DATAMINING AND ISLAMIC KNOWLEDGE EXCTRACTION: ALHADITH AS A KNOWLEDG RESOURCSE

Kawther A.Aldhlan Ahmed M. Zeki Akram M. Zeki

Faculty of Information and Communication Technology
International Islamic University Malaysia

k_aldhlan@hotmail.com amzeki@gmail.com akramzeki@yahoo.com

Abstract **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 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.**

Key words : Islamic knowledge , intelligent system , Data mining.

I INTRODUCTION

Access to the underlying knowledge in the Islamic sources requires interpretation and inference; much knowledge is encoded via subtle use of words, grammar, allusions, links and cross-references. For over a thousand years, scholars have sought to extract knowledge and laws from the text, and have built up a much of analyses, interpretations and inference chains.

Computer Science and Artificial Intelligence presents the opportunity to re-analyze the text data, extract and capture the underlying knowledge in a Knowledge Representation and Reasoning formalism, and enable automated, objective inference and querying [1]. Current study attempts to understand the need for extracting Islamic knowledge and the role of the data mining and the other intelligent techniques to represent and extraction this knowledge, present study adopts Hadith as a knowledge resource study to explore the diverse intelligent methods that were applied on them.

II THE NEED FOR EXTRACTING ISLAMIC KNOWLEDGE

Recent years have witnessed the rapid increase of data all around the world, therefore, the need for extracting Islamic knowledge become more important to distinguish the true knowledge and the fabricated one. Moreover, current systems can answers "factoid" questions from the source text, but many potential questions are more difficult and contentious to answer via text-match, requiring a new Knowledge Representation and Reasoning formalism capable of capturing complex, subtle knowledge encoded in the Classical Arabic, and inferencing in new ways which mirror the thousand-year–old traditions of scholarly analysis.

In principle, researchers can use any book as training data for Knowledge Extraction research. However, 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. That mean the answers are always "correct" in that they are consistent with and supported by evidence from the source text, have a demonstrable chain of inference

III *SPECTRUM OF DATA, INFORMATION AND KNOWLEDGE*

There are three theories of knowledge [2][3]. First, the idealistic theory like Plato believes that knowledge was a function of the recollection of previous information. Second, the materialistic theory that believes in five senses. They consider sense perception as the source or means of knowledge. Third, the Islamic theory that believes in the existence of matter as well as soul. By that, knowledge is a complex concept and it is not easy to define because it is not easily understood, perceived, and measured. It has absolute truth, or ground truth, which describes the rich truths of real situation experiences [4][5].

However, most people have some understanding of what knowledge is. Knowledge, information, and data are not interchangeable concepts. A brief comparison of data, information, and knowledge based on literature are tabulated in Table 1. It shows that data, in and of itself is a symbol, is out of context and with no value until processed into useful forms. By adding meaning, values, and searching for context to make sense of data, this context reveals the structure or relationship (or both) that organizes the data into information. Knowledge is the process of making sense of information. Examples of knowledge are patents, recipes, formulas, instructions, and designs. Without the dimension of context, culture, tacit, and time, knowledge will be little more than information. Thus, knowledge has more to do with who is interpreting the information (their own principles and values) than the objective information on which it is based.

IV *THE KNOWLEDGE DISCOVERY PROCESS*

To meet the requirements of huge data sets, the research areas of knowledge discovery in databases (KDD) and data mining have emerged in the recent years, with multiple books [6][7][8] and numerous papers [9][10][11], surveys and theses [12][13] to mention a few. Often the terms data mining and knowledge discovery are used interchangeably, however, in the strict sense data mining is one step in the KDD process, which is defined as follows in [9]: Knowledge discovery in databases is the non-trivial process of identifying novel, potentially useful, and ultimately understandable patterns in data. Figure1 gives an overview on the KDD process which comprises five major steps:

1. Data Selection. As a first step the data needs to be carefully selected. Selection criteria include e.g. data availability, quality, type, format, and semantics. The selection of high quality data semantically corresponding to the goal of the discovery process is essential for the following steps.

2. Preprocessing. The target data often requires preprocessing. Suitable strategies on scaling and normalization of the features and strategies to handle missing attribute values have to be selected and applied.

3. Transformation. To reduce the dimensionality of the data, dimensionality reduction techniques, which derive transformed representations of the original features. Feature selection techniques reduce the dimensionality by identifying features which are useful for the goal of the discovery process.

