepare the data for the analysis tool. The *CRISP*-DM methodology is described in terms of a hierarchical process model, consisting of sets of tasks described at four levels of abstraction [8] (from general to specific): phase, generic task, specialized task and process instance. At the top level, the data mining process is organized into six phases: (1) business understanding, (2) data understanding, (3) data preparation, (4) modeling, (5) evaluation, (6) deployment [9].

Business understanding is to understand the project objectives and requirements from a business point of view, then convert this knowledge into a data mining problem definition, and then design a preliminary plan to achieve the objectives. Data understanding begins with an initial data collection then continues with activities in order to be familiar with the data, to determine data quality problems, to discover first insights into the data or to discover interesting subsets to form hypotheses for hidden information. Data preparation covers all tasks to build the final dataset from the initial raw data, such as table, record and attribute selection as well as transformation and cleaning of data for modeling tools. Data preparation tasks can be performed multiple times, and not in specific order. Modeling is to select and apply various Data Mining algorithms as modeling techniques and to calibrate their parameters to optimal values. Evaluation is to evaluate the results against business objectives. Deployment is to deploy the resulting model to be used by the customer whenever it is required.

## II. DATA PREPARATION

The data was collected and stored in MS-Excel 2010 and formatted according to the required structures; then it was converted to a CSV (Comma Separated Value) format to be loaded into WEKA.

### A. Data collection tools

The process of collecting data can be relatively simple according to the type of tools used to collect the data. Data collection tools are used to collect information that can be used in many aspects such as evaluation of project performance. The collected data can also be reused for analysis purposes after its refinement and cleaning. There are several methods that can be used to collect data. The data collection methods should be good enough to collect useful data in order to have better evaluations for the research. Selecting specific methods depend on the nature of the task, as well as the type of the required data. In this paper we selected three methods to collect Quran questions and answers: (1) a web-based tool, created by a group of scholars interested in the Islamic field, (2) eliciting questions and answers from Islamic experts: a group from Holy Mosque in Mecca who are leaders in the field of Islamic studies, this group gives answers to questions from Muslims who come to the Holy Mosque, and seek their expert advice, (3) some samples of questions and answers gathered from previous research.

### B. Collected questions and answers related to the holy Quran

Data collection is an important part of many research projects. Having enough data to learn from is of great significance for good performance. The issue of training data quantity and quality is key in machine learning research.

Increasing training set size can be more significant than developing better learning algorithms in terms of impact for improving object classification performance [10].

The selection of such Islamic data requires great care to give an accurate answer for a given question, which can be accepted religiously and universally. For instance, the answers should be evidenced as mentioned in Holy Quran as well as in Hadith books. In this paper we propose to focus on Frequently Asked Questions (*FAQs*) using the above tools in order to collect the questions and answers. Collecting questions and answers from authoritative and credible sources is an important issue; and low accuracies or wrong answers are not acceptable in the religious field especially in the domain of the holy Quran. Since there are not enough existing resources specifically designed for Quran questions and answers, we propose to merge different data subsets to comprise the Quran questions and answers dataset, as well as to have different questions from a range of different sources. Four web resources were used as raw data sources for our questions and answers dataset. The first web resource is the questions and answers generated by TurnToIslam.com [11], which is widely acknowledged to be one of the best places to learn about Islam, as it contains a huge library that covers many topics about Islam in many languages. It answers questions, shares Videos, Polls, Events and more. It aims at strengthening and uniting the nations and helping to show the beauty of Islam to the world, as well as building a kind, friendly community, on Islamic values.

The second web resource is the questions and answers generated by Islamic Knowledge/Come towards Islam [12] which is another widely-used web site, which contains monthly archives covering many topics concerned with Islam such as questions and answers about the Quran, understanding Islam, Islamic facts, holy Quran chapters, teachings of the prophet Muhammad, haram (forbidden) food and drinks, Ramadan, women in Islam and many more topics. We examined its archives running from March 2011 till February 2015. This web site also provides the Quran text translated into many languages, as well as an Islamic resource for reading and listening to the Quran online with translation in various languages.

The third web resources is All-Quran [13] web site, which aims to have the holy Quran available to everyone in the world by having an easy way of audio streaming for a variety of Quran reciters and audio translations. It contains a tab for Islamic FAQ, in addition to further information about the holy Quran. It has been a commercial-free website since it was created.

The fourth web resources is The Siasat [14] web site, which also provides questions and answers about the holy Quran. It is written in three languages: English, Urdu, and Hindi.

