Classification of Hadiths using LVQ based on VSM Considering Words Order

Mohamed Ghanem
Department of
Computer Science,
Faculty of Sciences,
Ibn Tofail University,
Kenitra, Morocco

Abdelaaziz Mouloudi
Department of
Computer Science,
Faculty of Sciences,
Ibn Tofail University,
Kenitra, Morocco

Mohammed Mourchid
Department of
Computer Science,
Faculty of Sciences,
Ibn Tofail University,
Kenitra, Morocco

ABSTRACT

The religion of Islam is based on a sacred text called Qur'an, a divine speech expressed in Arabic language. Qur'an constitutes the main root of Islam jurisprudence which has a second source of inspiration known as Hadiths. As the Muslim's life is governed by those holy texts, need of their authenticity is required. Using VSM (Vector Space Model), we can represent Hadiths as a vector of words. The Term Weighting obtained by multiplying term frequency by the inverse document frequency does not take into account the word order, however, order of narrators is critical to classify Hadith. In this paper we propose a new method considering the words order (in our case the narrator's order), to classify Hadiths into four categories: Sahih, Hasan, Da'if and Maudu'. We use in this purpose LVQ (Learning Vector Quantization). We got good results for classifying Sahih and Maudu' categories.

General Terms

Hadith categorization, Algorithms.

Keywords

Arabic Natural Language Processing, Learning vector quantization, Term Weighting, Text categorization, Vector Space Model.

1. INTRODUCTION

The behavior of the prophet of Islam (acts or saying) is known as Hadiths. The chain of narration is called Sanad and the so called Matn is the content.

Islamic thought builds its jurisprudence on Hadith's collection after the Qur'an recommendations. Originally, writing narration of Hadith was forbidden due to some political and religious reasons. But after the death of the prophet, fearing that some of hadiths are being lost, Umar Ibn Abd al-Aziz (2 November 682– 31 January 720) initiated the writing project of Hadith to guaranty integrity and uniformity of the text. Therefore, some irregularities were noticed and necessity of thorough authentication study of narration is mandatory before its use in jurisprudence. Each Narrator of the Sanad indicates the person from whom he heard the Hadith, and hierarchically, we got the originator of the Matn. That's why, clerics and jurists take into consideration narrator's order to classify Hadiths. First narrators were prophet companions, whom understood and know both the generality and the particularity of Hadiths. They forward what they learned from the prophet to those after them as they were commanded. Then the generation after them, the followers, received it and then conveyed it to those after them, and so on.

Classification of text in categories is an important issue which attracts many researchers in machine learning, statistics and information retrieval techniques.

There are several methods used to classify text according to a set of one or more preexisting categories, such as Support Vector Machine (SVM), K Nearest Neighbor (KNN), Artificial Neural Networks (ANN), Nave Bayes Classifier, and Decision Trees.

EL Kordi M., et al. [1] used Naive Bayes to classify Arabic web documents.

Mesleh A., [2] classified Arabic language articles using support vector machines.

Sawaf H., et al. [3] used statistical methods to classify political, cultural and economical Arabic documents. Their best accuracy was 62.70%.

Unfortunately; little work has been practiced on electronic books of sacred texts.

Martín-Valdivia M.T., et al. [4] proposed a categorization system based on neural learning, using the polyglot Bible.

Harrag F., and El-Qawasmah E., [5] proposed the application of Artificial Neural Network for the classification of Hadith into categories like faith, judgments and worships.

Kashif B., et al. have created the system: Muhadith [6], to classify hadiths into three categories: Marfu' "elevated", Mauquf "stopped" and Maqtu'- "severed".

To our knowledge there are no systems for automatic classification of Hadith into the four main categories: Sahih (Authentic), Hasan (Good), Da'if (Weak) and Maudu'(faked); which is the purpose of this paper.

