# Semantic Network based Classifier of Holy Quran

Naseem Shahzadi
Barani Institute of Information Technology
3rd Floor, Umair Plaza, 6th road Chowk, Rawalpindi, Pakistan

Atta-ur-Rahman
Barani Institute of Information Technology
3rd Floor, Umair Plaza, 6th road Chowk, Rawalpindi, Pakistan

Mohammad Jamil Sawar
Barani Institute of Information Technology
3rd Floor, Umair Plaza, 6th road Chowk, Rawalpindi, Pakistan

## ABSTRACT
Automated Text Categorization (ATC) is a useful technology to build such software tool that can categorize a document to one of many predefined categories. Unfortunately there is no as such classifier for Holy Quran, one of the most important documents of the universe. This is because Holy Quran is written in Arabic and Arabic language processing is yet not that mature that language processing can be done. This paper aims on building a software tool that can categorize any verse of Holy Quran to one of predefined categories and if it is not falling in one of predefined categories it can automatically define a new category. New category will be added to database for future categorization process. Moreover, this categorization is not just based upon word count rather it is based upon word as well as meaning of the words in the verse and there correspondence in semantic network. The semantic network for this task is created and used in this classification. This tool will help those people who want to know the theme of a verse of Holy Quran.

## General Terms
Automatic Text Categorization, Semantic Networks

## Keywords
Automated Text Categorization, Semantic Network, Semantic Search

## 1. INTRODUCTION
Automated text categorization (ATC) [1] is one of key technologies used in natural language processing [2], where the unseen documents are categorized in one of the predefined categories. This helps in classification of electronic document. This concept is also coined as text mining mostly used in digital libraries, where most concise search is made for a user's request of a document.

Semantic Networks [3] are one of basics of knowledge representation [4]. It is a graphical representation of concepts. As a graph is consisted of vertices and edges or mathematically,

$$G = f(V, E) \quad (1)$$

where V is set of vertices and E is set of edges. In semantic network V is consisted of concepts and E is consisted of semantic relationships between those concepts.

A statistical classifier for the Holy Quran is proposed that classifies a verse in one of many categories based upon a linear classifier [5]. In this tool concept of Automated Text Categorization (ATC) [4] is used. This tool is a good approach to analyze the relevance of a Quranic verse based upon a score function but there were few limitations.

First, that classifier is for only two Surah that is Fatiha and Yaseen, second this classifier demands for a special corpus of entire Holy Quran which is not available to-date, third this classifier was based upon a word count not semantic relevance.

In this paper, we aim to address all three limitations in existing classifier. That is, this classifier is for entire Holy Quran not for just few Sura'h, and secondly there is no requirement of a special corpus. In fact, electronic translation of Holy Quran is used. In this regard we took two translations for sake of testing of the algorithm one is by Pickthall and other is by Yusuf Ali taken from reliable resources. In these translations, there followed a sequential index of each verse that is [Sura'h#: Verse#], that works like a composite key to address any verse of Holy Quran. Third limitation is addressed by a Semantic Network of Holy Quran. It is a customized Semantic Network based upon knowledge representation of domain experts. The Semantic Network is already proposed by the same authors [6].

Rest of the paper is organized as follows; section 2 contains the Semantic Network introduction, section 3 contains the classification process, section 4 contains algorithmic implementation of classifier and section 5 gives a comparison of classifier with another classifier while section 6 concludes the paper.

## 2. SEMANTIC NETWORK
This section describes the steps involved in making a semantic network, semantic search etc. There are two major aspects of this research. First is to make a semantic network and second is to build a classifier based upon the semantic network created in first step. Following diagram depicts the steps involved in creating a semantic network.

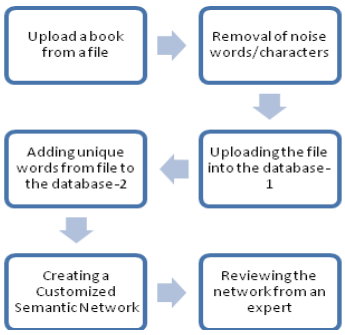


**Figure 1. Steps to create a semantic network**

Following is given the brief description of each step shown in flowchart.