As mentioned above, beside these four web resources, some questions and their answers were gathered from Islamic experts at the Holy Mosque in Mecca, and from previous research projects.

### *C. Preparing Comma Separated Value file*

To prepare a *CSV* file to be loaded into *WEKA* the following has been done. A representative sample of questions and their answers were selected from the collected questions and answers dataset, to include representative samples covering the above methods and sources that were used to collect them, and the questions types, as well as the length of the answer. There are some questions that need a long explanation for their answers. The selected dataset required cleaning prior to data usage; therefore the data was cleaned by removing control characters, and resolving formatting problems concerned with some characters such as double quotes, single quotes, comma, apostrophe, etc.

Since Excel can produce and use *CSV* files, these questions and their answers are entered in one sheet of an Excel workbook file, so that each question and its corresponding answer is a record (row). Our table consist of two columns with headings "question" and "answer". Then the Excel file was saved as comma delimited values *CSV* file, by selecting the format "*CSV* (comma delimited value)" from "save as types" in the Save As dialog box.

## III. LOADING THE CORPUS INTO WEKA

To load the *CSV* file, from *WEKA* chooser *GUI* we selected Explorer application button, and then the Preprocess panel, which is used to choose and modify the data being acted on. After that we chose the Open File button to display a dialog box allowing browsing for our *CSV* data file. From the dialog box we selected Open button to load our file into *WEKA*. WEKA also enables us to load data from other locations by selecting the desired button. The Open URL button allows asking for a Uniform Resource Locator web-address where the data is stored. The Open DB button is used when we want to read data from a database. The Generate button enable us to generate artificial data from a variety of DataGenerators. Using the Open File button we can read files in a variety of formats: *ARFF*, *CSV,* C4.5, or serialized instances format, these format have the extensions .arff, .csv, .data and .names, .bsi. *WEKA* has converters for some file formats such as Spreadsheet files with extension .csv, C4.5's native file format with extensions .names and .data, etc., This list of formats can be extended by adding custom file converters to the *WEKA* core converters package. The appropriate converter is used based on the file extension. If WEKA cannot load the data, it tries to interpret it as ARFF. If that fails, it pops up the generic object editor box, which is used throughout *Weka* for selecting and configuring an object. In this case the *CSVLoader* for .csv files is selected by default and the "*more"* button gives us more information about it.

After loading the CSV file into WEKA's Explorer, this dataset was processed into vectors of word frequencies using the StringToWordVector filter, which converts a string attribute to a "bag of words", a vector that represents word occurrence frequencies. The StringToWordVector filter produces numeric attributes that represent the frequency of words in the value of each string attribute. The set of words

(the new attribute set) is determined from the full set of values of all the strings in the full dataset. The list of all attributes, statistics and other parameters can be utilized as shown in Fig.1. There are 30 instances and 196 attributes in the "Quran question and answer" sample relation file. The processed data can be further analyzed using different data mining techniques like, clustering, association rule mining, and visualization algorithms. In our model we use 4 clusters. Fig. 2 shows the attributes which are clustered, the number of clusters, and the number of instances each cluster contains. Clustering is used for data in which no class value is specified. In clustering, relevant attributes in the data are selected to decide the cluster. In some algorithms the number of clusters can be specified by setting a parameter in the object editor. For probabilistic clustering methods, *WEKA* measures the log-likelihood of the clusters on the training data: The larger this quantity, the better the model fits the data. Increasing the number of clusters normally increases the likelihood. Fig 3 shows *WEKA* cluster visualizer in which the attributes are clustered into 4 groups. The Visualize panel helps to visualize a dataset itself. It displays a matrix with a two-dimensional scatter plot.

## IV. RESULTS

From Fig. 2, it is shown that 18 instances were clustered in cluster 0, 1 instance in cluster 1, 8 instances in cluster 2, and 3 instances in cluster 3. From TABLE 1, it is evident that cluster 2 has questions of "how many" with number answers, and in cluster 3 the questions contain some of the same words, for example the words "name", "prophet", "mentioned", and "Quran " were found in the questions. Cluster 1 has questions containing words that did not appear in any other question like the words "Islamic", "view", and "Abortion". Cluster 1 contains the rest of the questions. These results can be used in further analysis.

