Information Mining from Muslim Scriptures

Abdul Rauf Saeed
PUCIT, University of the Punjab
Lahore, Pakistan
bsef09m037@pucit.edu.pk

Syed Waqar Jaffry
PUCIT, University of the Punjab
Lahore, Pakistan
swjaffry@pucit.edu.pk

Abstract

There are billions of believers of various religions in the world and Islam is the second largest religion having 1.6 billion followers. The primary written sources of religious beliefs and practices of Muslims are the Quran and the Hadith (saying and practices of their prophet Muhammad Peace Be Upon Him). Written text of the Quran and the Hadith books is of manageable size and hence state of the art text mining techniques can easily be applied on it. In this paper first a comprehensive review of existing applications offering various type of the Quran and the Hadith information retrieval is presented then a framework based on text mining techniques is proposed. Finally an application is developed to demonstrate the Quran and the Hadith information retrieval framework. This application is evaluated with the help of an end user assessment questionnaire. It is recorded that end users have observed salient advantages of the designed application.

1 Introduction

There are several living religions in the world and the Islam is the second largest religion with respect to its followers (Ali, S. R., Liu, W. M., Humedian, M., 2004). Millions of Muslims (believers of Islam) receive their inspirations for the primary source of their religious beliefs and daily practices from the holy book "Quran" and sayings of their prophet Muhammad (PBUM) called "Hadith". Textual data stored in the Quran and the Hadith books is not huge and could be processed efficiently using the state of the art text analytic techniques. These machine based text processing techniques have the potential to summarize, analyze and effectively present different concepts present in these books. In this paper an initial effort based on research and implementation of a framework is presented which can help in processing and understanding this data. This paper also includes end users evaluation and author's conclusion and future work regarding this ongoing project. It is believed that this work would be useful for the researchers, and implementers who want to conduct similar research and development.

2 Related Work

The related work done on this field is very limited and can be extended further. Currently, most of the Quran and Hadith information retrieval systems are not well-formed as these systems search on basis of book name, volume number and the Hadith number and in case of the Quran it is the Surah (chapter) name or number and verse number. These kinds of systems are using primitive type of parameterized information retrieval. There are various systems like "The Islamic Search", "SearchTruth", "Allah.pk", "Islamic City", "IslamicSearch.org" and "IntoIslam" etc. The features of these systems and their working are briefly discussed as follows.

- The Islamic Search (TheIslamicSearch, 2013): It is an Islamic information retrieval system that takes a query from the user and it did not give results from Quran or Hadith but from the web by searching query with respect to some Islamic keywords (such as "Islamic", "Hadith", "Quran", "Islam" etc.). So, basically, it is not a information retrieval system based on the Quran and the Hadith but it is based on Google web search engine.
- SearchTruth (SearchTruth, 2013): It is a system that searches on the basis of exact keyword and the substring matching. For example, if we will search word "ski" instead of "sky", it will give Hadith or verse contains "asking".

- Allah.pk (Allah.pk, 2013): It is a search engine that searches the results on the basis of exact keyword matching. If the user enters wrong spelling or any word that is not present in the Quran or the Hadith, it will display no record found. Its plus point is that user can search in Arabic also and its results both in Arabic verse and its English translation.
- Islami City (IslamiCity, 2013): is a search engine that searches on the basis of keywords and give results. The searching only involves exact string matching in translation of the Quran. In results, the Arabic verse, its translation and phonetic verse is shown as well.
- Into Islam (intoIslam, 2013): is search engine similar to "The Islamic Search" in which user's query is searched through Google search API.

These are the well-known information retrieval systems based on the text of Quran and Hadith. It is observed that these all systems lack even primitive preprocessing and result ranking techniques used in information retrieval systems. Web Search Engines like Google are the examples of information retrieval systems. Hence the framework presented in this paper uses stemming and related algorithms to create inverted index and term frequency and inverted document frequency (TF-IDF) for result ranking. Furthermore proposed system uses synonyms for user query expansion to generated relevant results.

