1-1-2014

MQVC: Measuring Quranic Verses Similarity and Sura Classification Using N-Gram

Mohammed Akour
Yarmouk University

Izzat Alsmadi
Boise State University

Iyad Alazzam
Yarmouk University

Publication Information

Akour, Mohammed; Alsmadi, Izzat; and Alazzam, Iyad. (2014). "MQVC: Measuring Quranic Verses Similarity and Sura Classification Using N-Gram". *WSEAS Transactions on Computers, 13*, 485-491.

MQVC: Measuring Quranic Verses Similarity and Sura Classification Using N-gram

Mohammed Akour
CIS Department
Yarmouk University
Irbid-Jordan
Mohammed.akour@yu.edu.jo

Izzat Alsmadi
Computer Science Department
Boise State University
USA
alsmadi@gmail.com

Iyad Alazzam
CIS Department
Yarmouk University
Irbid-Jordan
eyadh@yu.edu.jo

Abstract: Extensive research efforts in the area of Information Retrieval were concentrated on developing retrieval systems related to Arabic language for the different natural language and information retrieval methodologies. However, little effort was conducted in those areas for knowledge extraction from the Holly Muslim book, the Quran. In this paper, we present an approach (MQVC) for retrieving the most similar verses in comparison with a user input verse as a query. To demonstrate the accuracy of our approach, we performed a set of experiments and compared the results with an evaluation from a Quran Specialist who manually identified all relevant chapters and verses to the targeted verse in our study. The MQVC approach was applied to 70 out of 114 Quran chapters. We picked 40 verses randomly and calculated the precision to evaluate the accuracy of our approach. We utilized N-gram to extend the work by performing experiment with machine learning algorithm (LibSVM classifier in Weka), to classify Quran chapters based on the most common scholars classification: Makki and Madani chapters.

Key-Words: - Text Classification, Quranic Verses Similarity, Madani and Makki Chapters classification.

1 Introduction

Evaluating similarities in documents is widely used for applications related to information retrieval, natural language processing, etc. Examples of applications in which there is a need to sort, cluster or classify documents based on the amount or level similarity includes: Databases, web, and search engines indexing, documents automatic clustering and classifications, and so on. The easiest way for illustrating text as concept/term vectors is bunch of words technique, where each document or article is represented as a set of vectors. Each vector composed of boolean value for each word reveals in the document/article. For each word appears in the document, its associated boolean value will be assigned 1 value otherwise it will be assigned 0 value.

In this paper we develop an approach that used Term Frequency, Inverse Document Frequency (TF-IDF) measures to induce similar verses in Holy Quran based on vector similarity. Moreover, Normalization, Stemming and Stop Word removing were used to improve the retrieval of relevant chapter and verses. Our approach referred to as Automated Quranic Verses similarity measurement (MQVC). To demonstrate MQVC and to measure the accuracy of the MQVC, we consider and compare the work of a Quran Specialist who manually identified all relevant Chapters and verses to the targeted verse in our study. The experiments show how MQVC is highly accurate for retrieving relevant verses. The Holly Quran chapters are described as either Makki or Madani. Classifying the chapters is based upon the exposure of the majority of the chapter's verses before or after the Hijra [1], 86 of them Makki and 28 Madani. We randomly picked 70 of the 114 chapters in our study.

In the second part of our experiments, we employed N-gram to extend the work by performing experiment with machine learning algorithm (LibSVM classifier in Weka), to classify Quran chapters based on the most common scholars classification Makki and Madani chapters.

The rest of the paper is organized as follows: related works are discussed in section 2. In section 3, methodology is described. Section 4 experiment results are explained. Subsection 4.1 presents detailed description of MQVC clustering and classification experiments based on N-gram, and finally conclusions and future works are given in section 5.