## 2.1 Upload a book from file

Many translations of the Holy Quran are available by different authors, electronically. Most of them are given in MS-Word or text document format. The verses from different Sura'h are given in following style.

[Sura'h#:Verse# Verse translation]

So in this phase one can add a book translation of Holy Quran by any author given in above format by browsing option in the software. Upon uploading entire book will be saved in a database table. A snapshot of the file is given in the figure below.

```
001.001  In the name of God, Most Gracious,
001.002  Praise be to God, the Cherisher and
001.003  Most Gracious, Most Merciful;
001.004  Master of the Day of Judgment.
001.005  Thee do we worship, and Thine aid w
001.006  Show us the straight way,
```

Figure 2. Snapshot of file containing Quranic Translation

## 2.2 Removal of noise characters

In this step the extra characters like dots, commas, semi-colons etc are erased from the uploaded file by an intelligent parsing algorithm based upon string matching as well as pattern matching.

## 2.3 Uploading file into database-1

In this step file with erased characters in uploaded in database-1. This database contains all the sura'h and verses in a chronological order as it appears originally in the source file.

## 2.2 Adding unique words to database-2

In this step unique (distinct) words are picked from source file and added into a word database. Since multiple files may be incorporated in the database at the same time and there are a number of repeated words so this step just unify them.

## 2.3 Creating a Semantic network

User is given a facility to create his/her own customized semantic network. For example, in many translations there used word "Allah" while in another "God", similarly, "mosa" and "moses", "esa" and "Jesus" etc. So in this customized semantic network one can associate them using a link 'alias of' etc. Similarly, 'is a' is very useful link since most of the relationships can be expressed by using this link. In this way, one can create semantic network of any dimension using his/her own link descriptions and way to connect the concepts.

Figure 3 shows an example of semantic network being used where three concepts are inter-related with each other using various relationships.

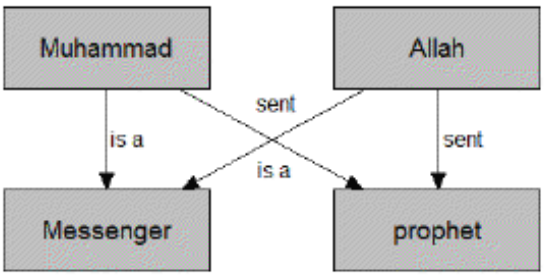


**Figure 3. Semantic network example**

## 2.4 Expert Review

Though there is given a customized approach in creation of semantic network, yet this is a religious semantic network, scholar has to consult an expert or a teacher to verify the semantic network. This is because when this tool will become operational then religious institutions can use it for research etc so authentication aspects are provided.

# 3. SEMANTIC NETWORK BASED CLASSIFIER

The classification procedure is depicted in the following flow diagram.

