Knowledge Discovery in Al-Hadith Using Text Classification Algorithm

Khitam Jbara

Department of Computer Science, King Abdullah II School for Information Technology, The University Of Jordan, P.O. Box 710481 Amman 11171 Jordan.
ktjlc2000@yahoo.com

Abstract: Machine Learning and Data Mining are applied to language datasets in order to discover patterns for English and other European languages, Arabic language belongs to the Semitic family of languages, which differs from European languages in syntax, semantic and morphology. One of the difficulties in Arabic language is that it has a complex morphological structure and orthographic variations. This study is conducted to examine knowledge discovery from AL-Hadith through classification algorithm in order to classify AL-Hadith to one of predefined classes (books), where AL-Hadith is the saying of Prophet Mohammed (Peace and blessings of Allah be upon him (PBUH)) and the second religious source for all Muslims, and because of its importance for Muslims all over the word knowledge discovery from AL-Hadith will make AL-Hadith more understandable for both Muslims and nonmuslims.
[Khitam Jbara. Knowledge Discovery in Al-Hadith Using Text Classification Algorithm. Journal of American Science 2010;6(11):485-494]. (ISSN: 1545-1003).

Keywords: AL-Hadith, classification, stem, feature, class, expansion, training set.

1. Introduction

Information Retrieval (IR) is the discipline that deals with retrieval of unstructured data, especially textual documents, in response to a query, which may itself be unstructured like sentence or structured like Boolean expression. The need for effective methods of automated IR has grown in the last years because of tremendous explosion of the amount of unstructured data (Greengrass, 2000).

Text mining is a class of what is called nontraditional (IR) strategies (Kroeze, et al., 2003).The goal of these strategies is to reduce the required effort from users to obtain useful information from large computerized text data sources. Also text classifications (TC) is a subfield of data mining which refers generally to the process of deriving high quality of information from a text, which is typically derived through the dividing of patterns and trends through methods such as statistical pattern learning.

However; text classification is one of the most important topics in the field of natural language processing (NLP), where the purpose of its Algorithm is to assign each document of text dataset to one or more pre-specified classes. More formally if di is a document of set of documents D and {c1,c2,…,cn }is The set of all classes, then text classification assigns one category cj to a document di and in multi-subjects classification di can be assigned to more than one class from a set of classes.

Text classification techniques are used in many applications, including e-mail filtering, mail routing, spam filtering, news monitoring, sorting through digitized paper archives, automated indexing of scientific articles, classification of news stories and searching for interesting information on the web (Khreisat, 2006).

Also, an important research topic appears in this field called Automatic text classification (ATC) because of the inception of the digital documents. Today, ATC is a necessity due to the large amount of text documents that users have to deal with (Duwairi, 2006).

According to the growth of text documents and Arabic document sources on the web, information retrieval becomes an important task to satisfy the needs of different end users; while automatic text (or document) categorization becomes an important attempt to save human effort required in performing manual categorization.

In this paper, a knowledge discovery algorithm for AL-Hadith is proposed in order to classify it to one of predefined classes (books), this algorithm is consists of two major phases; the training phase and Classification phase. Experiments will be conducted on a selected set of AL-Hadith from Al-Bukhari book, where thirteen books were chosen as classes in order to run these experiments. The evaluation of the proposed algorithm is carried out by comparing its results to Al-Bukhari classification.

This paper is organized as follows; related work is represented in section 2, while section 3 represents the proposed classification system, and section 4 analyze experiments and results, finally section 5 demonstrates conclusion.

2. Related Work

Most of nowadays classifiers were built for English or European languages. For example, Zhang (2004) builds a Naïve Bayes (NB) classifier, which calculates the posterior probability for classes then the estimation is based on the training set that consists of pre-classified documents, in his system testing phase the posterior probability for each class is computed then the document is classified to the class that has the maximum posterior probability.

Isa, et al. (2008) explore the benefits of using enhanced hybrid classification method through the utilization of the NB classifier and Support Vector Machine (SVM). While Lam, et al. (1999) built a neural network classifier addressing the classifier drawbacks and how to improve its performance.