2 RELATED WORKS

In this section, we will list a subset of examples of documents clustering or classification specifically for Arabic language. A special focus is given to research papers that evaluated similarity, clustering or classification for the Nobel Quran. Text categorization or text classification is the process of allocating a text to single or more groups of predetermined listing [2]. A broad diversity of approaches and techniques have been proposed, designed and implemented in the area of text classification such as Bayesian Categorization [2] K Nearest Neighbor Categorization [3], Decision Tree Categorization [2], Rule Based Categorization, Support Vector Machines [4] and Neural Networks [5].

Karamcheti [6] performs a comparison between k-Nearest Neighbor methodology and Naive Bayes through implementing two classification engines for text classification for both. They evaluate the effectiveness based on the standard recall and precision for a group of documents. The results show that the classification engine of k-Nearest Neighbor is better than classification engine of Naïve Bayes engine. The work by Lee [7] illustrates the improvement of semi-supervised and supervised learning methods to similarity based text classification systems. Supervised methods to text classification need huge amount of training documents in order to increase the performance and effectiveness. [7] Proposes new approach in text categorization, it decreases the number of training documents in order to obtain the same level of effectiveness. A model called Hierarchical Mixtures of Experts (HME) used for text classification [8]. The model employs the principle of divides and conquers in building and training machine learning algorithms. The model is assessed through using linear classifier and neural networks.

Abdul-Baquee and Eric presented a large corpus: QurAna that was created from the original Holly Quranic text [9]. The corpus is annotated with antecedent references of pronouns. Authors described in detail the annotation scheme and process. They used and measured verse distance using Vector Space Model, and considered each verse of the Quran as a separate document. Abdul Baquee developed an online tool that is used to find similar Quran verses based on TF-IDF for term weighting and vector similarity measurement [10]. Author removed stop words where the total of 33,931 words in Quran was collected. Dost and Ahmed [11] addressed the constructional characteristics of the Suras and Ayats classification in the Holy Quran based on location of revelation: Makki or Madani. They provided a probabilistic approach to study Makki, Madani and Mixed Suras of the Holy Quran. Their approach was based upon the word-size and word-length of Ayats.

Statistics of Holy Quran verses were studied and explained in the literature, such as number of suras and ayat, the ayat-size by words, and ayat-size by letters etc. This gives a comprehensive insight into the structure of ayats. Al-Dargazelli [12] had identified numerical patterns of the Holy Quran based on verses; he has only used basic descriptive statistics and computed frequency tables of verses. He has given number of verses for each Sura and identified the Suras by Makki and Madani.

Akour et al [13] presented an approach and a new question answering system (QArabPro) for reading comprehension texts in Arabic. In order to implement the QA system they developed an IR system to search and retrieve relevant documents. The IR system was utilized Salton's statistical VSM.

3 Holy Quran Verses Corpus pre-processing and Methodology

Before building the Quran corpus and finding the similarity among verses the whole extracted chapters and verses is filtered. Arabic function words were detached (i.e. prefixes, suffixes, pronouns, and prepositions) and then our simple stemmer extracted the root of the remaining texts. The holy Quran is composed of 114 chapters. Each chapter consists of a number of verses. The total number of verses in the whole Quran is 6,214 verses.

A verse includes words with or without frequencies. The chapter is a set of verses that allows duplicates of the same word. The general assumption is that, frequent terms in a chapter are more important and shows some major subject. We use Salton's equations [14] to measure Holy Quran terms frequency. The weight of each keyword is calculated as follows:

Term Weight = wi = tfi * log (D/dfi) , Where: tfi represents the term frequency (i.e. term counts) or number of times a term i occurs in a verse. dfi = verse frequency or number of verses containing the term i, and D = total number of verses in the corpus

The dfi/D ratio is the probability of selecting a chapter, then a verse that contains the queried verse from the whole chapters. This perspective reveals a global probability over the entire corpus. Thus, the log(D/dfi) term is the inverse verse frequency, IDFi accounts for global information.