4. Data Mining. Depending on the goal of the discovery process, a suitable data mining method is selected. The decision on the data mining algorithm is not easy, since for most tasks there are a huge variety of possibilities. The selected algorithm is applied on the preprocessed and transformed data. For most data mining algorithms it is also a non-trivial task to find appropriate parameter settings.

5. Interpretation and Evaluation. The results of the data mining algorithm are analyzed and interpreted. If the results are not satisfactory, there may be the need to go back one or more steps. In fact, the KDD process is an iterative process. In addition to feature selection, (belonging to the transformation step). The next section gives an overview on the various kinds of data mining methods.

Supporting Literatures	Data	Information	Knowledge
[14]	symbols, numbers, or characters	process, informed mental state, commodity, product, or thing	as a thing
[4]	set of discrete, objective facts about events - structured records of transactions in organizational context	document or audible or visible communication - has meaning the "relevance and purpose" - data becomes information when its creator adds meaning by adding value	a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information
[15][16]	record, store and maintain attributes	when add value in some way	when it adds insight, abstractive values, and better understanding
[17]	symbols represent objects, events and/or their properties - out of context - no value until processed into useful forms	a function of processed or structured data containing both the data and its relationship - provide objective descriptions - content oriented	- has both collective and personal components - has tacit and explicit nature - is the process of making sense of information
[23]	Physical process	It is a constituent element of all physical processes and hence cannot treated as something epiphenomenal to the economic process. It must be engaged in on its own terms	Extracted information from data through Effective cognitive strategies

Table1: A brief comparison of data, information, and knowledge

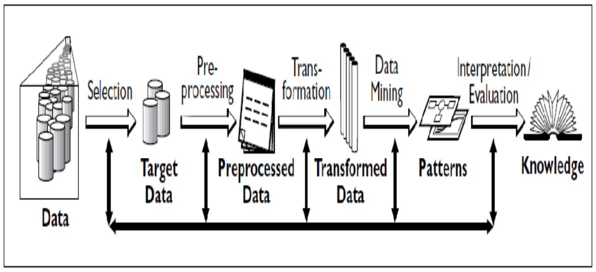


Figure1: Overview of the steps constituting the KDD process

V DATA MINING TECHNIQUES

As a crucial step of the KDD process, data mining requires the selection of a suitable data mining algorithm. The goal of the discovery process. Following a common characterization [8] the diverse data mining methods can be categorized as follows:

• **Clustering:** Find a partitioning of the objects of the data sets into groups (clusters) while maximizing the similarity of the objects in a common cluster and minimizing the similarity of the objects in different clusters.

• **Outlier Detection**: Find objects in the data set which are exceptional, i. e. which do not correspond to the general characteristics or model of the data.

• **Classification**: Learn a function, model or other method from a subset of the data objects to assign a data object to one of several predefined classes Association Analysis: Find subsets of the attributes or subsets of attribute ranges which occur frequently together in the data set (called frequent item sets). Derive so called association rules from the frequent item sets. The association rules describe common properties of the data.

• **Evolution Analysis**: Discover and describe regularities or trends for objects with properties that change over time.

• **Characterization and Discrimination**: Summarize general properties of the data set or of a set of features (characterization). Compare different subsets of the data to other subsets (discrimination).

VI AL HADITH AS KNOWLEDGE RESOURCSE

There are several aspects that can drive as a knowledge form alhadith. For instance term legislative and jurisprudential, alsyrh Alnabwia (prophet's life), Islamic military basis, Hadith classification …etc.

Current study attempts to computerize Al-Ahadith to classify them into their categorized. Some of the previous studies were conducted different approaches to reach the same goal of this study. Ghazizadeh *et al.* in [18] used an expert system software to implement the fuzzy system where the data knowledge base has been designed and the essential rules have been extracted to determine the rate of validity of Hadith, two inference engines have done to get out the result. The output of the first engine is the Rank of each narrator and it will be an input for the second inference engine. The output of the second engine is Hadith validation rate. They divide their system in rules in two blokes .First rule blokes focus on the personal characteristics of narrator and the second rule blocks concentrate on Hadith. Their study sample was collected from KAFI ,volume1 .And the results deduced from designed expert system were compared with expert view points. The comparison showed that the system was correct in 94% cases.

However, there are two main data mining techniques that can be used to achieve this outcome and they are called Supervised and Unsupervised Classification techniques.