3 Methodology

This section provides a brief about the methodology followed in this research which includes data collection, preprocessing and the methods used for storage, searching and ranking of the results.

3.1 Data Collection

Three books of Hadith and the Quran are used. Hadith books include Sahih-Bukhari, Sahih-Muslim and Sunan-abi-Dawood. Texts for these books are taken from (Bihar Anjuman, 2013).

3.2 Data Pre-processing

In data pre-processing, first these books are converted from pdf files to text files in order to use these files in programming easily (Convertonlinefree.com, 2013). An online converter is used to convert the pdf files to text files. Furthermore, names of narrators are extracted from these Hadith books after tokenizing the words in file. Porter Stemmer algorithm (Porter, 1980) is used to stem the words.

After creating an index for Narrators the primary data set is again used to extract candidate keywords by applying RAKE keyword extraction algorithm. RAKE (Rapid Automatic Keyword Extraction) algorithm is used for extracting keyword from individual documents. In this algorithm, text is basically split on the basis of stop words and then a score is assigned to each phrase in the document. Stop words are the words that have higher frequency in text but they are just useless and cannot be used as a keyword for searching in the text such as (is, am, a, are, in etc.). After keyword extraction, the inverted index is created. Inverted index is a structure in which data is stored as key/value pair. In this structure, extracted keywords are placed as key and the list of indexes (locations in respective data file) is placed as value.

3.3 Method

There are three basic modules in this framework that are Hadith searching, Quran searching and searching from the both modules.

The first module is to search the Hadith. In this module, the user selects the book, narrator name and then enters query. As the user clicks 'search' button, the narrator name and query words are passed to a function where the query words are preprocessed. Preprocessing consists of conversion of text to upper case, removal of stop words, obtaining list of synonyms for each word and then stemming of each word. When the query is preprocessed the system gets the query word by word and gets the index(es) of each word and of its synonyms from the inverted index table. After getting the indexes, the system searches for the index values in Hadith books. After getting index value, it extracts whole Hadith and check the Narrator name. If it gets the combination of "NARRATED" and "NARRATOR NAME" in any line, it starts storing the Hadith in a variable and adds the variable in a list and again start searching for next index until the end of indexes. Now the list contains all the Hadith that is narrated by narrator selected by the user and related to user query. The list is then iterated for calculating the TF-IDF of all selected Hadith. The formula for calculating TF-IDF is mentioned below:

$$TF\text{-}IDF = TF(t, d) * IDF(t, D)$$

TF (t, d) = frequency of token *t* in document *d*

IDF (t, D) = log (|D| / {d in D, t in d})

Where,

t = token (word)
d = a specific document
D = all documents

In the above formula TF is equal to the ratio of frequency of each query word and its synonym in each Hadith to overall frequency of word in all Hadith. IDF is calculated by taking ratio of total Hadith to the count of Hadith in which the specific word found and then taking *log* of this ratio returns the IDF. Now, TF-IDF is calculated by taking product of TF and IDF for each word and Hadith. The list data is then sorted in descending order of their TF-IDF score and the final results are shown to the user.

In Quran searching, the user selects the Surah of Quran and then enters the query. The query is then pre-processed as mentioned above and the TF-IDF is calculated between query words and verses of the Quran. The verses are then sorted in descending order of TF-IDF score and finally the results are shown.

In searching from Quran and Hadith, both above mentioned methods are applied together. Detailed follow chart of the system is depicted in figure 1.