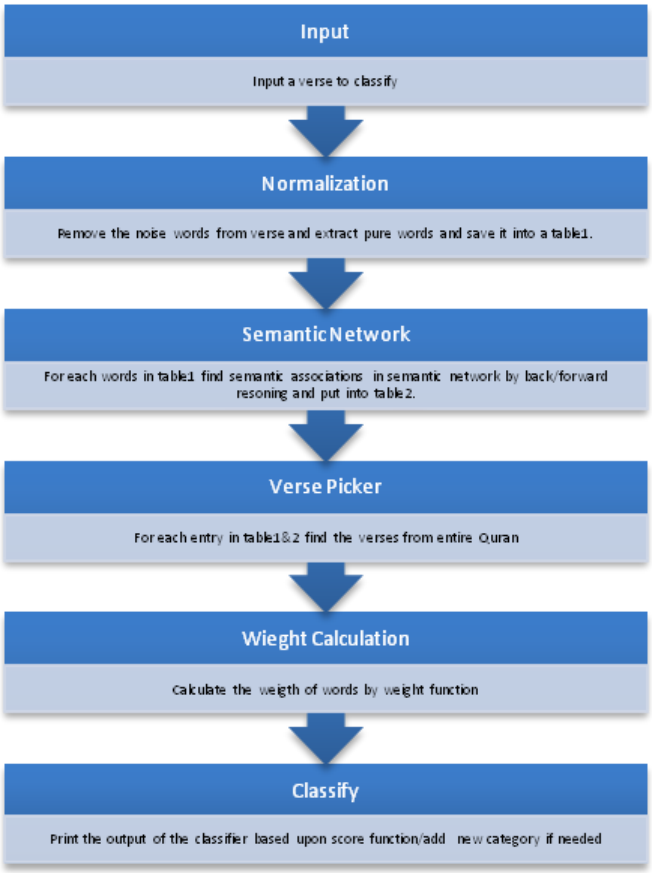


**Figure 4. Classification Flow Diagram**

The detail of the each step is given below.

## 3.1 Input a verse

A verse to be classified can be input by selecting the Sura'h number and then verse number respectively by using two combo boxes, when a Sura'h number is selected from first

combo box, according verse numbers are populated in second combo box.

## 3.2 Normalization

In this phase, all noise words and characters are removed and pure words tokens are obtained. The noise characters list is comprise of dots, commas and other punctuation symbols are eliminated also noise words like “in”, “the”, “then” etc are removed. Then pure words from the verse are added in a list1.

## 3.3 Semantic Search

For each words in previous list found after normalization phase, a semantic search is made based upon the Semantic Network already designed using the steps explained in section 2. After searching in entire semantic network all relevant words that may be alias or synonyms or any semantic association for the basic word tokens obtained in previous step, will be populated in list2. The union of list 1 and 2 is given a notion of feature set F.

$$F = L_1 \bigcup L_2 \qquad (2)$$

## 3.4 Verse Picker

For each word in list1 and 2 verses will be picked from entire Holy Quran and a verse list will be populated.

## 3.5 Weight Calculation

In this phase weight for all classes will be calculated by the weight function given below. Weight of class ‘i’ can be described as relative score of class ‘i’ in that feature set normalized by total score of all classes in that feature set.

$$W(C_i) = \frac{f_i}{f} \qquad (3)$$

fi=frequency of ith class for each word

f=frequency of all classes for each word

## 3.6 Classification

The verse will be classified by a maximum function applied upon individual weights. In this way, maximum weight class will be coined as “Class of the verse”. Mathematically this can be expressed as follows.

$$Sum(C_i) = \sum_{j=1}^{M} W(C_i); 1 \le i \le N \qquad (4)$$

Where M is total number of features extracted and N is total number of classes. So above equation can be stated as sum of all weights of class ‘i’ for entire feature set. Once having this we can easily classify the verse by the following equation.

$$C_{win} = Class(Max(Sum(C_i))); 1 \le i \le N \qquad (5)$$

Where $C_{win}$ is the winning class that can be expressed as the class that maximizes the sum over all feature set.

If all scores are less than a certain threshold a new class will be added to class table. Since if no class score greater than a threshold means there is no class in that class set. Following are given some screenshots of the application.