A set of experiments ran to determine the accuracy of our technique in retrieving the most accurate verses.. Figure 1 provides an overview of the MQVC approach. The bolded processes in Figure 1 represent the second part of our experiments, detailed description will provided later in subsection 4.1. After removing stop words and performing the normalization, our approach extracts the key terms and feeds MQVC builder with them to build the corpus. We developed a GUI that helps users to type the entire required verse or even partial to retrieve the most similar one's along with the percentage. We used Precision as an evaluation measure.

4 Results and Discussion

In this section we present the result of the first part of MQVC experiments. We compared the retrieved verses with a manual third party oracle, i.e., we consider and compare the work of third party who manually identified similar verses. We calculated the precision for each verse that targeted in the experiment, and then we calculate the mean precision. Although table 1 shows how MQVC approach provides good accurate precision measures, measuring similarity based on the key terms leads to missing several similar verses as these verses contain similar concepts and knowledge not only key terms.

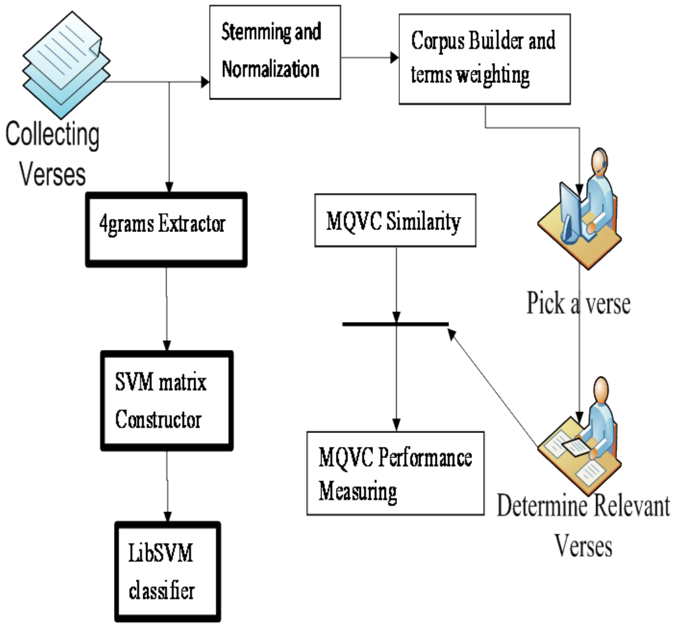

Collecting Verses
Stemming and Normalization
Corpus Builder and terms weighting
4grams Extractor
MQVC Similarity
Pick a verse
SVM matrix Constructor
MQVC Performance Measuring
Determine Relevant Verses
LibSVM classifier

Figure1. MQVC Approach

Table 1. Mean precision of the MQVC approach

Chapter	Verse	Precision	Chapter	Verse	Precision
2	1	94%	22	57	88%
3	1	90%	23	21	87%
4	68	84%	24	18	88%
5	68	80%	25	50	88%
6	4	85%	30	14	87%
7	42	82%	31	21	85%
8	20	83%	32	25	95%
9	55	93%	33	45	90%
10	17	85%	34	20	83%
11	67	81%	41	68	85%
12	57	83%	42	31	95%
13	9	80%	43	25	86%
14	10	80%	51	20	85%
15	11	86%	52	39	84%
16	15	83%	55	1	82%
17	48	95%	56	32	85%
18	67	89%	63	2	80%
19	54	93%	67	23	90%
20	17	88%	68	30	89%
21	41	92%	69	2	90%