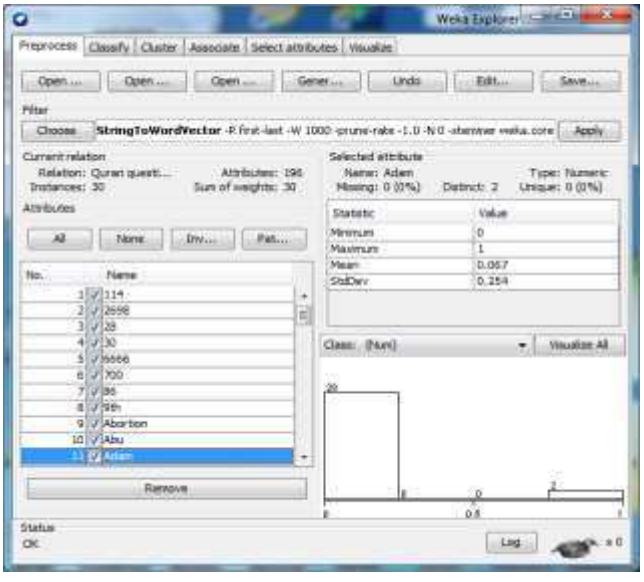


Fig. 1 The processed CSV file in WEKA

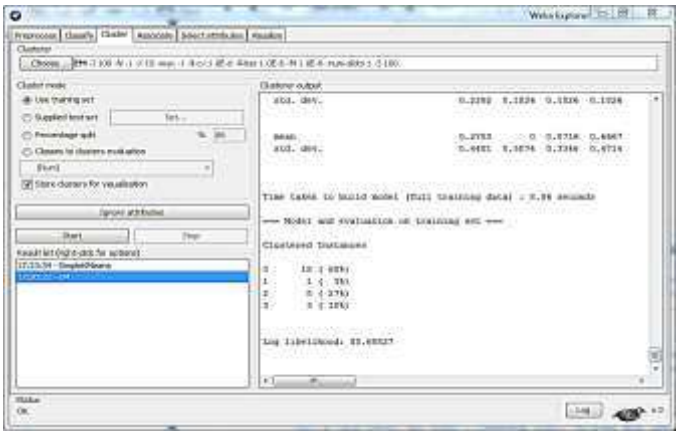

Fig. 2 The clusters and their instances

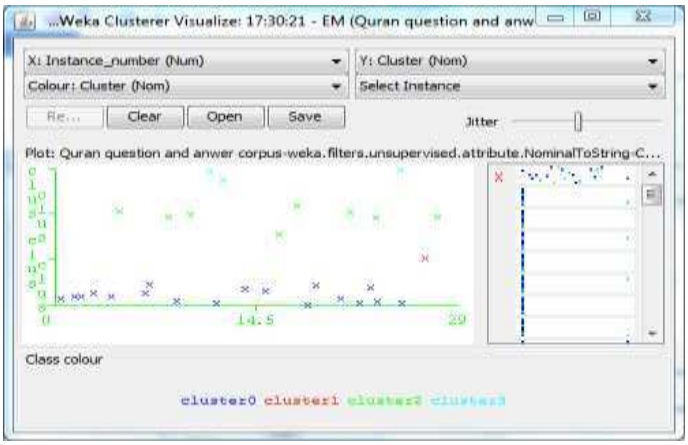


Fig. 3 Visualizing the Quran questions and answers dataset

TABLE 1 Example questions and answers in each cluster

| Cluster no. | Question | Answer |
|---|---|---|
| Cluster 0 | Who revealed the Quran? | Allah revealed the Quran |
| | On which night was the Quran first revealed? | LailatulQadr |
| | Through whom was the Quran revealed? | Through Angel Jibraeel |
| | Who took the responsibility of keeping the Quran safe? | Allah himself |
| | In which Surah (chapter) the law of inheritance is mentioned? | Surah Nesa |
| | What are the conditions for holding or touching the Quran? | One has to be clean and to be with ablution |
| | Why do Muslims believe that the Prophet Muhammad is the final prophet? | Muslims believe that the Prophet Muhammad is the final prophet on the grounds that the Quran and hadith state so |
| | To whom was the Quran revealed? | To the last Prophet Muhammed |
| | Which is the longest Surah (Chapter) in the Quran? | Surah-al-Baqarah |
| Cluster 1 | What is the Islamic view on Abortion? | Islam considers abortion as murder |

| | | |
|---|---|---|
| | | and does not permit it |
| Cluster 2 | How many verses are there in the Quran? | 6666 |
| | How many parts are there in the Quran? | 30 |
| | How many Makki Surah (chapter) are there in the Quran? | 86 |
| Cluster3 | What is the name of the Prophet that mentioned and discussed most in the Quran? | Moosa (Alahis-Salaam) |
| | Who is the relative of the Prophet Muhammed (Sallallahu Alaihi Wasallam) whose name is mentioned in the Quran? | Abu Lahab |

