VFAST Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2014 ISSN(e): 2308-8168; ISSN(p): 2411-6335
Volume 2, Number 1, January-December, 2014

pp. 89-95

A REVIEW of QURANIC WEB PORTALS THROUGH DATA MINING

Mohammad Khubeb Siddiqui[#], Shams Naahid, Mohammad Nazrul Islam Khan
College of Computer Engineering and Sciences,
Salman bin Abdulaziz University,
Kingdom of Saudi Arabia
College of Computer, Qassim University
[1]m.khubeb@sau.edu.sa, s.azamali@sau.edu.sa

ABSTRACT*: In the present era abundant web portals are available over the internet. In the proposed work we concentrate on data mining of Quranic web portals. To know and obtain awareness about Islam, numerous Quranic web portals are being accessed worldwide. Data mining is one of the emerging technologies which analyzes raw data using supervised and unsupervised techniques to find the hidden pattens. This paper is intended to study the access pattern of some of these websites region wise using classification based data mining under which ROC plots have been depicted. The AUC of depicted ROC of considered Islamic web portals are obtained and have been distinguished as to which portal's prediction is more appropriate. Alexa's web-site is an effective tool for obtaining the required data about each of these Quranic web portals regions wise. The study is focused to analyze this data and find the reasons for certain preferences.*
Keywords-Islamic awareness; Quran research; data mining;Classifcation;ROC

Introduction:There has been an unprecedented increase in the amount of data (raw) everyday. Extracting desired and meaningful information from this data involves lot of efforts, and the technique that has been specifically used to extract knowledgeable information, is termed as Data Mining which is popularly known as Knowledge Discovery in Databases (KDD) refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process [1]. The usage of data mining is not limited to one field but extensively applied in different fields of human endeavor, including marketing, banking, engineering and various field of science.

In the present research we use data mining for the review of Quranic web portals and their predictions. Islam is a religion being practiced by 23% (as of 2011) of the total world population with exponential growth rate day in day out. The followers of Islam are called Muslims who believe in one God (Allah) and blissfully follow the teachings of the final prophet Muhammad (PBUH). The Qur'an is the religious book for Muslims. According to Muslim faith Allah (SWT) revealed the Qur'an to the Prophet Muhammad (Peace Be Upon Him) 1444 years ago in classical Arabic language. The Qur'an contains valuable information and provides one stop solution for humanity. It is free from contradictions and discrepancies [6]. The World Wide Web contains various Quranic web portals which contain large amounts of information from the Quran and about the religion of Islam. Some of the websites are preferred over the other in various countries. We study the pattern of usage of some of these Quranic websites. The following are the websites taken into consideration in this research analysis:

~~www.islamicity.com~~*:* This website was launched in 1995 by Human Assistance & Development International (HADI). The website's mission is to share with the world an understanding of Islam and Muslims and promote peace, justice and harmony for all people. It has evolved into one of the world's leading online source of Islamic information and one of the largest Muslim e-Community, offering a wide range of information and services. The website contains various sections such as Quran search which provides a search engine for the words from the Quran, Hadith (saying and acts of Prophet Mohammad (PBUH)) database. Various other sections which provide

the original Arabic script and translation of the Quran in English helping Muslims and non-Muslims understand the Quran and its purpose in a simple and understandable manner. *www.quran.com:* The core competency of this website is to make the verses of Noble Qur'an easily accessible in many languages with features that allow users to browse, search and listen to recitations of the glorious Qur'an. The website provides the full text of the Quran along with the translation in English. The website also host a section which connects to other hyperlink for audio and translation.

www.tanzil.net: This website is a Quranic project launched in early 2007 to produce a highly verified Unicode Quran text to be used in Quranic websites and applications. The mission of this project was to produce a standard Unicode Quran text and serve as a reliable source for this standard text on the web. The website contains the full text of the Quran chapter wise and provides recitation by renowned Islamic scholars.*www.quranexplorer.com:* This website was started in 2005 by a group of Muslims living in USA with a goal to spread the message of Holy Quran in the whole world and advocate peace and tranquility by making people cognizant of the true message of Islam. The website has a strict policy of making sure that all the material comes from a reliable source and has been proof-read by a Muslim scholar ('alim). The website has a section for translation of the Quran in eight different languages with audio available for a few. Also various other sections with information of Hadith and Quran search along with translations from different scholars. Live Quran Tutoring is one the niche characteristic of this website to enable learner is taught by renowned Islamic scholars.

We use the Website of Alexa Web Information Company, which is a part of Amazon.com Company which provides website analytics for all websites country wise. The country traffic rank is a measure of how a website is doing among internet users in a particular country relative to other sites over the past month. The rank by country is calculated using a combination of the estimated average daily unique visitors to a site and the estimated number of page views on that site from users in that country over the past month.
The dataset is formed by grouping the countries into Geographic regions.