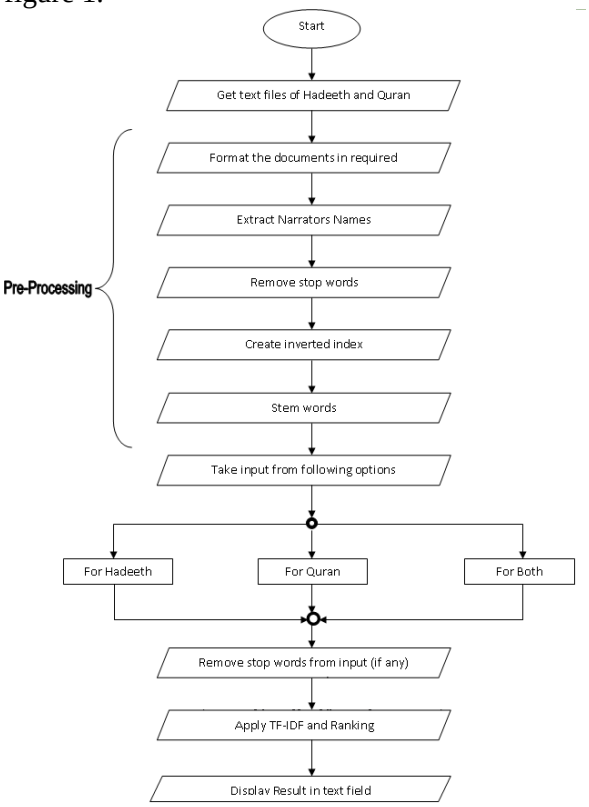

Figure 1. System Flow Diagram

4 Results

Results are measured on the basis of efficiency and usability. The following sections provide the details about results of both types.

4.1 Efficiency

For efficiency, time is noted for five cases, when number of query words range from one to five. The graph in figure 2 shows the time taken in seconds (y-axis) for the number of words in user query (x-axis). Here it can be observed that relationship between query size and the time taken by the system is linear.

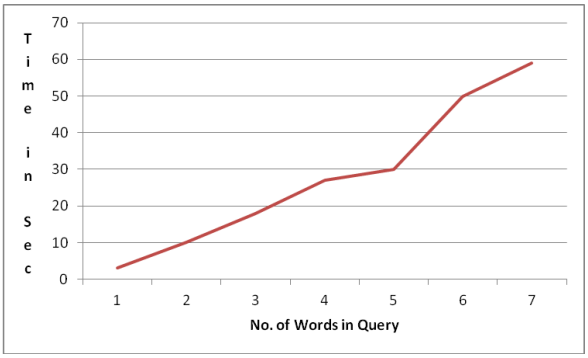

Figure 2. Performance

4.2 Usability

Usability results generated through a survey on Schneider Man's Principles based questionnaire (Shneiderman's, B. 2004). Questionnaire contains six questions with three options scaled to 0, 0.5 and 1. The following table shows results of questionnaire filled by ten participants, here column header should be read as, A: User ID, B: Layout Consistency, C: Easy to Understand, D: Organization of Actions, E: Error Prevention, F: Memorability, G: User Control, H: Informative Feedback, I: Use of Up to date Art.

A	B	C	D	E	F	G	H	I
1	1	1	1	1	0.5	1	0.5	1
2	1	1	1	1	1	0.5	0.5	0.5
3	1	0.5	1	1	0.5	0.5	1	0.5
4	1	0.5	1	1	1	0.5	0.5	1
5	1	1	0.5	0.5	0.5	0.5	0.5	1
6	0.5	0.5	1	1	0.5	0	0	0.5
7	1	1	1	1	0.5	0.5	1	0.5
8	1	1	1	1	1	1	1	1
9	0.5	0.5	0	0	0.5	1	0.5	0.5
10	1	1	0.5	0.5	1	0.5	0.5	1

