2nd International Conference on E-Learning & Knowledge Management Technologies (ICEKMT 2012)
Date : April 16-17, 2012
Venue : Hotel Corus , Malaysia
http://www.icekm.com

Knowledge Extraction In Hadith Using Data Mining Technique

Kawther A.Aldhaln
Dept. of Computer Science
University of Hail
Hail, Saudi Arabia
k_aldhlan@hotmail.com

Akram M. Zeki
Dept. of Information Systems
International Islamic University Malaysia Kuala Lumpur, Malaysia
akramzeki@iium.edu.my

Ahmed M. Zeki
Dept. of Information Systems
University of Bahrain
Bahrain
amzeki@uob.edu.bh

Abstract :

Muslims believe that the Sunnah of the Prophet Muhammad(SAAW) is the second of the two revealed fundamental sources of Islam, after the Holy Qur'an. Hadith can provide a Gold Standard "ground truth" for AI (Artificial Intelligent) knowledge extraction and knowledge representation experiments. In the present study, the extracted Islamic knowledge represented the focal point of the research, three famous books in Hadith science framed the corpus of the study. This study attempted to explore new approach to classify Hadith using data mining techniques to classify Hadith according to its validity degree (Sahih, Hasan, Da'eef and Maudo'), the proposed Hadith classifier model was built through learning process, DT classifier modeling had been represented by the tree structure model, and the attributes of the instances originally were obtained from the source books. Whilst some attributes were indicated as null values, or missing values. A novel mechanism called missing data detector (MDD) was employed to handle these missing data. This mechanism was generated based on the Isnad validity methods in Hadith science. The results of the research were compared with the resource books, concurrently with the point of view of the experts in the Hadith science. The findings of the research showed that the performance of DT Hadith classifier had significant effect with DDM, the CCR was sharply increased from (50.1502 %) to (97.597%) Furthermore, the favorable obtained results indicated that the DT Modeling is a viable approach to classify Hadith due to the ease of rules induction and results interpretation

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. Contrary to more traditional data analysis methods, which begin with a hypothesis and then test the hypothesis based upon the data, data mining approaches deal the problem from the opposite direction. Thus, data mining is discovery-driven rather than assumption-driven [1]. As the process searches through the data, patterns are automatically extracted. In general, data mining objectives can be placed into two categories: descriptive and predictive [2]. The goal of descriptive data mining is to find general patterns or properties of elements in data sets. This often involves aggregate functions such as mean, variance, count, sum, etc. In other words, descriptive data mining reports patterns about the data itself. Predictive data mining, however, attempts to infer meaning from the data in order to create a model that can be used to predict future data. This is often done by grouping data elements based on similarities and then analyzing the properties those data elements have in common. The common properties should be a reasonable predictor for the given result.

Data mining methods include neural networks [3], decision trees (DT) [4], and 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[5]. In the research case, the ignoring of missing values may cause incorrect Hadith classification that mislead to reject or accept Hadith. Thus, the current study aims to explore new approach to classify Hadith according to the validity of its Isnad (Sahih, Hasan, Dae'f and Maudo') using data mining technique. The target approach proposed 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 is collected from three books namely: 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 experts in Hadith science.

The rest of the paper is organized as follows: literature review in the second part, followed by brief description of the research approach, whilst, the experiment procedures are presented in the fourth section, the results and conclusions are illustrated in the last part.

II. LITERATURE REVIEW

A few researches are conducted to implement Takhreej Al-Hadith [6]. Takhreej Al-Hadith is the process that grades Hadith according to its validity degree. In this regards, Ghazizadeh *et al.* [7] 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 domain experts. The comparison showed that the system was correct in 94% cases. Meanwhile, Hyder & Ghazanfer [8] developed a graphical 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 advanced data-mining techniques to assist Hadith research and research in Islamic history and literature. Their way to represent Hadith 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 evaluated 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 [9][10] 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 manually drive out all the rules that are needed to cluster the data instances.