Hyder & Ghazanfer in [19] defined a graph theoretic representation of the chain of narrators of Hadiths 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, data-warehouses besides using of advanced data-mining techniques to assists 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 that Annotation of 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 finding 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.

Attempts of Alrazo in [20][21] to computerzied hadith by using unsupervised classification. 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. These sub-groups are called clusters.

He discribed his methodolgy to computeriz hadith as follow :

1- Collect all resources that related to classify hadith from traditional books and analyze them to figure out the facts that irrefutable and the general rules which can be used to construct knowledge.

2- Translate rules into suitable computer code.

3- Construct Hadith knowledge model based on the anlyst facts and rules.

4- Evaluate the authenticity of the model .

According to (Alrazo) the Hadith knowledge resources contains Data, information and knowledge as shown in table 2.

	Information Resources within hadith science	Required Processing	Value
Data	All Hadith resource (Sahih books, Sunn, Musnad…etc)	1- Create database for hadith Narrators. 2- Analsis resources to findout the rules of classification according to Muhadeth methodology.	Strong correlated with the hadith context and its reference.
Information	Rules of classification according to the experts point view.	1-Form rules as a computer code . 2- construct rules that associative with narrators database.	Addition value in processing hadith context.
Knowledge	Produces new hadith knowledge through mining the text of hadith.	Constructs biases for the expert system that classify hadith .	Addition value according to nature of knowledge goals .

Table 2 : Components of Hadith knowledge resources.

Alrazo intended to describe Hadith knowledge by using Rule – Based method. Which is based on IF-→ then. However, using unsupervised learning required to drive out all the rules that is needed to classify or cluster the data instances, It kind of machine learning but it is not enough to consider it as an intelligent machine.

In contrast, supervised learning classification can learn from the data instants in the training data set, to get out the associative rules to classify the data in the test set, the current study intends to classify hadith according to their validity include Sahih(sound, true), hasan (good, agreeable), da'eef(weak) and Maudoo' (fabricated).

Four kinds of hadith validity will be considered as the independent variables, while the attributes of each class will be considered as dependant variables of the experiment.

To accepts any hadith, the hadith must satisfy the following conditions [22] :

- All narrators in Isnad were renowned for their honesty.
- All narrators in Isnad were renowned for their accuracy.
- There is not interrupting in the Isnad .
- There is no irregular statement in the Hadith Maten .
- There is no defective in the Hadith Maten.

The current approach runs under two assumptions:

1- Rejal Ahadith database is existed that mean :-

a- The reliability of each narrator must be determined and the evaluation of the reliability must be determined before inputting in the attribute tables, four values will indicate in this table (reliable, weak, abandoned and liar).

b- The Exactitude of each narrator must be determined and the evaluation of the reliability must be determined before inputting in the attribute tables, three values will indicate in this table (excellent, good, poor).

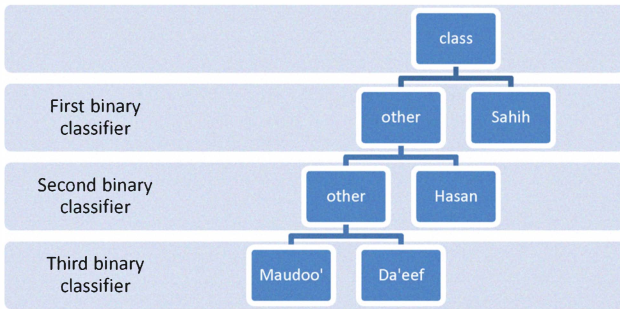
c- The continuity of the sand chain must be determined before inputting in the attribute table, two values will indicate in this table (True, false).

2- Within the scope of the study, the research will assume that Maten is regular and free of defectives. So that table 3 describes the different attributes of the data instances.

Depended variables					Independent variable (class)
Continuose	Irregular	Defective	The reliabilty evaluation of the narrators	The exactitude evaluation of the narrators	
True	False	False	Reliable	Excellent	Sahih
True	False	False	Reliable	Good	Hasan
True	False	False	Reliable	Poor	Da'eef
True	False	False	Weak	Excellent	Da'eef
True	False	False	Weak	Good	Da'eef
True	False	False	Weak	Poor	Da'eef
True	False	False	Abandoned	Excellent	Da'eef
True	False	False	Abandoned	Good	Da'eef
True	False	False	Abandoned	Poor	Da'eef
True	False	False	liar	Excellent	Maudoo'
True	False	False	liar	Good	Maudoo'
True	False	False	liar	Poor	Maudoo'
False	False	False	Reliable	Excellent	Da'eef
False	False	False	Reliable	Good	Da'eef
False	False	False	Reliable	Poor	Da'eef
False	False	False	Weak	Excellent	Da'eef
False	False	False	Weak	Good	Da'eef
False	False	False	Weak	Poor	Da'eef
False	False	False	Abandoned	Excellent	Da'eef
False	False	False	Abandoned	Good	Da'eef
False	False	False	Abandoned	Poor	Da'eef
False	False	False	liar	Excellent	Maudoo'
False	False	False	liar	Good	Maudoo'
False	False	False	liar	Poor	Maudoo'