Table 1. Usability Survey Results

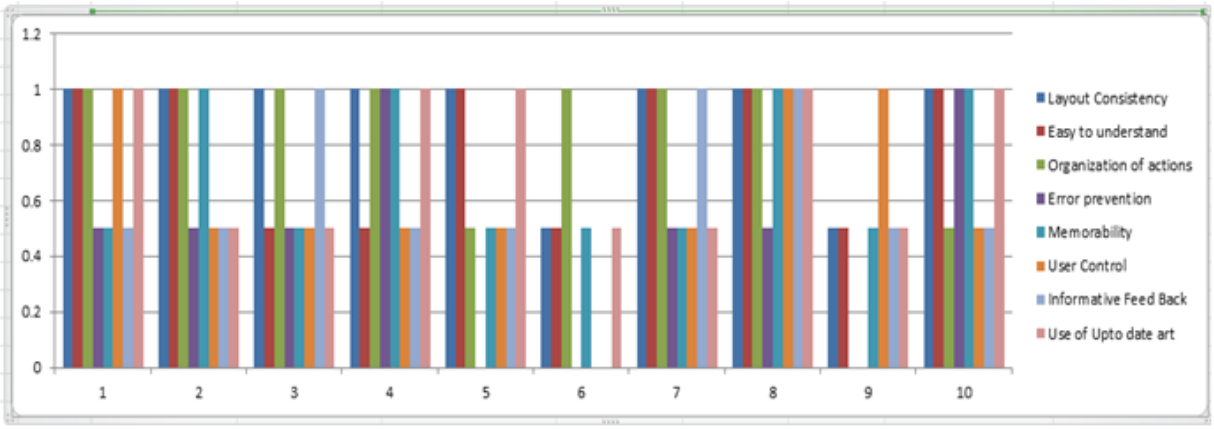

Figure 3. Usability Statistics

Results of this survey shows that users are more satisfied with first four indictors (B: Layout Consistency, C: Easy to Understand, D: Organization of Actions, E: Error Prevention) as compared to last four (F: Memorability, G: User Control, H: Informative Feedback, I: Use of Up to date Art). In future extensions specific consideration would be given to improve these factors. A usability graph generated from above table id shown in figure 3. This graph shows the variations of different questions by different participants.

5 Discussion

This work is a step to contribute in the field of text mining to process religious information particularly from Islamic Scriptures e.g. Quran and Hadith. In this paper searching of the Quran and Hadith with the options of synonyms matching and stemming the words has been introduced. For example, if a text contains 'happy' for five times and 'glad' for twenty times then user would be interested to view all these twenty five results, if s/he queries 'happy'. Also if search for the word 'pray' is generated then system will give results which contain the words like 'prays', 'praying' and 'prayed'. These two things make this work novel as current Quran and Hadith search engines do not offer such functionalities.

This research work can spawn several ideas and can provide a basis for further research on text mining in religious text. The extraction of information from Quran is a very vast task as it is believed that Quran and Hadith covers guidelines related to all aspects of human life including natural and related sciences beside the religious guidelines of morality and ethics. Hence this research of extracting information from Quran and Hadith has importance beyond religious boundaries.

6 Recommendations and Future Work

Some of the enhancements in current work include the better visualization of query results and the addition of more search options that can make user's experience more informative and rewarding. Such extensions might include, parametric, and chronological search support. Furthermore perspectives of different narrators and historians might be included and mined to provide different opinions of religious scholars on user's desired subject of search. Also current search can be extended from byword phrases to NLP query interface where the user can enter any query and system can detect the syntax as well as semantics of query in order to make the search results more relevant. In future this system can also be extended and multi-lingual support can be provided so that users can easily view and search results according in their native language. Another interesting future extension is to make a comparison among different Hadith and Verses to find out their mutual conflict or resemblances The Quran and the Hadith text can also be clustered among different hierarchies which can provide relations between concepts described in the text.

References

Alkhatib, M. 2010. *Classification of Al-Hadith Al-Sharif Using Data Mining Algorithm*, Proc. of European, Mediterranean & Middle Eastern Conference on Information Systems, Abu Dhabi.

Al-Kabi, M.and Al-Sinjilawi, S. 2007, *A comparative study of the efficiency of different measures to classify Arabic text*, University of Sharjah Journal of Pure and Applied Sciences, volume 4, No. 2.