III. RESEARCH APPROACH

The current study uses supervised learning algorithms in order to classify Hadiths, the sample of the study includes (999) Hadiths from Sahih Al-Bukhari, Jami'u Al-Termithi and Silsilat Al-Ahadith Al-Dae'ifah w' Al-Mawdhu'ah. Hadith database consists of Hadiths and their attributes, some attributes are obtained from the resources where they announced clearly, while other are indicated as missing values where they not clearly mentioned in the resources. The study proposes novel mechanism called missing data detector (MDD) to handle these missing values, this approach is constructed based on the validity methods of the Isnad in Hadith science.

Furthermore, the sample is divided into two datasets; two third (66.7%) frames the training dataset to build the Hadith classifier model(HC), while the rest of the sample (33.3%) is used to assess the performance of the Hadith classifier model (HC). Moreover, the experiment applied C4.5 algorithm to induce the rules of classification. Fig.1 illustrates the research framework using missing data detector (MDD).

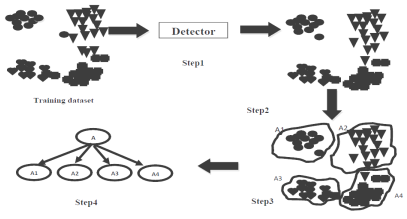

Figure1: Research frame work

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 into four classes A1, A2, A3 and A4. Some objects from A1, A2 and A3 have missing attributes that may classify them into an incorrect class.

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

Step2: Some objects are correctly classified, while some others are still incorrectly classified.

Step3: Applying Decision Tree (DT) algorithm to classify Hadith.

Step4: Building the tree and inducing the rules.

A. Hadith database

According to Tahan [11] 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.

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'
:	:	:	:	:	:	:
:	:	:	:	:	:	:
:	:	:	:	:	:	:
:	:	:	:	:	:	:

TABLE 1 The Attributes of the Training Dataset

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 predictive power of the proposed HC . Finally, evaluation phase, where the classification results are compared with the point of view of the experts in Hadith science. As seen in Fig.2

V. THE EXPERIMENT PROCEDURES

A. Data Pre-processing

As mentioned earlier, the dataset is 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.

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 punctuation and diacritical 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.