Bellot, et al. (2003) propose an approach that combines a named entity recognition system and an answer retrieval system based on Vector Space model and uses some knowledge bases, while Liu, et al. (2004) focus on solving the problem of using training data set to find representative words for each class, also (Lukui, et al. 2007) explore how to improve the executing efficiency for classification methods.

On the other hand, Yu-ping, et al. (2007) propose a multi-subject text classification algorithm based on fuzzy support vector machines (MFSVM).

In the Arabic language field, AL-Kabi, et al. (2007) present a comparative study that represents the efficiency of different measures to classify Arabic documents. Their experiments show that NB method slightly outperforms the other methods, while AL-Mesleh (2007) proposes a classification system based on Support Vector Machines (SVMs), where his classifier uses CHI square as a feature selection method in the pre-processing step of text classification system procedure.

El-Halees (2006) introduces a system called ArabCat based on maximum entropy model to classify Arabic documents, and Saleem et al. (2004) present an approach that combines shallow parsing and information extraction techniques with conventional information retrieval, while Khreisat (2006) conducts a comprehensive study for the behavior of the N- Gram Frequency Statistics technique for classifying Arabic text document.

Hammo, et al. (2002) design and implement a Question Answering (QA) system called QARAB.EL-Kourdi, et al. (2004) build an Arabic document classification system to classify non-vocalized Arabic web documents based on Naïve Bayes algorithm, while AL-Kabi, et al. (2005) represent an automatic classifier to classify the verses of Fatiha and Yaseen Chapters to predefined themes, where the system is based on linear classification function (score function), and (Hammo, et al. 2008) discuss the enhancement of Arabic passage retrieval for both diacritisized and non-diacritisized text, they propose a passage retrieval approach to search for diacritic and diacritic-less text through query expansion to match user's query**.**

3. Proposed Classification System

The proposed system consists of four phases; first one is the preprocessing phase. Second phase is the training phase where the learning database is constructed which contains the weights of features representing a class. The input for this phase is a set of pre-classified documents. Third phase is the classification phase in which the resulted training database of previous phase is used with the classification method to classify targeted Hadith, also a query expansion occurs in this phase and the output of it will be the class (book) of targeted AL-Hadith. Finally, data analyzing and evaluation phase. These phases are shown in figure 1.We can define the corpus that contains a set of Ahadith (plural of AL-Hadith) as in definition 1.

Definition 1: Corpus Definition
Suppose corpus $\mathbf{C} = \{H_1, H_2, H_3, \ldots, H_n\}$. Where $\mathbf{H_i}$ represents the *i*th tested Hadith in **C** , *n* is the number of tested Hadith in the **C** and i: 1…..*n*.
Suppose $\mathbf{H_j} = \{w_1, w_2, w_3, \ldots, w_m\}$. Where $\mathbf{w_d}$ represents the *d*th word in AL-Hadith $\mathbf{H_j}$, *m* is the number of words in $\mathbf{H_j}$ and *d*: 1….*m*.

Figure 2 shows an example of Hadith from the book of food that will be used in the illustration of each step of the proposed system.

3.1 Preprocessing phase

In this section the preprocessing techniques are introduced, preprocessing will be conducted on each Hadith used in the training and testing sets. This stage is necessary before the classification phase can be applied to discover knowledge from AL-Hadith and it consists of several sub phases:

1. **Removing Sanad**: this process is done manually and aims to remove Sanad which is a part of AL_Hadith that refers to the chain of names of persons who have transmitted AL-Hadith.
2. **Tokenization**: which aims to divide AL_Hadith into tokens (words); AL-Hadith tokenization was

easily resolved since each token can be identified as a string of letters between white spaces.

3. **Removing punctuation and diacritical marks**: removing diacritical and punctuation marks is important since those marks are prevalent in Ahadith and have no effect on determining AL_Hadith class.
4. **Removing stop words**: Stop words are words that found in AL-Hadith and have no discriminative meaning (AL-Kabi, et al., 2005). In the proposed system a list of stop words is built manually and it consists of Arabic pronouns, prepositions, names of people (companions of Prophet Mohammed) and places were mentioned in AL-Hadith corpus. Then after removing stop words from AL_Hadith, the remaining words (terms) are considered as features.
5. **Stemming:** In this step the stems of features are extracted, stem extraction implemented is considered as light stem extraction which depends on removing some prefixes or suffixes from the word to relate the word to its stems, we used the stemming algorithm proposed by (Al-Serhan, et al., 2003). The result of stem extraction was filtered to eliminate the incorrect (roots less than three characters). The resulted stems will be used in the query expansion process which will be discussed in details in section 3.3.2, Table 1 shows all steps of preprocessing for AL-Hadith that is presented in figure 2.