**Figure 5. Graphical User Interface of the Application**

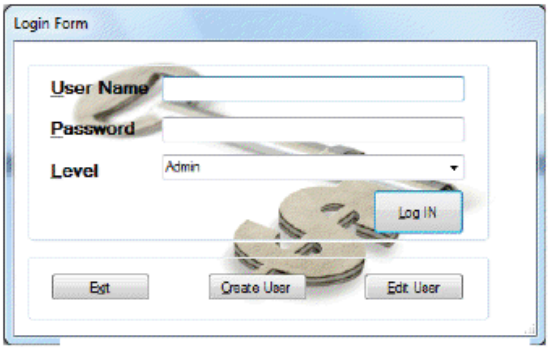


**Figure 6. Login Form**

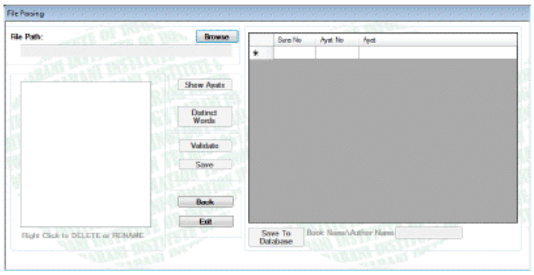

**Figure 7. File Parser and concept picker**

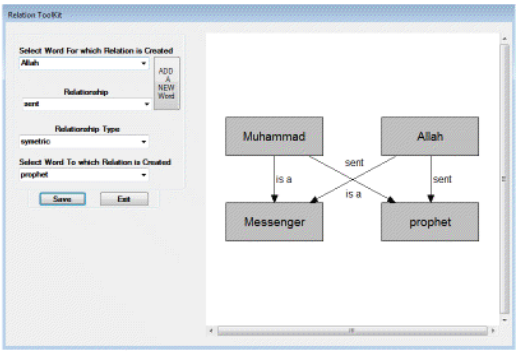


**Figure 8. Semantic Network Toolkit**

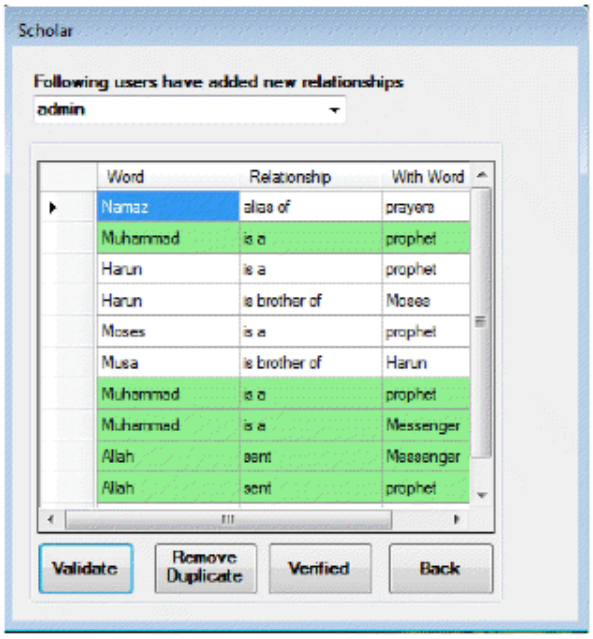


**Figure 9. Expert Verification**

Figure 5 contains the Graphical User Interface (GUI) of the application designed. It has a menu base interface rather that buttons etc. Menu can be seen at top left corner of the window.

Figure 6 shows the login form, where there are two types are user one is scholar, other is expert (admin). Once user is logged-in and create a semantic network or changes in existing semantic network means add new links new relationships, the updated semantic network will not be useful until expert verifies it. This is handled by a flag in database which is initially set to off. Once the verification is done the flag is turned on now the semantic network is useful in subsequent searches.

Only administrator is capable of creating new users (scholar) this is done for sake of an unauthorized use of the religious repository.

Figure 7 shows the interface for a file parser and concept picker. This provides a step by step mechanism. First of all a new file will be uploaded one have to enter the common file separator used by that translator.

In our examples Quanic translations contains a dot (.) separator while in King James Bible colon (:) is used as a common separator. Parser will trim off these separators and show aya't in next grid by clicking on "Show Grid" button. After that "Distinct Words" button will become active by clicking it, all distinct words in the uploaded file will be enlisted in left side pane. Also the noise words and characters will be eliminated like 'is', 'in', 'a' etc. Then lastly upon clicking the save button these unique words will be saved in the database.

Figure 8 shows relationship toolkit that helps in creating new dimensions of semantic network of any depth. New links can be created also old links can be repeated. Like in above example, "Muhammad is a messenger" and "Muhammad is a Prophet". Similarly, new links can be added like "Yaseen is an alias of Muhammad" etc. This can go arbitrarily long. In fact, as the customized semantic network (CSN) will grow our customized semantic search (CSS) will be even efficient, so this application provides an evolutionary mechanism to build a rich semantic network of the religious literature we mentioned above.

Figure 9 shows the verification process of new links added by the user. Duplicate links will be eliminated by "Remove Duplicate" button. Upon validation this process will be completed.

## 4. ALGORITHM AND IMPLEMENTATION

Table-1 contains the list of parameters used for experimentation.

