Improving Knowledge Extraction Of Hadith Classifier Using Decision Tree Algorithm

[1]Kawther Aldhaln, [2]Akram Zeki
Information system dept.
IIUM
KL, Malaysia
[1] k_aldhlan@hotmail.com
[2]akramzeki@yahoo.com

Ahmed Zeki
Information technology dept.
UOB
Al-Manama, Bahrain
amzeki@gmail.com

Hamad Alreshidi
Instruction technology dept.
UOH
Hail, S.A
mr_hamad15@hotmail.com

***Abstract*— Decision tree algorithms have the ability to deal with missing values. While this ability is considered to be advantage, the extreme effort which is required to achieve it is considered a drawback. With the missing values the correct branch could be missed. Therefore, enhanced mechanisms must be employed to handle these values. Moreover, ignoring these null values may cause critical decision to user. Especially for the cases that belong to religion. The present study proposed Hadith classifier which is a method to classify such Hadith into four major classes Sahih, Hasan, Da'ef and Maudo' according to the status of its Isnad (narrators chain). This research provided a novel mechanism to deal with missing data in Hadith database. The experiment applied C4.5 algorithm to extract the rules of classification. The findings showed that the accurate rate of the naïvebyes classifier has been improved by the proposed approach with 46.54%. Meanwhile, DT classifier had achieved 0.9% better than naïvebyes classifier.**

Keywords- Data mining; Decision Tree; Hadith classifier; Missing data; supervised learning algorithm.

I. INTRODUCTION

Data mining is the process of finding patterns that lie within large collections of data. Data mining is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [1]. In addition, data mining techniques enable knowledge to be extracted from data in statistical models to see how variables relate to each other and to better understand the underlying phenomena among them. Data mining methods include neural networks [2], decision trees (DT) [3], cluster analysis, market basket analysis, and regression analysis, among others.

The tree structured modeling is a data mining technique used to recursively partition a dataset into relatively homogeneous subgroups in order to make more accurate predictions on the future instances. Moreover, decision tree algorithms have the ability to deal with missing values, while this ability is considered to be advantage, the extreme effort which is required to achieve it is considered a drawback. The algorithm must employed enhanced mechanisms to handle missing values. However, ignoring these missing data may cause critical decision. In the research case, the ignoring of missing values may cause incorrect Hadith classification that mislead to reject or accept Hadith. Thus, current study is conducted to propose approach to classify Hadith according to the validity of its Isnad (Sahih, Hasan, Dae'f and Maudo'). The target approach using a novel mechanism to deal with missing data in the Isnad attributes. The experiment of the study consists of two phases; training phase and testing phase. The sample of the study is collected from three books in Hadith includes Sahih Al-Bukhari, Jami'u Al-Termithi and Silsilat Al-Ahadith Al-Dae'ifah w' Al-Mawdhu'ah. The evaluation of the proposed algorithm is carried out by comparing the results of classification with the point of view of the expert in Hadith science

II. LITERATURE REVIEW

With respected to the efforts that are provided in computerized Hadith, even for the software that are produced by commercial companies, a few researches are conducted to implement Takhreej Al-Hadith [4]. Takhreej Al-Hadith is the process that grades Hadith according to its validity degree. In this regards, Ghazizadeh *et al.* in [5] used expert system to implement the fuzzy system where the data knowledge base was designed and the essential rules were extracted to determine the validity grade of Hadith, their deduced results were compared with the point of view of expert. The comparison showed that the system was correct in 94% cases. Meanwhile, Hyder & Ghazanfer [6] defined a graph theoretic representation of the chain of narrators and an aligned database structure suitable for storing the biographical data of the narrators and other historical events. Their study aimed to use computer science concepts for algorithmic research, database queries, and data-warehouses besides using of advanced data-mining techniques to assist Hadith research and research in Islamic history and literature. Their way to represent Hadith

978-1-4673-1090-1/12/$31.00 ©2012 IEEE

was amenable for cross verification and analysis in a computationally feasible manner, they found the nodes and arcs with various kinds of weights and then evaluating the aggregate averages over different paths and over the entire graph to yield numerical grades of evaluations. According to their findings the classifications of Hadith are qualitative, and these kinds of aggregate functions would enable quantitative grading of these classifications. Such quantitative grades would make it easier to compare and contrast criteria for evaluations.

Alrazo [7][8] used unsupervised classification to implement the knoweldge of Hadith. Unsupervised learning classification is the process in which the available data instances are divided into a given number of sub-groups, based on the level of similarity between the instances in a certain group. Alrazo intended to describe Hadith knowledge by using Rule– Based method. However, using unsupervised learning required to drive out all the rules that are needed to cluster the data instances.