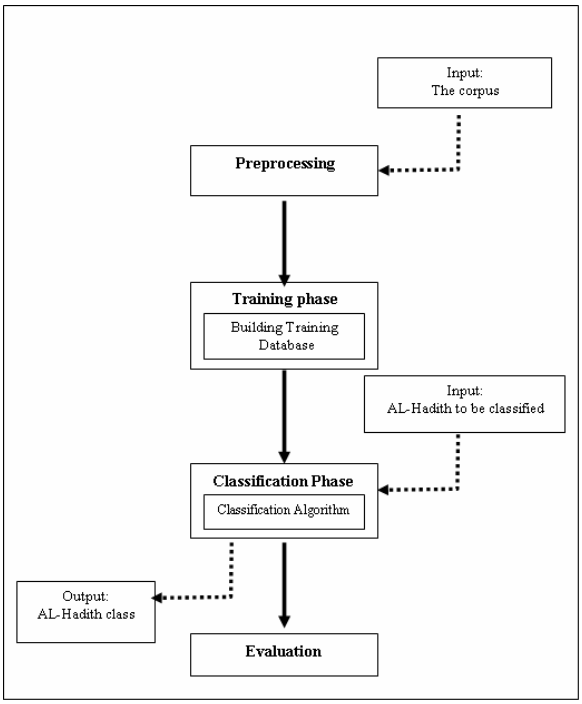

Figure 1. An overview of proposed system phases.

حدثني إسحق بن إبراهيم: أخبرنا رَوْح بن عبادة: حدثنا ابن أبي ذئب، عن سعيد المَقْبُري، عن أبي هريرة رضي الله عنه:

أنه مر بقوم بين أيديهم شاة مَصْلِيَّة، فدعوه، فأبى أن يأكل وقال: خرج رسول الله صلى الله عليه وسلم من الدنيا ولم يشبع من خبز الشعير.

Figure 2. Example of AL-Hadith from the book of food.

Table1. Results of preprocessing phase steps for AL-Hadith in figure 2

Step	Result of the step
Removing Sanad	أنه مر بقوم بين أيديهم شاة مَصْلِيَّة، فدعوه، فأبى أن يأكل وقال: خرج رسول الله صلى الله عليه وسلم من الدنيا ولم يشبع من خبز الشعير.
Tokenization	{"أنه","مر" ,"بقوم"," بين"," أيديهم", "شاة"," مَصْلِيَّة","،","," فدعوه","،","," فأبى","," أن", " يأكل", "وقال", ": ","خرج","رسول", "الله","," صلى","," الله","," عليه","," وسلم", "من", "الدنيا", "ولم", "يشبع","," من","خبز","," الشعير","."}
Removing Punctuation and Diacritical Marks	{ انه , مر, بقوم, بين, ايديهم, شاة, مصلية, فدعوه, فابى, ان, ياكل, وقال, خرج, رسول, الله ,صلى, الله, عليه, وسلم, من, الدنيا, ولم, يشبع, من, خبز, الشعير}
Removing Stop Words	{ مر، بقوم، ايديهم، شاة، مصلية، فدعوه، فابى، ياكل، خرج ، الله ،صلى ، عليه، وسلم ، الدنيا ، يشبع ، خبز، الشعير}
Stemming (valid stems)	{ايدي،دنيا،شعير}

3.2 Training Phase

Supervised classification exploits the predefined training documents that belong to specific class to extract the features that representing a class. Therefore, every class will have a feature vector representing it, and then these features will be reduced using one of the features selection

techniques. Feature vectors will be used later by the classification algorithm in the testing phase.

Supervised classification has its difficulties; one main problem is how to be sure that trained document actually belongs to a specific class. In this study this problem is resolved by conducting it on a set of AL-Hadith that has been classified by the famous AL-Hadith scientist AL-Bukhari who gave us a good base to evaluate the proposed algorithm.