4.1 Clustering and Classification based on N-Gram

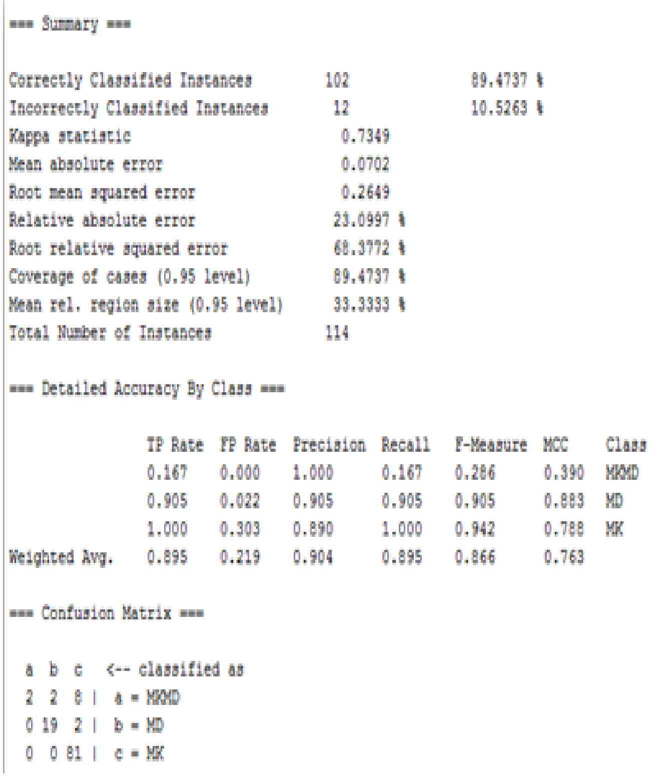
In the first part of MQVC experiments we measure and evaluate verses similarity in Holy Quran based on VSM. In this section we utilized N-gram to extend the work by performing an experiment with a machine learning algorithm (LibSVM classifier in Weka), to classify Quran chapters based on the most common scholars classification Makki and Madani chapters. SVM is known to be a very good classification algorithm. Features selected represent the most popular ngrams based on their repetition in Quran chapters and verses.
In this section, we describe an approach to extract N-gram terms from the holly Quran.

To provide an evidence of the usefulness of such process, we investigate the Issue of classification Sura's into Madani, Makki or Both – classified according to religion scientists. The methodology of extracting grams and how we get benefit of utilizing N-gram in distinguishing Madani, Makki , or both Sura's is summarized as follows:

1. Removing all stop words from Quran Corpus.
2. As a first step we consider 4-grams as an example based on random selection. In the future work we will include using other gram sizes as well as using popular terms.
3. Extract all 4grams from Quran complete content.
4. Select the top 1000 4grams. This selection is based on the number of repetition of those 4grams in the complete Quran corpus. Top 1000 grams based on their frequency are selected.
5. Construct an SVM matrix, where columns represent top 1000 4-grams, and rows represent Quran Sura's. Values in rows represent frequency of the particular 4gram in the particular Sura.
6. Add a class column to label each Sura into: Makki (MK), Madani (MD) or Both (MKMD). This information is collected from Islamic websites based on religious knowledge.
7. Use LibSVM classifier (available in WEKA 3.7 [15] to evaluate prediction accuracy and get prediction metrics.
8. The result is compared with study was done by Sharaf [9, 10].

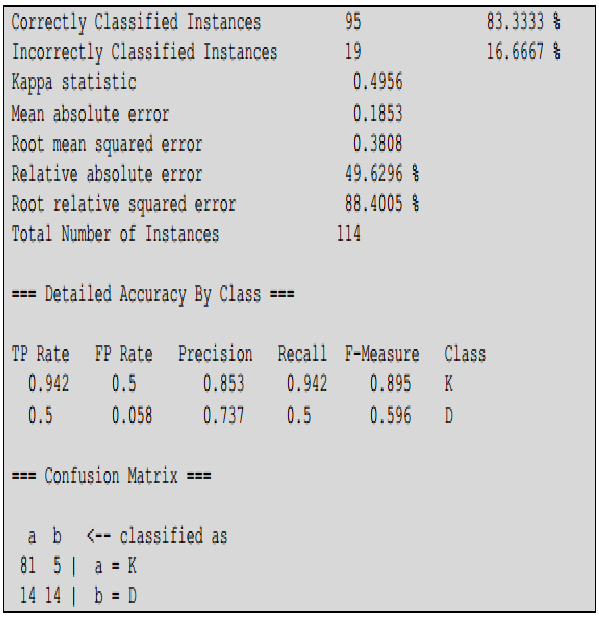
Figure 2 shows summary of the classifier results. In the work of Sharaf [9, 10], picking the minimal, meaningful, and critical differentiating attributes was the main challenge. In order to cope with this challenge, he proposes and discusses the selection and the classification of 14 attributes. From the complexity perspectives, Generating N-gram attributes in order to classify Quran Sura's and in comparison with Sharaf [9, 10] procedures for choosing the most differentiating attributes was easier. Several machine learning research works utilized TP, TN, FP, FN measures to calculate and evaluate the accuracy of their works [9, 10, 16, 17, 18].