III. MATERIAL AND METHODS

The current study attempts to reach the same goal of classification using supervised learning algorithms, 999 Hadiths from Sahih Al-Bukhari, Jami'u Al-Termithi and Silsilat Al-Ahadith Al-Dae'ifah w' Al-Mawdhu'ah are framed the sample of the study, the attributes of the Hadith database are calculated according to the validate methods of Hadith science. The sample is divided into two parts (75%) as training dataset to build the classification model, while the rest of the sample (25%) is used to evaluate the performance of the Hadith classifier model. Moreover, the experiment applied C4.5 algorithm to extract the rules of classification. Fig.1 illustrates the research framework using Missing Data Detector method (MDD).

The summary of the process in Fig.1 are as follows: There is a training data set including four classes. Different shapes denote different classes. The whole training data set is portioned in to four classes A1, A2, A3 and A4. Some objects from A1, A2 and A3 have missing attributes that may classify them into incorrect class.

Step1: Applying the proposed mechanism into the training dataset to detect the missing attributes.

Step2: Applying DT algorithm to classify Hadith.

Step3: Some objects are correctly classified, while other objects are still in the incorrect class.

Step4: Building the tree and inducing the rules.

A. Hadith database

According to Tahan [9] there are five conditions must be satisfied to validate the Isnad of Al- Hadith: (1)All narrators in Isnad were renowned for their honesty.(2) All narrators in Isnad were renowned for their accuracy (3)There is no interrupting in the Isnad (4)There is no irregular statement in the Hadith Maten (5)There is no defective in the Hadith Maten. Therefore, the experiment corpus consistes of five basic features (link, defective, irregular, grade of reliability, grade of preservation). Table 1 shows the attributes with the possible values.

TABLE 1
The Attributes of the Training Dataset

ID	link	Irregular	Defective	Grdae Of Reliability	Grade Of Preservation	Class
1	True	False	False	True	True	Sahih
2	True	False	False	True	False	Hasan
3	False	False	False	True	True	Hasan
4	False	False	False	True	False	Hasan
5	True	False	True	True	True	Hasan
6	False	False	False	True	False	Daee'f
7	True	False	False	True	Poor	Daee'f
8	True	False	False	Daee'f	True	Daee'f
9	True	False	False	Daee'f	poor	Daee'f
10	True	False	False	False	True	Daee'f
11	True	False	False	Any	Poor	Daee'f
12	False	True	False	Null	Any	Maudo'
13	False	Any	Any	Matrook	Any	Maudo'
14	False	Any	Any	Monker	Any	Maudo'
15	False	Any	Any	Liar	Any	Maudo'
:	:	:	:	:	:	:
:	:	:	:	:	:	:
:	:	:	:	:	:	:
:	:	:	:	:	:	:

IV. THE PROPOSED CLASSIFICATION APPROACH

The proposed approach consists of four phases; first one is the data pre-processing. Followed by the training phase, the input of this phase is a set of pre-classified documents, while the output is the Hadith classifier model. Whilst, the third phase is the classification (testing) phase which is responsible to test the prediction ability of the proposed classifier. Finally, evaluation phase. As seen in fig.2

V. THE EXPERIMENT PROCEDURES

A. Data Pre-processing

As mentioned earlier, the dataset was collected from different books, therefore, data pre-processing is conducted on each Hadith in the training and testing sets to reduce redundancy and to uniform the style of Hadith.