Training phase consists of two main stages; first one is executed once to produce Inverse Document Frequency (IDF) matrix for the corpus while the second one is executed for each training set.

3.2.1 Corpus IDF Matrix

After conducting the preprocessing phase a list of features for each Hadith in the corpus is produced and will be used in the classification process. Building the IDF matrix for AL-Hadith corpus is done only one time and it will be used in the classification process every time the IDF value for a feature is needed. The IDF value for a given feature is computed according to equation (1).

$$IDF_i = \log\left(\frac{N}{DF_i}\right) \quad (1)$$

Where

N: number of Ahadith in the corpus.

DF_i: Number of Ahadith in the corpus containing feature i.

Fewer documents containing a given feature will produce a larger IDF value and if every document in the collection contains a given feature, feature IDF will be zero, in other words the feature which occurs in every document in a given collection is not likely to be useful for distinguishing relevant from non-relevant documents. Table 2 shows the IDF matrix structure.

Table 2. Corpus IDF matrix.

	Pre-defined Classes(Books)					
Feature	**Book1**	**Book2**	**Book3**	**….**	**Bookc**	**Feature redundancy**
Feature1	Log (N/DF_1)	log (N/DF_1)	log (N/DF_1)	…..	log (N/DF_1)	([DF_1]/N)*100
Feature2	Log (N/DF_2)	log (N/DF_2)	log (N/DF_2)	…..	log (N/DF_2)	([DF_2]/N)*100
Feature3	Log (N/DF_3)	log (N/DF_3)	log (N/DF_3)	…..		
Feature4	Log (N/DF_4)	log (N/DF_4)	log (N/DF_4)	…..		
Feature5	Log (N/DF_5)	log (N/DF_5)	log (N/DF_5)	…..		
…..				…..		
…..				…..		
…..				…..		
…..				…..		
…..				…..		
……				…..		
Feature n	Log (N/DF_N)	log (N/DF_N)	log (N/DF_N)	…..	log (N/DFc)	([DF_n]/N)*100

3.2.2 Weight calculations for training sets features

The proposed system depends on using a set of Ahadith as training documents to extract representative words for each book (class) to compute their weights. The weight of a given feature in a given document is calculated as (TF×IDF) because this weighting schema combines the importance of TF and IDF at the same time, and the features training weights is computed according to equation (2).

$$TW_{bi} = TF_{bi} \times IDF_i \quad (2)$$

Where:

TW_{bi}: feature i training weight in training set b.

TF_{bi} : feature i frequency in training set b.

IDF_i: feature i inverse document frequency calculated earlier (IDF matrix).

Features that will be considered to calculate their training weights must satisfy the feature redundancy threshold 45, that's mean that feature redundancy must be less than 45.

Table 3 shows training weights for features in the training set b in general, while Table 4 shows training weights for features in a training set from the book of food.

Table3.Training weights for features in training set b.

Table 4. Training weights for features in a training set from the book of food.

The book of food (training set No.1)			
Feature	IDF	TF	TW
ياكل	1.80	8	14.39
الدنيا	1.84	1	1.84
شاة	1.97	4	7.90
مر	2.12	1	2.12
ايديهم	2.22	1	2.22
خبز	2.34	4	9.37
فابى	2.42	1	2.42
الشعير	2.52	2	5.04

3.3 Classification Process

The classification process consists of four steps as shown in Figure 3. First step is computing query weights where feature's weight in targeted AL-Hadith is found. Second step is the expansion process where the stems are used to expand the query. Third step is calculating the similarity coefficient for each feature in AL-Hadith to be classified, and the final step is finding the cumulative similarity for AL-Hadith over the predefined classes (books).