As defined in [7] the four categories mean:

1) Sahih : means to be actual.
2) Hasan : whose persons who carry its narration have been reported to be less pious than persons reporting Hadith Sahih.
3) Da'if (The Weak): weakness is either pragmatic or abstract, it mean here the abstract weakness.
4) Maudu' (faked, forged): a text that goes against the established norms of the Prophet's sayings, or its reporters may contain falseness.

Sahih and Hasan are practised by Muslims. However, Da'if is inadmissible but can be practised if it does not contradict with the Islamic doctrines. Maudu' is considered unacceptable in any condition.

The closest research to the goal of this paper is that of Hernández-Reyes E., et al. [8]. To take into account the word order, they proposed a document representation in which each document is represented as a set of its maximal frequent sequences, and they compared the result of the clustering based on their method with those obtained with the vector space model. But we cannot compare our results, because in our case each term must appear only once in each text, hence the term frequency is always equal to 1.

2. VSM AND LVQ

In order to cluster a document collection, we first need to represent them in a suitable way to be compared. Many researchers use the Vector Space Model (VSM) which is is considered an effective model in information retrieval [9] proposed by Salton in 1975 as information representation scheme [10]. This model represents the documents as a word vector in which the features correspond to the words of the documents. A weight assigned to a term represents the relative importance of that term.

(Salton and McGill, 1983) [9] proposed an approach for term weighting; they calculated the frequency of occurrence of words to represent the vectors of texts.

They used the standard *tf* × *idf* equation, where *tf* is the frequency of the term in the document, and *idf* is the inverse document frequency defined as:

$$\mathrm{idf}_{\mathrm{i}} = \log_2\left(\frac{M}{\mathrm{df}_{\mathrm{i}}}\right)$$

Where df_i (document frequency) is the number of documents in the collection in which term *i* occurs, and *M* is the total number of documents.

If a term *t* exists in all documents:

$$idf_t = \log_2\left(\frac{M}{M}\right) = 0$$

This means that this term is not useful for differentiating a document from another one.

If it exists just in one document, its *idf* will have the maximum value:

$$idf_t = \log_2\left(\frac{M}{1}\right) = \log_2(M)$$

Thus, the weight is calculated by the following equation:

$$w_{ij} = tf_{ij} \times idf_i$$

Where tf_{ij} (term frequency) is the number of occurrences of term *i* in document *j*.

To represent categories by term weight vectors, the VSM is applied for each category using its representative document.

Using the Euclidean distance, the similarity between document *j* and category *k* is:

$$sim(d_j, c_k) = \left\| d_j - c_k \right\| = \sqrt{\sum_{i=1}^{N} (w_{ij} - c_{ik})^2}$$

Where *N* is the number of terms in the whole collection, W_{ij} is the weight of term *i* in document *j* and C_{ik} is the weight of term *i* in category *k.*

In this research we used Learning Vector Quantization (LVQ), a classification method that uses a competitive supervised learning algorithm [11]. In LVQ, there are only two layers [12], the output one has neurons equal to the number of classes.

3. PROPOSED METHOD

The data set used in this work contains a collection of 160 Hadiths already classified by experts of the Hadith sciences.

As shown in the table below, the number of training and testing set is uniform for all categories.

Table 1. Division of the corpus into training and test sets

Corpus Categories	Training corpus	Test corpus
Sahih	30	10
Hasan	30	10
Da’if	30	10
Maudu’	30	10
TOTAL	120	40
Percentage	75%	25%

3.1 Preprocessing phase

Before starting learning or classification, the dataset must be processed in order to remove the non useful information. For that the operations below are performed :

1. ***Removing Matn***: this process is done manually and aims to remove Matn, which is the second part of Hadith (the text).
2. ***Removing verbs*** (manually): like حدثنا, أخبرني (tell us)...
3. ***Removing the word*** “عن” (from).
4. ***Names standardization***: As shown in Table 2.