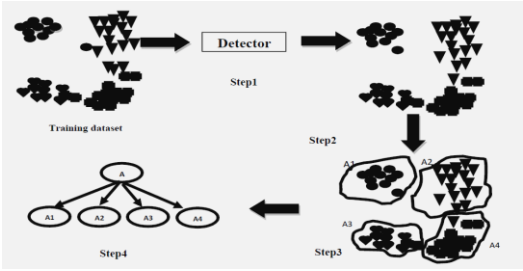

Figure1: Research frame work

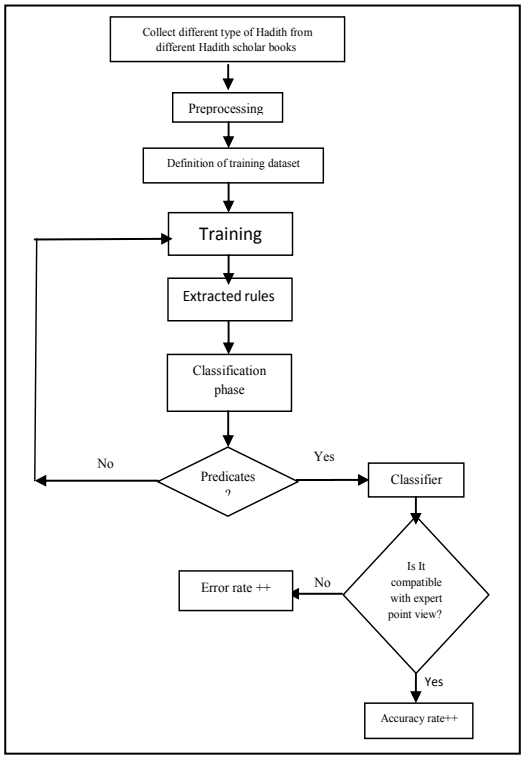

Figure2: The proposed classification phases

This phase includes:

1) *Attaching Isnad*: Some Hadith were separated from their Isnad either for suspicion in the narrator chain or redundancy. This process aimed to attach the Isnad at the beginning of the Maten to facilitate the narrators' chain scanning.

2) *Removing punctuation and diacritical marks:* Removing diacritical and punctuation marks is important since these marks are prevalent in AL-Hadith and have no effect on determining the class of Hadith.

3) *Adding special character:* Adding special character to distinguish between the narrators while scanning Isnad. Table 2 shows the results of the pre-processing stage.

TABLE 2
Results of Preprocessing Phase

Step	Result of the step
Attaching Isnad	عن عبد الله بن سعد الرقي حدثتني والدتي مروة بنت مروان قالت حدثتني والدتي عاتكة بنت بكار عن أبيها قالت: سمعت الزهري يحدث عن سالم بن عبد الله عن ابن عمر أن رسول الله صلى الله عليه وسلم قال: (ما ترك عبد شيئا لله لا يتركه إلا لله إلا عوضه منه ما هو خير له فى دينه ودنياه)
Removing punctuation and diacritical marks	عن عبد الله بن سعد الرقي حدثتني والدتي مروة بنت مروان قالت حدثتني والدتي عاتكة بنت بكار عن أبيها قالت سمعت الزهري يحدث عن سالم بن عبد الله عن ابن عمر أن رسول الله صلى الله عليه وسلم قال ما ترك عبد شيئا لله لا يتركه إلا لله إلا عوضه منه ما هو خير له فى دينه ودنياه
Adding special character	عن .عبد الله بن سعد الرقي, قال حدثتني والدتي .مروة بنت مروان, قالت حدثتني والدتي .عاتكة بنت بكار, عن ابيها قالت سمعت .الزهري, يحدث عن .سالم بن عبد الله, عن .ابن عمر, أن رسول الله صلى الله عليه وسلم قال ما ترك عبد شيئا لله لا يتركه إلا لله إلا عوضه منه ما هو خير له فى دينه ودنياه

B. Experiments Specifications

The target approach is supervised classification. The training dataset is used to be applied by learning algorithm, in purpose to build Hadith classifier model .In the experiments author uses (75%) of AL-Hadith database as training set , while the rest (25%) of the sample is used as testing set. Two algorithms of learning are chosen to run using the same corpus after and before applying the detector method these are C4.5 and naïvebayes**.**

C. Attributes selection

The attributes are selected according to the information gained criteria. Table 3 illustrates the ranking of the features according to this criterion.

TABLE3
The Information Gained Of the Hadith Features

Feature	Information gain after splitting
Link	0.8711
Irregular	0.7927
Defective	0.704
Reliability _ Grade	1.0201
Preservation _ Grade	1.10296

D. Detection Of Missing Attributes

The present study proposed enhanced mechanism to handle the missing attributes in the Hadith database. This mechanism is based on the validate methods of the Isnad [9]:

1) *The status of reliability attribute in the Isnad chain*: Each narrator must be reliable and well known in the narration of Hadith. There are a lot of terms that indicate the reliability status of the narrator. Table 4

summarized these terms and the definitions regarding to the research goals.

2) *The status of the narrators' retention or preservation in the Isnad chain:* In this process the approach determines the value of the preservation for each narrator in the Isnad chain. Table 5 illustrates the terms of narrator's retention.

3) *The status of the link attribute in Isnad chain* : There are three methods to evaluate the status of the Isnad link (a) Tracing the student and the teachers for each narrator. (b) Check the time period between two consecutive narrators. (c) Check the place of each narrator and his journey.