**Table1. Parameters and testing material**

| Sr. # | Parameter Name | Description |
|---|---|---|
| 1 | M | Total number of features |
| 2 | N | Total number of classes |
| 3 | fc | Class Frequency |
| 4 | Wt | Weight |
| 5 | Max | Maximum Frequency |

The proposed algorithm is given in figure-10. The tools and technologies used for development of the software are;

- Microsoft Visual Studio .Net 2008
- Microsoft SQL Server 2005

## 5. COMPARISON

The comparison of proposed classifier with the existing classifier [5] can be given as follows. The comparison is based upon certain characteristics like domain, dependencies, searching criteria, feature set and cardinality of class set. Comparison shows that proposed classifier is more accurate, versatile and robust compare to the existing one.

**Table-2 Parameters and testing material**

| Characteristic | Statistical Classifier | Proposed Classifier |
|---|---|---|
| Domain | 2 Sura'h (Fatiha and Yaseen) | Entire Holy Quran |
| Dependency | A Specific Corpus | No dependency |
| Search Criteria | Word matching | Word + Semantic Search |
| Feature set | Limited to key words only | Vast due to keyword and semantics |
| Class set | Fixed | New classes can be added manually as well as automatically |

```
Algorithm Classify (Input: Surah number, verse number, output: Category)
Begin
Step1: Choose Sura'h and Verse number
Step2: Normalize the verse and obtain word tokens
Step3: For each word token Do
        Begin
Step3.1: Find all semantic words (associations) from Semantic Network
Step3.2: Find all verses that contains words from both token words and semantics (features)
Step3.3: For each verses obtained in step3.2 Do
          Begin
    Step3.3.1:  Find all classes (themes) related to that verse
End
fc=sum(all frequencies for all features)
Step3.4: For each class in class_set Do
Begin
   Step3.4.1:Wt[feature,class]=frequency of class/fc
End
Step4: For class 1 to N Do
        Begin
          For feature 1 to M Do
          Begin
          Step4.1:Sum[class]= Sum[class]+Wt[feature,class]
          End
        End
Step5: max=Sum[1]
For class 2 to N Do
Begin
  If(Sum[class]>Max)
          Max=Sum[class]
End
Step6: If(Max<Threshold)
          Add new class named most frequent feature
        Return Class with Max sum
End
```

**Figure 10. Proposed Classification Algorithm**

# 6. CONCLUSIONS

In this paper a Semantic Network based Statistical classifier of Holy Quran is proposed, designed and tested as a product. A Customized Semantic Network is proposed and used that is based upon different Quanic translations. In this way a meaningful classification of any verse of Holy Quarn is made possible. Existing work was limited due to certain restrictions of language etc while the proposed scheme is not. This application is useful for Islamic scholars who can use it for referencing and indexing in religious literature, as well as for a common person.

# 7. ACKNOWLEDGEMENTS

This research work is supported by Higher Education Commission (HEC) of Pakistan and Barani Institute of Information Technology (BIIT) Pakistan.

# 8. REFERENCES

[1] Fabrizio Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47, 2002.

[2] Bates, M. (1995). Models of natural language understanding. Proceedings of the National Academy of Sciences of the United States of America, Vol. 92, No. 22 (Oct. 24, 1995), pp. 9977–9982.

[3] Allen, J. and A. Frisch (1982). "What's in a Semantic Network". In: *Proceedings of the 20th. annual meeting of ACL*, Toronto, pp. 19-27.

[4] John F. Sowa, Alexander Borgida (1991). *Principles of Semantic Networks: Explorations in the Representation of Knowledge.*

[5] Mohammed Naji Al-Kabi, Riyad Al-Shalabi, "Statistical Classifier of Holy Quran Verses (Fatiha and Yaseen Chapters)", *Journal of Applied Sciences* 5(3):580-583, 2005.

[6] Naseem Shahzadi, Atta-ur-rahman, Adil Shaheen, "Semantic Network based Semantic Search of Religious Repository", *International Journal of Computer Applications*. Vol. (36), No. 9, December 2011.