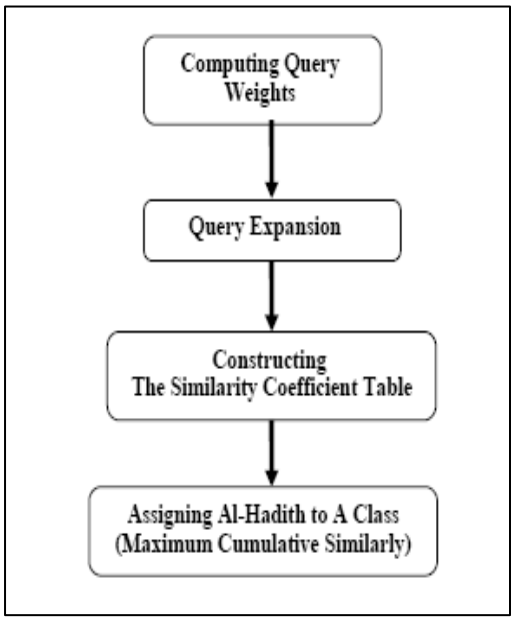

Figure 3. Classification process steps

3.3.1 Computing Query Weights

A feature weight in the query (specific Hadith) is calculated according to equation (3) :

$$Q_hW_i = TF_{hi} \times IDF_i \qquad (3)$$

Where:

Q_hW_i: feature i weight in AL-Hadith h (Hadith to be classified).

TF_{hi} : feature i frequency in AL-Hadith h

		Bookb	
Feature	IDF	TF	TW
Feature1	IDF1	TFb1	TWb1 = TFb1 *IDF1
Feature2	IDF2	TFb2	TWb2 = TFb2 *IDF2
Feature3	IDF3	TFb3	TWb3 = TFb3 *IDF3
Feature4	IDF4	TFb4	TWb4 = TFb4 *IDF4
…..			
Feature n			TWbn = TFbn *IDFn

IDF_i: inverse document frequency calculated by equation (1).

Query weights as shown in Table 5 will be computed for each feature in AL-Hadith to be classified. Feature frequency (TF) depends on AL-Hadith features occurrence while Inverse document frequency (IDF) is a global value referenced from IDF matrix.

Table 5. Query Weights Table for Mined Hadith

No.	Feature	IDF	Feature Redundancy	TF	QW
1	الدنيا	1.84	1.44	1	1.84
2	الشعير	2.52	0.30	1	2.52
3	الله	Feature redundancy >45	80.02	2	0
4	ايديهم	2.22	0.61	1	2.22
5	بقوم	2.64	0.23	1	2.64
6	خبز	2.34	0.45	1	2.34
7	خرج	1.53	2.95	1	1.53
8	شاة	1.97	1.06	1	1.97
9	صلى	Feature redundancy >45	68.89	1	0
10	فابى	2.42	0.38	1	2.42
11	فدعوه	3.12	0.08	1	3.12
12	مر	2.12	0.76	1	2.12
13	مصلية	3.12	0.08	1	3.12
14	وسلم	Feature redundancy >45	68.58	1	0
15	ياكل	1.80	1.59	1	1.80
16	يشبع	2.64	0.23	1	2.64

3.3.2 Query Expansion

The process of query expansion depends mainly on using the stems of features to expand the searching area. The stems for all features in the training set and AL-Hadith to be classified were produced in the preprocessing phase.

In the expansion process the newly added stems in the expanded training set will have the same weights for its origin feature. In other words, if we have the couple {(W,S) , (W, TW)} where S is the stem of word W and TW is the training weight for W from the training weights Table then the weight for stem S will be the same weight of W.

The same procedure is applied to stems in expanding the query set where stem S will have the same weight of its origin word W from the query

weight Table. The extended query weights for AL-Hadith sample are shown in Table 6.

	Pre-defined Themes			
Feature	**Book1**	**Book2**	**…**	**Book13**
Feature1	$Sim1=Q_bW_1 *T_1W_1$			$QbW_1 *T_{13}W_1$
Feature2	$Sim2=Q_bW_2 *T_1W_2$			$QbW_2 *T_{13}W_2$
Feature3	$Sim3=Q_bW_3 *T_1W_3$			$QbW_3 *T_{13}W_3$
Feature4	$Sim4=Q_bW_4 *T_1W_4$			$QbW_4 *T_{13}W_4$
Feature5	$Sim5=Q_bW_5 *T_1W_5$			$QbW_5 *T_{13}W_5$
Feature n	$Sim\ n=Q_bW_n *T_1W_n$			
	$\sum_{i=1}^{n} Sim$			

3.3.3 Constructing the Similarity Coefficient Table

In the proposed system the cosine similarity coefficient is used, where the similarity between two documents (document (D) & query (Q)) is actually the cosine of the angle (in N-dimensions) between the 2 vectors and can be calculated according to equation (4) (Baarah, 2007):

$$sim(D,Q) = \sum_{i=1}^{n} (W_{di} \times W_{qi}) \quad (4)$$

Where, denote the query feature and n is the number of feature in the query Hadith.