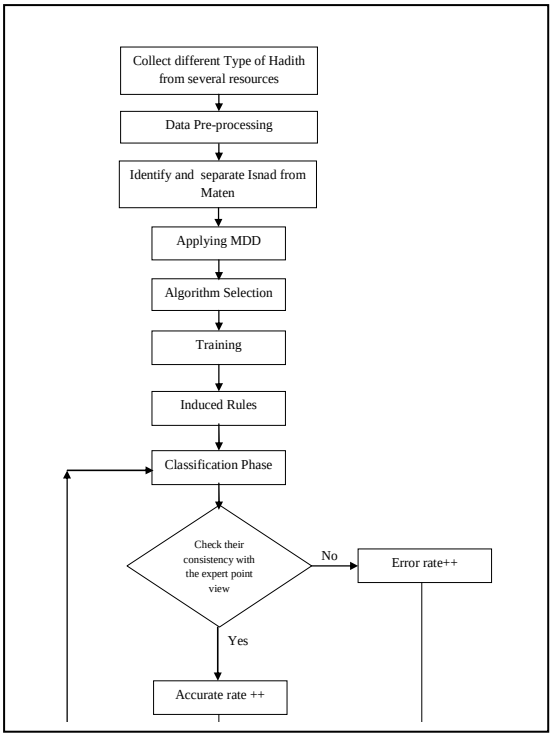

Figure2: The proposed classification phases

TABLE 2 Results of Preprocessing Phase

Step	Result of the step
Attaching Isnad	عن عمر بن يحيى قال: حدثنا شعبة الحجاج عن ثور بن يزيد عن خالد بن معدان عن معاذ بن جبل مرفوعا (قلوب بني آدم تلين في الشتاء وذلك لأن الله خلق آدم من طين، والطين يلين في الشتاء)
Removing punctuation and diacritical marks	عن عمر بن يحيى قال حدثنا شعبة الحجاج عن ثور بن يزيد عن خالد بن معدان عن معاذ بن جبل مرفوعا قلوب بني آدم تلين في الشتاء وذلك لأن الله خلق آدم من طين والطين يلين في الشتاء
Adding special character	عن .عمر بن يحيى, قال حدثنا .شعبة الحجاج, عن .ثور بن يزيد, عن .خالد بن معدان, عن .معاذ بن جبل, مرفوعا قلوب بني آدم تلين في الشتاء وذلك لأن الله خلق آدم من طين والطين يلين في الشتاء

B. Experiments Specifications

The target approach uses training dataset in purpose to build Hadith classifier model (HC) which frame two thirds of the sample. While the rest is used as testing dataset to evaluate the performance of HC. In addition, two algorithms of learning are chosen to run using the same corpus after and before applying MDD, these algorithms are C4.5 and naïvebayes.

C. Attributes selection

The attributes are selected according to the information gained criteria. Information gain (IG) has been used in C4.5 [14]. This is based on the entropy measure, see eq.1 (originally from information theory):

$$I_{entropy} = -\sum_i p(i)\log(p(i)) \qquad (1)$$

Information gain is defined by eq.2:

$$IG = -\sum_{J=1}^{J} p(j)\log_2\ p(j) + \sum_{m-1}^{M} p(m)\ \left[\sum_{J=1}^{J} p(j/m)\log_2\ p(j/m)\right] \quad (2)$$

Where J is the total number of classes and M is the total number of splits.

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. Missing Data Detector (MDD)

The present study proposed enhanced mechanism to handle the missing attributes (MDD) in the Hadith database. This mechanism is based on the validate methods of the Isnad [11]:

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 journies.

TABLE 4 : Hadith Terms Used to Indicate the Narrator's Reliability in the study

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 in the study

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

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 MDD 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 (3).

$$CCR=(NCP/NOP)*100 \quad (3)$$

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

2) Error Rate(ER):

Equation (4) represents the mathematical form of the number of incorrect prediction.

$$ER=(NWP/NOP)*100 \quad (4)$$

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 (5).

$$TPR=TP/(TP+FN) *100 \quad (5)$$

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 [12].

$$TNR= TN/(TN+FP) \quad (6)$$

$$Sp = 1- FP \quad (7)$$

Specificity measures the proportion of negatives which are correctly identified.

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 DM. And an ROC value result is (0.996) which indicates that the classifier with MDD performed well with sharp increase of CCR (97.597%).

TABLE 6 : The detailed accuracy of HC before and after MDD

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

V. CONCLUSION

In summary, 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 result 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 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 Mohaddithеen 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] Radivojevic, Z., Cvetanovic, M., & Milutinovic, V. "Data Mining: A Brief Overview and Recent IPSI Research". Annals of Mathematics, Computing, and Teleinformatics , 2003, 1 (1), 84-91.

[2] De Raedt, L., Blockeel, H., Dehaspe, L., & Van Laer, W. " Three companions for data mining in first order logic". In S. D. Lavrac (Ed.), Relational Data Mining, 2001, (pp. 105-139). Verlag: Springer.

[3] 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.

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

[5] Berson, A., Smith, S., Thearling, K. Building Data Mining Applications for CRM . 2000, USA: The MCGraw-Hill.

[6] Aldhlan, K. A., Zeki, A. M., & Zeki, A. M. Encyclopedias of Hadith software: The current status and future view. The Third National Information Technology Symposium "Arabic and Islamic Contents on the Internet", 2011(6-7 March). Riyadh, S.A: KSU.

[7]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.

[8] 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.

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

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

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

[12] 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.

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

[14] *Quinlan, J.R. C4.5: Programs for Machine Learning, 1993, San Mateo, CA, Morgan Stanley.*