Table3: The attributes of training data set

Current study adopts Multi Binary classifier as shown in figure 2.



VII CONCLUSION AND FUTUER WORK

Data mining has powerful techniques to extract the knowledge, Islamic knowledge is special case, because of any fault in the extraction could misled Muslims as well as. none Muslim. However, It essential to use this tool to mining all data and information that had been released in the holy Qur'an, Hadith and Islamic traditional books. Which is difficult to do manually. The techniques that would be used are different regarding to the purpose of the knowledge.

Unsupervised learning classification and supervised learning classification are the two famous techniques that used to classify hadith. The first techniques required to indicates all rules that necessary for classification. On the other hand, supervised learning can learn from training set to induct the rules, post test are the second step to evaluate the accuracy of the model. Farther study will conduct other algorithms rather than decision tree to construct classification model.

REFERENCES

[1] E. Atwell, K.DukeS, A. Sharaf, N.Habash, B.Louw and B. Abu Shawar, "Understanding the Quran:a new Grand Challenge for Computer Science and Artificial Intelligence", New York, USA.

[2]A.As-Sadr, " Our Philosophy. USA: Routledge and Taylor & Francis Group" , 1987.

[3] P Cornford," Plato's Theory of Knowledge", Indianapolis, Indiana: Bobbs-Merrill Company, Inc,1957

[4] T.H.Davenpor and L.Prusak, " Working Knowledge: How Organizations Manage What They Know", Harvard Business School Press, 2000.

[5]P.F.Drucker, "The essential Drucker: Selections from the management works of Peter F. Drucker", New York: Harper Business,2001.

[6] M.Ester and J. Sander, "Knowledge Discovery in Databases". Springer Verlag , Berlin, Sept.2000.

[7] I.H.Witten and W.Frank, " Data Mining - Practical machine learning tools and techniques with java implementations". Morgan Kaufmann, 2000.

[8] J.Han and M.Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2000.

[9] U.M.Fayyad, G.Piatetsky-Shapiro and P.Smyth, " Knowledge discovery and data mining: Towards a unifying framework". in Proc of KDD Conference, pp. 82–88, 1996.

[10] R.T.Ng and J.Han, " Efficient and effective clustering methods for spatial data mining. in Proc. of VLDB Conference, pp. 144–155,1994.

[11] S.Papadimitriou ,H. Kitagawa, P.B.Gibbons and C.Faloutsos, "LOCI: Fast OutDetection Using the Local Correlation Integral", in Proc. of ICDE Conference, pp. 315,2003.

[12] P.Kr¨oger, "Coping with new challenges for density-based clustering", PhD thesis. LMU Munich, 2004.

[13] J.Y.Pan, "Advanced Tools for Multimedia Data Mining", PhD thesis. Carnegie Mellon Univerity, Pittsburgh, PA, 2006.

[14]E.J. Quigley and A.Debons, " Interrogative theory of information and knowledge". SIGCPR conference on Computer personnel research. New Orleans, Louisiana, United States: ACM, 1999.

[15] I. Spiegler, "Knowledge Management: A New Idea or a Recycle Concept", Communications of Association for Information Systems(13) 4, 2000 .

[16] I. Spiegler, "Technology and knowledge: bridging a "generating" gap". Information & Management, Volume 40, Issue 6, pp. 533-539, 2003.

[17]P. Kaipa, "Knowledge architecture for the twenty-first century", Behaviour & Information Technology, 19 (3), pp.153-161,2000.

[18] 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 , PP.354-359, 2008.

[19] 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, PP. 14, 2008.

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

[21] 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/

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

[23] M. Boisot and A.Canals," Data, information and knowledge: have we got it right?", J Evol Econ ,vol.(14)pp. 43–67,2004.