From the accuracy perspective, in comparison with Sharaf [9, 10] results, it can be clearly seen that our classification outperforms earlier one (our MK main prediction metrics: TP: 1, FP 0.303, while theirs TP: 0.942, FP 0.5). The good performance requires TP to be as closer to 1 as possible and FP as much close to zero as possible. MD metrics: Ours: TP: 0.905, FP: 0.022, theirs: TP: 0.5 FP: 0.058. In addition, in our study we classify the SuraSura's based on MKMD as well while they did not. Moreover, in the terms of recall and precision, our methodology outperforms their as it achieved the highest recall in both MD and Mk classes. Figure 3 shows Sharaf [8,9] Weka classifier results.



```
=== Summary ===

Correctly Classified Instances         102               89.4737 %
Incorrectly Classified Instances        12               10.5263 %
Kappa statistic                          0.7349
Mean absolute error                      0.0702
Root mean squared error                  0.2649
Relative absolute error                 23.0997 %
Root relative squared error             68.3772 %
Coverage of cases (0.95 level)          89.4737 %
Mean rel. region size (0.95 level)      33.3333 %
Total Number of Instances              114

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC    Class
                 0.167    0.000    1.000      0.167    0.286      0.390  MKMD
                 0.905    0.022    0.905      0.905    0.905      0.883  MD
                 1.000    0.303    0.890      1.000    0.942      0.788  MK
Weighted Avg.    0.895    0.219    0.904      0.895    0.866      0.763

=== Confusion Matrix ===

  a  b  c   <-- classified as
  2  2  8 |  a = MKMD
  0 19  2 |  b = MD
  0  0 81 |  c = MK
```

Figure 2. Weka Classifier Results



```
Correctly Classified Instances          95               83.3333 %
Incorrectly Classified Instances        19               16.6667 %
Kappa statistic                          0.4956
Mean absolute error                      0.1853
Root mean squared error                  0.3808
Relative absolute error                 49.6296 %
Root relative squared error             88.4005 %
Total Number of Instances              114

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.942     0.5       0.853      0.942     0.895    K
  0.5       0.058     0.737      0.5       0.596    D

=== Confusion Matrix ===

  a  b   <-- classified as
 81  5 |  a = K
 14 14 |  b = D
```

Figure 3. Sharaf [9, 10] Weka Classifier Results

5 Conclusions and Future Work

We focus in this study on evaluating similarity of text and documents in Arabic Language. In particular we used text of the holy book of Muslims, The Quran, as the subject of the experiment. Verses of the Quran are used as the queries to search for and evaluate similarity. More than 2000 different verses from the Quran are used in this experiment. The Quran contains 114 chapters and total of 6236 verses in all chapters. For each verse used as a query, results retrieved most similar verses based on the algorithms evaluated the similarity in percentage between query verse and retrieved verse and also overall precision.