Table 2. Names standardization

Forms of the same name	Selected form
أنس بن مالك (Anas the son of Malik)	أنس بن مالك
أنس (Anas)	
ابي أيوب	أبو أيوب
أبو أيوب	
أبو سعيد الاشج	أبو سعيد الاشج
أبو سعيد عبد الله بن سعيد الاشج	
سمعان ابن مهدي	سمعان
سمعان	

An example of the preprocessing is shown in the table below.

Table 3. Preprocessing steps of a given Hadith

Step	Result of the step
The full hadith	حدثنا زهير بن حرب. حدثنا جرير، عن سهيل، عن عبد الله بن دينار، عن أبي صالح، عن أبي هريرة؛ قال: قال رسول الله صلى الله عليه وسلم: "الإيمان بضع وستون شعبة. فأفضلها قول لا إله إلا الله. وأدناها إماطة الأذى عن الطريق. والحياء شعبة من الإيمان". Tell us Zuhair Bin Harb, told us Greer, from Suhail, of Abdullah ibn Dinar from Abu Salih, from Abu Huraira; said: The Messenger (peace be upon him) said: “Faith has sixty-odd branches, the best of it is saying: “there is no god but Allah”, and the lowest of it is removing something harmful from the road, and modesty is a branch of faith”
Removing Matn	حدثنا زهير بن حرب. حدثنا جرير، عن سهيل، عن عبد الله بن دينار، عن أبي صالح، عن أبي هريرة؛ قال: قال رسول الله صلى الله عليه وسلم
Removing verbs	زهير بن حرب. جرير، عن سهيل، عن عبد الله بن دينار، عن أبي صالح، عن أبي هريرة؛ رسول الله صلى الله عليه وسلم
Removing the “عن” (from)	{زهير بن حرب، جرير، سهيل، عبد الله بن دينار، أبي صالح، أبي هريرة، رسول الله صلى الله عليه وسلم} {Zuhair Bin Harb, Greer, Suhail, Abdullah ibn Dinar, Abu Salih, Abu Huraira, The Messenger peace be upon him}

After preprocessing all the documents in the dataset, a total of 445 narrators are obtained.

In order to calculate the term weights, the standard *tf* × *idf* equation will be changed, because, the number of occurrences of term (narrator) *i* in a hadith *j*, is always equal to 1, and *idf* (inverse document frequency) does not take into account the order of narrators.

In our case, the “Term frequency” is not the number of occurrences of term *t* in document *d*; it is the number of occurrences of term *t* in position *p* as can be seen in table 4.

Table 4. Number of occurrences of term *t* in position *p* in the collection

Position Term	1	2	3	4	5
ابن عباس	0 times	0	0	0	3
أبي هريرة	0	0	1	5	7
الزهري	0	0	1	3	0
الليث بن سعد	0	3	3	0	0
زهير بن حرب	4	0	0	0	0
أنس بن مالك	0	1	2	2	4

For example, to calculate *idf* of the narrator (Anas):

$$\mathrm{idf}_{\mathrm{p}}(Anas) = \ln\left(\frac{120}{1+K}\right)$$

Where *k* is the number of occurrences of the narrator “Anas” in position *p* in the 120 Hadiths.

(1+k) to avoid division by zero if k=0 (the narrator ‘N’ never exists in the position ‘P’)

So for example, if the narrator "Anas" exists only once in position 7:

$$\mathrm{idf}_{7}(Anas) = \ln\left(\frac{120}{1+1}\right)$$

The table below shows the *idf* of some narrator according to position of appearance in the text.

Table 5. *idf* of term *t* in position *p*

Position Term	1	2	3	4	5
ابن عباس	4.78	4.78	4.78	4.78	3.40
أبي هريرة	4.78	4.78	4.09	2.99	2.70
الزهري	4.78	4.78	4.09	3.40	4.78
الليث بن سعد	4.78	3.40	3.40	4.78	4.78
زهير بن حرب	3.17	4.78	4.78	4.78	4.78
أنس بن مالك	4.78	4.09	3.68	3.68	3.17

4. RESULTS

A number of 445 of narrators are obtained from the result of the preprocessing step, so the vector space has a dimension of 445.