Group 1: Asia
Asia is the largest of all the continents with around 59% of all the population of the world living here. Literacy rate on an average for this continent is around 65%. The World Wide Web is a part of life for around 28% population in this continent. All the four websites considered presently have nearly 50% visitors from this continent.

Group 2: Europe
The European Union is a group of developed nations with around 12% of the world population. This group has the least population and a high literacy rate of more than 95%. About 63% of the total population in this continent use internet.

Group 3: Africa
Africa is the second largest continent with 15% of the total world population. The continent has an average literacy of 45%. Though internet is not very prevalent with only 16% of the population using internet, all the four considered websites have a few visitors from this part of the world.

Group 4: Americas, Caribbean and Oceania
This group is the largest in terms of geographical area with Canada, North & South America and Australia. The countries in this group have around 14% of the total world population. These countries have a high literacy rate of above 75% and the highest internet usage of about 56% of the population using internet.

LITERATURE SURVEY: The literature reveals that, [2] author studied a range of Artificial Intelligence and Corpus Linguistics research at Leeds University on Arabic and the Quran, and saw a great potential impact of Artificial Intelligence modeling of the Quran. which has produced a range of software and corpus datasets for research on Modern Standard Arabic and more recently Quranic Arabic .The work on Quranic Arabic corpus linguistics has attracted widespread interest, not only from Arabic linguists but also from Quranic students, and the general public. In [3], Qur'an, AL-Sunnah and Islamic traditional books are the rich resources for Muslims that used as the sole authoritative source of knowledge, wisdom and law. The challenge for computer scientists is to extract and represent these knowledge, wisdom and law in computer systems, this knowledge is directed or underlying, therefore, to build an intelligent systems which can answer any question with knowledge from Quran, Al-Sunnah and other Islamic books, special techniques for mining data must be used to deal with this issue, which can help society, both Muslim and non-Muslim, to understand and appreciate the Islamic religion, this paper attempts to understand how the new techniques in data mining can extract Islamic knowledge from its resources, and represent these knowledge in meaningful form for the user. Moreover, this study concentrates on Hadith as

knowledge resource, and proposes approach to classify Hadith to its categories using supervised learning classification. The finding of this study shows that there are several ways to extract knowledge from Hadith depending on the goal of the knowledge. In [4], Information repositories containing text data of different languages are abundant on the World Wide Web. Digital corpora of sacred text of Islam related to Quran containing Arabic language are also publicly available. The availability of these corpora and intelligent application to analyze them are the vital to better comprehend the religious text of Islam, and propose a method of representing the Quranic text corpus as a graph, and apply a frequent sub-path mining algorithm on it to generate frequent patterns. The research shows that how the frequent patterns can be used for subjective indexing and clustering similar verses of Quran.

Data Collection: The data for the Quranic web portal dataset was collected from (http://www.alexa.com/) [5]. The website provides web traffic data and statistics of a particular website in various countries. The website's traffic data is based on a global panel of Toolbar users. This panel represents a sample of all internet users. The panel consists of millions of people using toolbars created by over 25,000 different publishers, including Alexa and Amazon [5]. The tables had been designed in oracle 10g database named as 'quran_mining' having four attributes (sr_no,name, no_hit,region) and five rows. Attribute 'sr_no' is defined as serial number it's a primary key constraint, 'name' contains the name of quran webportals, 'no_hit' contains the percentage of visitors, and 'region' contains the name of the geographic area.

TOOLS AND TECHNIQUES: Many types of data mining tools are available, such as ODM, Weka, SPSS, etc. every tool has its own pros and cons. In the present work we employed Oracle Data Miner 10.2.0.3.0.1; build 2007 for the prediction of data it act as client and Oracle 10g database served as a server.

EXPERIMENTAL ANALYSIS

Classification: Classification technique is the prominent data mining technique to predict the discrete type values of dataset. The input data, also called the training set, consists of multiple records each having multiple attributes or features. The goal of classification is to accurately predict the target class for each case in the data. In present dataset the target attribute is 'webportal_name' attribute that includes the names of four distinct Islamic websites, it acts as discrete value.

Support Vector Machines (SVM):SVM is a powerful algorithm with strong theoretical foundations based on the Vapnik-Chervonenkis theory. SVM is implemented for classification, regression, and anomaly detection. It has strong regularization properties which refer to the generalization of the model to new data [7]. The Formula for linear SVM is stated as: $u=w^Tx_i + b$, Where **w** is a normal vector (weight coefficient vector), x_i is the input vector and **b** is the bias / intercept term. Based on that, we can get the class **u** where **u** is 1 or -1. The distance between a training vector $\mathbf{x_i}$ and the boundary is called the margin. According to the original theory by [9], we want to find the margin **m** where $\mathbf{w}^T\mathbf{x}_i + b \geq 1$for all $\mathbf{x}_i \in P$ $\mathbf{w}^T\mathbf{x}_i + b \leq -1$for all $\mathbf{x}_i \in N$ and in order to separate the elements which are in a positive or a negative class.