Ali, S. R., Liu, W. M., Humedian, M. 2004. Islam 101: *Understanding the Religion and Therapy Im-*

plications. Professional Psychology: Research and Practice, Vol 35(6), 635-642.

Allah.pk. 2013. *The Multilingual Quran & Hadith Search Engine*, retrieved: 10-July-13.

Bihar Anjuman. 2013. *Bihar Anjuman – My RAH-BAR*, retrieved: 17-June-13.

intoIslam.com, 2013. *intoIslam - Islamic Search Engine*, retrieved: 10-July-13.

IslamiCity.com, 2013. *Islam & The Global Muslim eCommunity*, retrieved: 10-July-13.

Jbara, K. 2010. *Knowledge Discovery in Al-Hadith Using Text Classification Algorithm*, Journal of American Science, 6(11).

Porter, M. 1980. *An Algorithm for Sufix Stripping*, Program, Vol. 14(3), pp. 130-137.

SearchTruth.com, 2013. Search Engine: Search in the Quran الى قران الى كريم and Hadith, retrieved: 10-July-13.

Shneiderman, B. 2004. Designing for fun: how can we design user interfaces to be more fun? interactions 11(5), 48-50.

TextMiningTheQuran.com. 2013. *Wiki textmining-thequran*, retrieved: 17-June-2013.

TheIslamicSearch.come, 2010. *ISLAMIC SEARCH powered by GOOGLE - Islamic Search Engine*, retrieved: 10-July-13.

Convertonlinefree.com 2013. *Free online PDF to TEXT converter*, retrieved: 17-June-2013.

Appendix A: Questionnaire for Usability Survey

1. Is Layout, color scheme, capitalization, font and menus consistency achieved in designing the interfaces?

   □ Yes
   □ No
   □ Up to Some Extent

2. Interface design and navigations are easy to understand for the diversified set of users?

   □ Yes
   □ No
   □ Up to Some Extent

3. For searching from anything, the actions performed are in sequence and organized (e.g. for searching, first click on searching option then select searching type then input fields and do search)?

   □ Yes
   □ No
   □ Up to Some Extent

4. Is there any mechanism for preventing any type of error (syntax or semantic error) like drop down list, autocomplete, suggestions etc.?

   □ Yes
   □ No
   □ Up to Some Extent

5. Is User feeling easy to do new things that what he/she is not told?

   □ Yes
   □ No
   □ Up to Some Extent

6. Is User feeling easy in reversal of step i.e. going back to previous states easily?

   □ Yes
   □ No
   □ Up to Some Extent

7. Is User feeling that he/she can easily control the system?

   □ Yes
   □ No
   □ Up to Some Extent

8. Is the display is simple i.e. user can memorize the steps easily?

   □ Yes
   □ No
   □ Up to Some Extent

9. Is the system giving response for every action either it is modest or complex depends on action?

   □ Yes
   □ No
   □ Up to Some Extent

10. Any up-to-date art/technique is used in this system?

    □ Yes
    □ No
    □ Up to Some Extent

Appendix B: System Screen Shots

Figure 4. Main Screen

Figure 5. The Hadith Search Interface

Figure 6. The Quran Search Interface

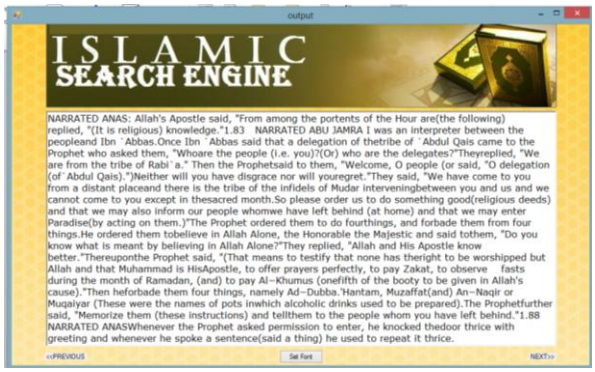

Figure 6. Search Results