The results after testing the proposed method on the dataset are given in tables 6 and 7.

Table 6. Euclidean distances between clusters centers

Category	Sahih	Hasan	Da’if	Maudu’
Sahih	0	31.182	31.639	28.829
Hasan	31.182	0	33.459	30.983
Da’if	31.639	33.459	0	29.823
Maudu’	28.829	30.983	29.823	0

Table 7. Classification results

Category	number of Hadiths tested	correct results	precision
Sahih	10	8	80%
Hasan	10	2	20%
Da’if	10	0	0%
Maudu’	10	10	100%

Hadiths which have narrators who are not in the training set, have zeros in the dimensions of these narrators; this justifies the results on Da’if Hadiths which most narrators are seen for the first time by our classifier; so they are close to the zero vector, hence they are close to Maudu’ cluster; so the program has classified them as Maudu’ (as can be seen in table 8).

Table 8. Euclidean distances between clusters centers and the zero vector

Categories	Null vector
Sahih	23.064
Hasan	25.269
Da’if	24.383
Maudu’	20.778

5. CONCLUSION

In this paper, we introduced a new document representation based on Vector Space Model and taking into account the words order, we applied our method to represent the Hadith, where the narrators order is critical. Hence the classification using Learning Vector Quantization gives good results for the main categories (Sahih and Maudu'). But all Hadiths, which have narrators who are not in the training corpus, were classified as Maudu'.

As for future works, some Hadiths contain: "Tell us Hafs and Khalid said: Tell us Omar and Zuhair..." In this case we must evaluate two narrators and consider them in the same position.

6. REFERENCES

[1] El Kourdi, M., Bensaid, A., Rachidi, T.E. (2004). Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. 20th International Conference on Computational Linguistics, Geneva.

[2] Mesleh, A. (2007). Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System. Journal of Computer Science, 3(6), pp. 430-435.

[3] Sawaf, H., Zaplo, J., Ney, H. (2001). Statistical Classification Methods for Arabic News Articles. the Arabic Natural Language Processing Workshop (ACL2001), Toulouse, France.

[4] Martín-Valdivia, M.T., García-Vega, M., Ureña-López, L.A. (2003). LVQ for text categorization using a multilingual linguistic resource. Neurocomputing, 55(3), 665-679.

[5] Harrag, F., El-Qawasmah, E. (2009). Neural Network for Arabic text classification. Applications of Digital Information and Web Technologies, ICADIWT'09, Second International Conference on the IEEE pp. 778-783.

[6] Kashif, B., Sajjad, Mohsin. (2012). Muhadith: A Cloud Based Distributed Expert System for Classification of Ahadith. Frontiers of Information Technology (FIT), 10th International Conference, pp. 73-78.

[7] Karim, N., Hazmi, N. (2005). Assessing Islamic information quality on the Internet: A case of information about hadith. Malaysian Journal of Library and Information Science, 10.2, 51.

[8] Hernández-Reyes, E., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., García-Hernández, R.A. (2006). Document representation based on maximal frequent sequence sets. Progress in Pattern Recognition, Image Analysis and Applications, pp. 854-863. Springer Berlin Heidelberg.

[9] Salton, G., McGill, M. (1983). Introduction to Modern Information Retrieval. New York.

[10] Al-Shalabi, R., Kanaan, G., Gharaibeh, M. (2006). Arabic Text Categorization Using kNN Algorithm. the Int. multi conf. on computer science and information technology.

[11] Kohonen, T. (1997). Learning vector quantization. In: Self-Organizing Maps, p. 203-217. Springer Berlin Heidelberg.

[12] Martín-Valdivia, M.T., Ureña-López, L.A., García-Vega, M. (2007). The Learning Vector Quantization Algorithm Applied to Automatic Text Classification Tasks. Neural Networks, 20(6), 748-756.