TABLE 4
Hadith Terms Used To Indicate The Narrator's Reliability

Hadith Term	The attribute value
صحابي، أوثق الناس ، ثقة ثقة ، ثقة حافظ ، إمام، ثبت ، عدل، ثقة	True
صدوق، لا بأس به، ليس به بأس، مقبول	False
صدوق سيء الحفظ، صدوق يهم، أو له أوهام، أو يخطئ، تغير بأخرة	False
رمي ببدعة ، رمي بالتشيع، رمي بالقدر، لين الحديث، مستور ، مجهول، ضعيف	Daeef
متروك، متروك الحديث واهي الحديث، ساقط	Matrook
منكر الحديث	Monker
متهم بالكذب ، متهم بالوضع ، كذاب ، وضاع	Liar

TABLE 5
Hadith Terms Used To Indicate The Narrator's Retention

Hadith Term	The attribute value
الضبط	True
خفيف الضبط	False
سيء الضبط	Poor

4) *The status of the defective attribute in the Isnad chain:* This process aims to evaluate the value of the defective attribute of the narrators' chain.

E. *Evaluation Strategy*

It is important to measure the performance of classification model to determine how well the model will perform with new cases. The model performance evaluated after and before applying the detector in the testing phase. Four important measurements are used:

1) Correct Classification Rate (CCR):

CCR is the number of correctly predicted scores by the classifier. It is also known as the accuracy of the classifier. This measurement is represented by (1).

$$CCR=(NCP/NOP)*100 \quad (1)$$

Where CCR, NCP,NOP are the Correct Classification Rate, Number of Correct Prediction and total Number of Predictions, respectively.

2) Error Rate(ER):

Equation (2) represents the mathematical form of the number of incorrect prediction.

$$ER=(NWP/NOP)*100 \quad (2)$$

Where ER, NWP and NOP are the Error Rate, Number of wrong Prediction and total Number of Predictions, respectively.

3) Sensitivity :

The True Positive Rate (TPR) -called also recall- given that the actual value is positive. As represented in (3).

$$TPR=TP/(TP+FN) *100 \quad (3)$$

Sensitivity measures the proportion of actual positives which are correctly identified.

4) Specificity:

The True Negative Rate (TNR) of the classification model given that the actual value is negative, the fraction value classified as true negative [10].

$$TNR= TN/(TN+FP) \quad (4)$$

$$Sp = 1- FP \quad (5)$$

Specificity measures the proportion of negatives which are correctly identified.

5) Receiver Operating Characteristic (ROC) Curve:

ROC curves provide a visual model that displays the trade-off between sensitivity and specificity. The ROC curve is produced by graphing the false positive rate (FPR) which is the same as "1-Specificity" against the true positive rate (TPR) [11].

VI. Results and Discussion

This section presents the main results of the experiment, then capped with a brief discussion. Table 6 illustrates the detailed accuracy by class.

It can be seen from this table that the average of sensitivity of the case (2) has sharply increased with score (97.6%). Furthermore, the average of specificity of the same trial recorded better results (99.4%) than case (2) which indicates that the proposed model performance improved by MDD. And an ROC value result is (0.996) which indicates that the classifier with MDD is performed well with sharp increase of CCR (97.597%).

Furthermore, table7 displays the comparison results between DT classifier and naïvebayes classifier before and after applying MDD.

The findings showed that the accurate rate of the classifier was improved by the proposed approach with (0.9%) above the CCR of the naïvebyes algorihm, on the other hand, the time complexity was effected with (0.05) seconds. In contrast to DT classifier, the time complexity to build naïvebyes classifier remained as it is (0.02) seconds.

TABLE 6
Hadith Terms Used To Indicate the Narrator's Retention

Measurement / Class	Case(1)Before MDD			Case(2) After MDD		
	SEN.	SEP.	ROC	SEN.	SEP.	ROC
Sahih	1	0	0.5	1	0.9994	0.997
Hasan	0	1	0.5	0.988	1	0.994
Da'eef	0	1	0.5	0.971	0.98	0.994
Maudo'	0	1	0.5	0.875	0.996	0.996
Weighted average	0.502	0.498	0.5	0.976	0.994	0.996
CCR	50.1502 %			97.597%		
ER	49.8498 %			2.4024%		