Receiver Operating Characteristic (ROC): The ROC curve allows us to explore the relationship between the sensitivity and specificity, thus allowing the determination of an optimal value. It is often a test is to be carried out, which provides a result on a continuous measure. The vertical and horizontal axis of ROC curve represents the true positive rate and false positive rate respectively.

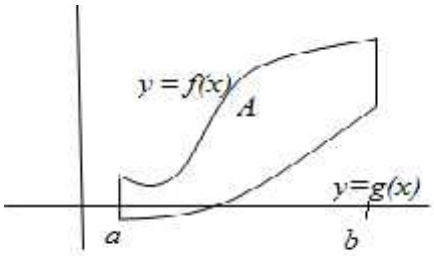
Area Under Curve (AUC): The area under curve measures the discriminating ability of a distinct classification model. The prediction can be based on AUC value. If the AUC value is larger the probability of positivity of the case is more compare to negativity of the case. The AUC is a portion of the area of the unit square whose value always lies between 0 and 1.0.
Area under Curve is calculated by mathematically:

Definition: A straight line on the coordinate plane can be described by the equation y = mx+b. where m is the slope of the line and b is the intercept. If a straight line on the coordinate plane can be described by the equation, $y = m(x-P_x) + P_y$, where m is the slope of the line and Px, Py are the coordinates of a given point on the line.

Remark: An equation of a straight line passing through origin is y=mx .

Area formula: If *f* and *g* are continuous functions on the interval [*a*, *b*], and if $f(x) \geq g(x)$ for all *x* in [*a*, *b*], then the area of the region bounded above by $y = f(x)$, below by $y = g(x)$, on the left by the line $x = a$, and on the right by the line $x = b$ is

$$A = \int_a^b [f(x) - g(x)]dx$$



by using above formulas ,we can easily calculate the area under curve given in fig.1-fig.4.

Confusion Matrix: The AUC works as efficiency measure of the unsupervised data produced by classification technique. During the testing we get correct and incorrect classification from each class. This result is formed as confusion matrix. A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix [8].

Accuracy

The accuracy for the predictive model is obtained by using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

RESULTS AND DISCUSSION

After applying Classification to the datasets respectively the following results are obtained:

www.islamicity.com

TABLE I: CONFUSION MATRIX FOR ISLAMICITY.COM

	Predicted Class		
A c t u a l			**i**
	O		0
	i		2

$$Predicitive\ Analysis = \frac{12 + 2}{12 + 0 + 2 + 2} = 0.87$$

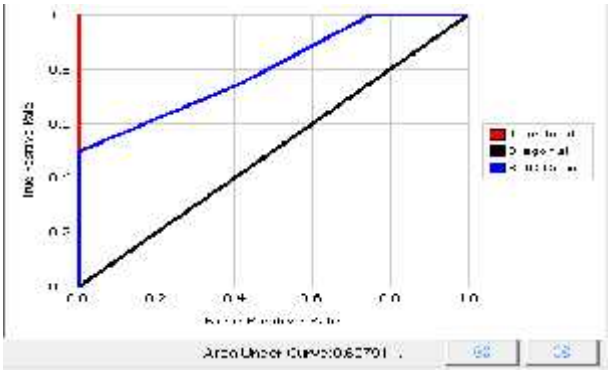


Fig. 1 ROC Plot for www.islamicity.com

www.quran.com

TABLE II: CONFUSION MATRIX FOR QURAN.COM

	Predicted Class		
A c t u a l			**q**
	O		0
	q		1

$$Predicitive\ Analysis = \frac{12+1}{12+0+3+1} = 0.81$$

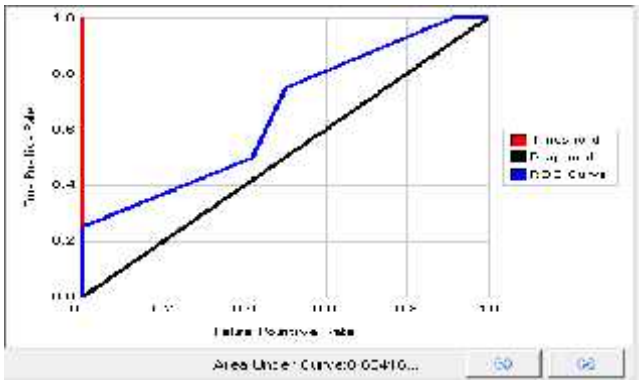


Fig. 2 ROC Plot for www.quran.com

www.tanzil.net

TABLE III: CONFUSION MATRIX FOR TANZIL.NET

	Predicted Class		
A c t u			**t**
	O		0
	t		1

$$Predicitive\ Analysis = \frac{12+1}{12+0+3+1} = 0.81$$

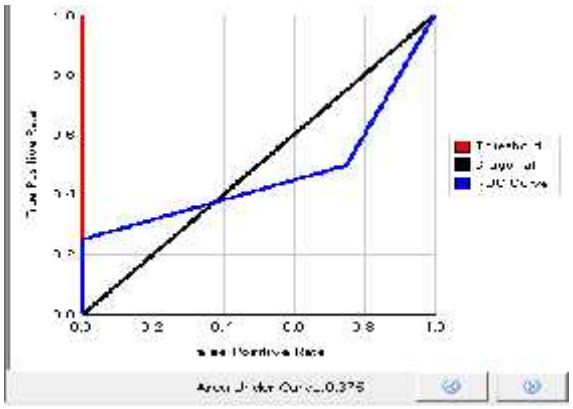


Fig. 3 ROC Plot for www.tanzil.net

www.quranexplorer.com

TABLE IV: CONFUSION MATRIX FOR QURANEXPLORER.COM

	Predicted Class		
A			**qu**

	Ot		0
	qu		2

$$Predicitive Analysis = \frac{12+2}{12+0+2+2} = 0.87$$

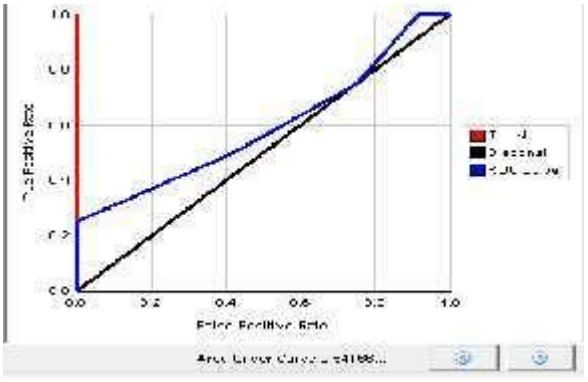


Fig. 4 ROC Plot for www.quranexplorer.com

TABLE V. RESULTS AFTER CLASSIFICATION

Quranic web portals	AUC	Accuracy
www.quran.com	0.60	0.81
www.islamicity.com	0.69	0.87
www.quranexplorer.com	0.54	0.87
www.tanzil.net	0.37	0.81

After applying classification to the dataset interesting results were obtained. The web portal islamicity.com has the highest AUC meaning the predictive model is accurate as it is closer to excellence rate. All the web portals have acceptable results except for tanzil.net which show astonishing result, in the predicted ROC plot AUC of
0.37 is above the threshold, rest of the curve
below 0.5. VII Conclusion

In this research paper, four Quranic web portals were taken into account to analyze the behavior of visitors. These websites are more prevalent and also the form of source of information pertaining towards the Islamic religion for the e-learners. A dataset was created by collecting the data from Alexa's web analytics based on geographic regions. Each Geographic location has its own characteristics. Various factors affect the usage of these websites like population, literacy, internet usage, religion, language, the Islamic months, the popularity of the website among the people and many more.
The message of the Quran can be made more easily available to maximum number of users using the information from the predicted model.

REFERENCES

[1] Han, J., Kamber, M., & Pei, J. (2006). Data mining, southeast asia edition: Concepts and techniques. Morgan kaufmann.

[2] Brierley, C., Sawalha, M., & Atwell, E. (2012). Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing. In *LREC* (pp. 1011-1016).

[3] Aldhlan, K. A., & Zeki, A. M. (2010, December). Datamining and Islamic knowledge extraction: alhadith as a knowledge resource. In *Information and Communication Technology for the Muslim World (ICT4M), 2010 International Conference on* (pp. H-21). IEEE.

[4] Ali, I. (2012). Application of a Mining Algorithm to Finding Frequent Patterns in a Text Corpus: A Case Study of the Arabic. *International Journal of Software Engineering and Its Applications*, *6*(3), 127-134.

[5] Provost, F., & Kohavi, R. (1998). Guest editors' introduction: On applied research in machine learning. *Machine learning*, *30*(2), 127-132.

[6] Vapnik, V. (2000). *The nature of statistical learning theory*. Springer Science & Business Media.

[7] Hennig-Thurau, T., Malthouse, E. C., Friege, C., Gensler, S., Lobschat, L., Rangaswamy, A., & Skiera, B. (2010). The impact of new media on customer relationships. *Journal of service research*, *13*(3), 311-330.

[8] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algo_svm.htm