We extend the work by employing N-gram and performing experiment with machine learning algorithm (LibSVM classifier in Weka), to classify Quran chapters based on the most common scholars classification Makki and Madani chapters. As a first step we consider 4-grams as an example based on random selection. We did a comparison with very relevant papers Sharaf [9, 10] in terms of ROC metrics. N-gram algorithm is widely used in other areas. Most feature selection methods for

Quran or relevant research focus on words for example top frequent words. We showed that Top N-grams can be more accurate in predicting the class although those n-grams do not necessary show complete or meaningful words.

In the future work we could use other gram sizes as well as using popular terms. The experiment revealed how using machine learning can play significant role in classifying Makki and Madani Sura's. Studying classification and classes' prediction is very popular in research in NLP or relevant fields. Our comparison focused on the very specific papers that we found used complex feature selection and classification algorithms. We showed that using the very simple N-gram approach mixed with SVM can produce comparable and better results in terms of ROC metrics without a significant effort. As future work we intend to employ N-gram and machine learning to retrieve similar verses and classifying verses based on related topics.

6 References

[1] Al-Hilali, M., Khan, M., 1997. "Translation of the Meanings of The Noble Qur`an in the English Language" King Fajd Complex for the Printing of the Holy Qur`an, Madinah, K.S.A

[2] Lewis, D.D. and Ringuette, M.. A comparison of two learning algorithms for text categorization. In Third Annual Symposium on Document Analysis and Information Retrieval, 81-93, 1994.

[3]Yang, Y. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 13-22, 1994.

[4]Vapnik, V., The Nature of Statistical Learning Theory ,Springer-Verlag, 1995.

[5]Wiener E., Pedersen, J.O. and Weigend, A.S. A neural network approach to topic spotting. In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), 1995.

[6] Karamcheti, A. (May 2010). A Comparative Study on Text Categorization. M.Sc Thesis, University of Nevada, Las Vegas.

[7]Lee, K. (September 2003). Text Categorization with a Small Number of Labeled Training Examples. PhD Thesis, School of Information Technologies, University of Sydney, Australia.

[8] Ruiz, M. (December 2001). Combining Machine Learning and Hierarchical Structures for Text Categorization. PhD Thesis, Computer Science Dept., University of Iowa, Iowa City, Iowa, USA.

[9] Sharaf, Abdul-Baquee and Atwell, Eric, (2012) "QurAna: corpus of the Quran annotated with pronominal anaphora", LREC 2012

[10] Sharaf, Abdul-Baquee and Atwell, Eric. (2012) "QurSim: A corpus for evaluation of relatedness in short texts", LREC 2012. [muhammad et al - 2012b- qurSim]

[11] Dost, M., Ahmad, M., Statistical Profile of Holy Quran and Symmetry of Makki and Madni SurrasPakistan, Journal of Commerce and Social Sciences Vol.1 2008.

[14] Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM., 18: 613-620.

[15] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[12] Al-Dargazelli,S. (2004). A Statistical Studies of Holy Quran. [Online] Available: http//www.quranicstudies.com/printout104.html

[13]M. Akour, S. Abufardeh, K. Magel, Q. Al-Radaideh. (2011) QArabPro: A Rule Based Question Answering System for Reading Comprehension Tests in Arabic. American Journal of Applied Science, June 2011, ISSN: 1546-9239 e-ISSN: 1554-3641.

[16]Schlemmer, A; Zwirnmann, H.; Zabel, M.; Parlitz, U.; Luther, S., "Evaluation of machine learning methods for the long-term prediction of cardiac diseases," Cardiovascular Oscillations (ESGCO), 2014 8th Conference of the European Study Group on , vol., no., pp.157,158, 25-28 May 2014

[17] Szénási, S., "Distributed Region Growing Algorithm for Medical Image Segmentation", International Journal of Circuits, Systems and Signal Processing, 2014, Vol. 8, No. 1, pp.173-181, ISSN 1998-4464

[18] Hemalatha, N.; Rajesh, M.K.; Narayanan, N.K., "A machine learning approach for detecting MAP kinase in the genome of Oryza sativa L. ssp. indica," Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on , vol., no., pp.1,6, 21-24 May 2014