Table 7 shows similarity coefficient for features in the mined AL-Hadith in general, while Table 8 shows the similarity coefficient for features in AL-Hadith illustrated in figure 2 against the training set from the book of food shown in section 3.2.2.

Table 6. Extended query weights table for mined Al-Hadith

No	Feature	IDF	Feature Redundancy	TF	QW
1	الدنيا	1.84	1.44	1	1.84
2	الشعير	2.52	0.30	1	2.52
3	الله	Feature redundancy >45	80.02	2	0
4	ايديهم	2.22	0.61	1	2.22
5	بقوم	2.64	0.23	1	2.64
6	خبز	2.34	0.45	1	2.34
7	خرج	1.53	2.95	1	1.53
8	شاة	1.97	1.06	1	1.97
9	صلى	Feature redundancy >45	68.89	1	0
10	فابى	2.42	0.38	1	2.42
11	فدعوه	3.12	0.08	1	3.12
12	مر	2.12	0.76	1	2.12
13	مصلية	3.12	0.08	1	3.12
14	وسلم	Feature redundancy >45	68.58	1	0
15	ياكل	1.80	1.59	1	1.80
16	يشبع	2.64	0.23	1	2.64
17	دنيا	Feature No. 1			1.84
18	شعير	Feature No. 2			2.52
19	ايدي	Feature No. 4			2.22

Table 7. similarity coefficient for features for mined Hadith in general.

3.3.4 Assigning AL_Hadith to a class

After constructing the similarity coefficient table for AL-Hadith to be classified against the predefined classes, the cumulative similarity weights for mined Hadith will be found against each of those classes. The cumulative similarity values indicate common features between AL_Hadith to be classified and the predefined books.

After finding the cumulative weight for the mined Hadith with correspondence to each predefined book (class), AL-Hadith will be assigned to the book with the maximum cumulative weight, because maximum cumulative weight is an indication of larger common features between the training set and the mined AL-Hadith features set.

Table 8. Similarity coefficient for features in the example Hadith against training set from the book of food.

No.	Feature	Similarity coeffecient
1	الدنيا	3.39
2	الشعير	12.69
3	الله	0.00
4	ايديهم	4.92
5	بقوم	0.00
6	خبز	21.93
7	خرج	0.00
8	شاة	1.97
9	صلى	0.00
10	فابى	2.42
11	فدعوه	0.00
12	مر	2.12
13	مصلية	0.00
14	وسلم	0.00
15	ياكل	1.80
16	يشبع	0.00
17	دنيا	3.39
18	شعير	12.69
19	ايدي	4.92
cumulative similarty		72.26

4. Experiments and Results

In this section, an overview is given for AL-Hadith corpus content that is used in this study to run the experiments of the proposed classifying algorithm, and details of experiments are also illustrated.

4.1 Content of AL-Hadith Corpus

AL-Hadith corpus that is used in running the experiments consist of thirteen books (classes). Ahadith were taken from Sahih AL-Bukhari which is the most well known Hadith book all over the Islamic world and the most trusted Hadith book for researchers in this field. Twelve of those books were included in AL_Kabi (2007) study and the Book of the (Virtues of the Prophet and His Companions) is added to the experiment in this study with 143 additional Ahadith.

Table 9 shows statistical information of books included in the experiments along with its name in English and Arabic as it was used by AL-Bukhari in his Sahih. The testing corpus has 1321 Hadith distributed over 13 books (classes).