TABLE 7
The Comparison Results Between DT And Naïvebyes Classifiers

	DT Classifier		Naïve bayes Classifier	
	Before MDD	After MDD	Before MDD	After MDD
Time complexity	0.01sec.	0.6 sec.	0.02 sec.	0.02 sec.
CCR	50.1502 %	97.597%	50.1502%	96.6967 %
ER	49.8498 %	2.4024%	49.8498	3.3033 %

VII. CONCLUSION

Sum of all, the researchers can use any book as training data for knowledge extraction research. The holy Qur'an, Hadith and Islamic books are special case. They stand out as the source of a large collection of analysis and interpretation texts, which could provide a gold standard "ground truth" for AI (artificial intelligent) knowledge extraction and knowledge representation experiments. In addition researchers must cross-check for compatibility and consistency with knowledge extraction results from the Islamic corpus. Some computational results may are incompatible with specific inferences, which will shed new light on traditional interpretations. On the other hand, new outcomes may resulted from these experiments, thus adding to the canon of Islamic wisdom. The system that would implement an Islamic knowledge must be reliable because it will be used by billions of Muslims, and non-Muslims.

In the present study, the extracted of Islamic knowledge represent the focal point of the research, three famous books in Hadith science represent the corpus of the study. The proposed Hadith classifier model was built through learning process, DT modeling had represented the structure model of the classifier, and the attributes of the instances originally were obtained from the source books. Whilst some attributes were indicated as null values, or missing data. A novel mechanism was employed to handle these missing data. This mechanism was generated based on the Isnad validity methods in Hadith science. As mentioned earlier, the implementation of the Islamic knowledge is very critical step due to its effects on the Muslim's life. Thus, the results of the research were compared with the resource books, concurrently with the point of view of the expert in Hadith science. The extracted knowledge represented the methods of Al-Imam Al-Bukhari, Al-Termithi and Al-Albani in Takhreej Al-Hadith, their approaches are slightly different. Therefore, it is difficult to claim that the proposed model represent all the Mohadeethen methods. The findings of the research showed that the performance of DT Hadith classifier had significant effect with the MDD. Whilst, the CCR was sharply increased from (50.1502 %) to (97.597%) Furthermore, the favorable results of the present research indicated that the DT Modeling is a viable approach to classify Hadith due to the ease of rules induction and results interpretation.

REFERENCES

[1] Hand, D., Mannila, H., and Smyth, P. *Principles of Data Mining,* Cambridge, 2001, MA: The MIT Press.

[2] Solomon, S., Nguyen, H., Liebowitz, J., & Agresti, W. Using data mining to improve traffic safety programs. Industrial Management and Data Systems , 5, 2006, pp. 621–643.

[3] Kotsiantis, S. B., Supervised Machine Learning: A Review of Classification Techniques. Informatica , 31, 2007,PP: 249-268

[4] Aldhlan, K. A., Zeki, A., & Zeki, A. Encyclopedias of Hadeeth: The current status and future direction . Siminar Warisan Nabawi Kali kedua ,2010 (p. 91). KL, Malaysia: universiti Sains Islam Malaysia.

[5]M.Ghazizadeh, M.H. Zahedi, M.Kahani,andB.M. Bidgoli, "Fuzzy Expert system in determining Hadith validity", advances in computer and information sciences and engineering ,2008, PP.354-359.

[6] S.I.Hyder and S.Ghazanfer, " Towards a database Oriented Hadith Research Using Relational, Algorithmic and Data-warehousing Techniques", The Islamic Culture, Quarterly Journal of Shaikh Zayed Islamic Center for Islamic and Arabic Studies, Vol. 19, University of Karachi, 2008,PP. 14.

[7]H.M. Alrazo, " النبوية للسنة المحوسب الانموذج Computerized frame of the Prophetic tradition', 17th National conferences for computer ,pp. 597-611.Madenh: scientific publishing center,2004.

[8] H.Alrazo, " الإسلامية المعرفة موارد على المعلوماتي التنقيب تطبيقات Data mining application on the Islamic knowledge reource", 2008 . Retrieved JAN 13, 2010, from Alukah : http://www.alukah.net/Culture/0/3123/

[9]M.Tahan," الاسانيد ودراسة التخريج أصول ", Riyadh: Al-Maref publishing ,1996.

[10] Kelly, H., Bull, A., Russo, P., & McBryde, E., Estimating sensitivity and specificity from positive predictive value, negative predictive value and prevalence: application to surveillance systems for hospital-acquired infections. *Journal of Hospital,Elsevier* , 2008, pp. 164-168.

[11] Fawcett, T. ,An Introduction to ROC Analysis. *Pattern Recognition Letters, Elsevier*, 2006, pp. 861-87.