## V. WEKA ERRORS WHILE LOADING DATA

To load data into *WEKA*, we have to put it into a format that *WEKA* understands. *WEKA* needs the data to be present in *ARFF* or *CSV* format in order to perform any tasks. When the format is incorrect while loading a certain file an error will occur; there are some reasons that caused that error for example wrong encoding file format or incompatible characters in the *CSV* Like a percentage sign (%), an apostrophe (‘), incorrect endings, and, any extra commas, etc.

## VI. CONCLUSIONS

The task of creating an integrated Quran question and answer corpus is an important issue, and we would like to encourage researchers to develop a Quran question and answer corpus as a shared task, which aims at improving the state-of-art of online Quran question answering systems. Creating a corpus for data mining with WEKA to extract Knowledge is becoming one of the key tasks for development issues and it plays a vital role for future planning and development. In this paper, we described the compilation of the Quran question and answer collection, through harvesting data from websites, Islamic experts, and existing research datasets. The merging and preparation of the corpus dataset involved removing control characters and solving the problem concerned with some characters, Tthen creating a *CSV* file format, and loading it into *WEKA* for further analysis.

## REFERENCES


[1] Ian H. Witten, Eibe Frank, Mark A. Hall.” Data mining-practical machine learning tools and techniques” (third edition) Morgan Kaufmann Publishers is an imprint of Elsevier

[2] B. Jagtap Sudhir, B. G. Kodge “ Census data mining and data analysis using WEKA” (ICETSTM – 2013) International Conference in Emerging Trends in Science, Technology and Management-2013, Singapore

[3] Motaz K. Saad, Ramzi M. Abed, “Distributed data mining on grid environment”, www.aasrc.org/aasrj American Academic & Scholarly Research Journal Vol. 4, No. 5, Sept 2012

[4] Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald.”WEKA Manual for Version 3-7-11” university of WAIKATO

[5] “Comma Separated Values (CSV) Standard File Format” http://edoceo.com/utilitas/csv-file-format , time of access 14/6/2015

[6] http://www.creativyst.com/Doc/Articles/CSV/CSV01.htm

[7] Csvreader.com, “CSV file format” http://www.csvreader.com/csv_format.php., time of access 17/6/2015

[8] Pete Chapman, Julian Clinton Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer and Rüdiger Wirth “CRISP-DM 1.0. Step-by-step data mining guide “

[9] Prof. Anita Wasilewska, Jae Hong Kil (105228510),

[10] Benjamin Sapp, Ashutosh Saxena, Andrew Y. Ng. “A fast data collection and augmentation procedure for object recognition”, Conference: Proceedings of the twenty-third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008

[11] Turn to Islam Community, “questions-on-Quran”, http://turntoislam.com/community/threads/100-questions-on-quran.10052, time of access 30/5/2015

[12] Islamic Knowledge/Come towards Islam “questions and answers about Quran”, https://islamicknowledge2all.wordpress.com/2011/10/30/question-and-answers-about- quran-3/, time of access 30/5/2015

[13] All Quran “Islamic material/frequently asked questions(FAQ)”, http://www.all-quran.com/islamic_material/frequently_asked_questions.html ,30/5/2015

[14] The siasat daily “Questions and answers about the holy Quran”, http://www.siasat.com/english/news/questions-answers-about-holy-quran?page=0%2C0 ,30/5/2015

[15] B. Abu Shawar, E. S. Atwell “Arabic question-answering via instance based learning from an FAQ corpus”, Proceedings of the CL2009 International Conference on Corpus Linguistics. UCREL. 2009

[16] E. Sherkat “A hybrid approach for question classification in Persian automatic question answering systems”

[17] Khalid Raza, “Application of data mining in bioinformatics” Indian journal of computer science and engineering. ISSN: 0976-5166, Vol 1 No 2, pp.114-118

[18] Wikipedia, “Comma separated value”, https://en.wikipedia.org/wiki/Comma-separated_values

[19] Data, 17/6/2015 Mining http://www.unc.edu/~xluan/258/datamining.html