Table 9. List of books in AL-Hadith corpus

Book (Class)Name	اسم الكتاب	Doc No.	No. of distinct features after stop words removal
The Book of Faith	كتاب الايمان	38	938
The Book of Knowledge	كتاب العلم	76	1946
The Book of Praying	كتاب الصلاه	115	2137
The Book of Call to Praying	كتاب الأذان	38	574
The Book of the Eclipse Prayer	كتاب الكسوف	24	715
The Book of Almsgiving	كتاب الزكاه	91	2267
The Book of Good Manners	كتاب الأدب	225	5258
The Book of Fasting	كتاب الصوم	107	1905
The Book of medicine	كتاب الطب	92	1895
The Book of Food	كتاب الطعام	91	1894
The Book of Pilgrimage (Hajj)	كتاب الحج	231	4885
The Book of Grievance	كتاب المظالم	40	906
The Book of the Virtues of the Prophet and His Companions	كتاب المناقب	143	3410

4.2 Classification Methods Applied to AL-Hadith Corpus

One of the researches in AL-Hadith classification field is done by (AL-Kabi, et al., 2007), in which AL-Kabi did not mention an accurate description of AL-Hadith corpus or the stop words list used in their expirements. Threrefore , in this study an implementation for their classification algorthim on the corpus used is done.

The following subsections represents in details the three methods have been implemented in this study.

4.2.1.AL-Kabi method : this method was propsed by AL-Kabi and his Colleagues on AL-Hadith classification(AL-Kabi, et al., 2007). This method is based mainly on using the stems of Ahadith words to calculate the IDF, the weighting of the feature in training phase and the classification process.

4.2.2. Word based classification (WBC): this method uses the words of AL-Hadith after going through the preprocessing phase without stemming stage. The words occurrences after preprocessing are used in the calculation of IDF and in the weighting process. Stems of the words are not used in this method neither in building the training database nor in applying the classification algorithm.

4.2.3. Stem expansion classification (SEC): It is the proposed method in this study. In which words and stems are used. Words are used in IDF and features weighting calculations for both training and query sets, but stems are used in query expanding process, the expansion process was discussed in details in section 3.3.2.

4.3 Experiments Specifications

Hadith corpus that is used in this study consists of 1321 Hadith distributed over thirteen books (classes). The system is considered as supervised classification since training sets are used to apply learning algorithm to build the leaning database which will be used for the classification algorithm.

In the experiments author uses (90%) of each Ahadith class as training set while the rest (10%) of each class is used as testing set for the classification system. Of course, for each class the (10%) Ahadith in the testing set are not included in the training set of Ahadith and the training phase calculation.

For each class five testing – training sets combination are chosen to run SEC algorithm, which means that for each class five separable experiments will be run, which gives variation for system testing.

It is important to mention that the same training - testing sets combination is used with the other two classification methods (AL-Kabi's,WBC) which is an important aspect to insure fair comparison among different methods against the proposed one.

4.4 Performance Measurements

In order to demonstrate the efficiency of any classification algorithm, measurements are needed to compare the proposed system's outcome with others. The most popular measurements in text classification algorithm are recall, precision and F-measure that are used in this study.

Recall and precision based on the concept of relevance. Precision is defined as the ratio of relevant documents retrieved to all documents retrieved while Recall is defined as the ratio of relevant items retrieved to all relevant items in a corpus.

There are obvious trade-off between recall and Precision. If the system retrieves all the documents in a corpus then the system will retrieve all relevant documents in the corpus, in this case the recall will be perfect.

On the other hand, since there are only small proportions of documents in a collection that are truly relevant to the given query, retrieving everything will give a very low precision (Greengrass, 2000).

Because of this trade-off between recall and precision a combination of good precision and good recall is needed. In the best case we would like to retrieve all the relevant documents and to discard non-relevant documents, this combination of recall and precision is found in F-measure (Harmonic mean).Precision, recall and F-measure are calculated according to equation presented by (Al-Mesleh, 2007) as follows

Precision (P) =A/ (A+ B).

Recall(R) = A/ (A+ C).

F-measure (Harmonic mean) = (2×P×R)/(P+R)

Where the meaning of parameters used in recall and precision calculations are shown in Table 10.

Table 10. Recall and Precision Parameters.

System says…	In reality, the document is…	
	Relevant	Irrelevant
document is relevant	A	B
document is irrelevant	C	D

4 .5 Comparisons and Results Analysis

In this section we introduce the comparisons between SEC and the other two methods (AL-Kabi's and WBC), in order to show the preference of the proposed system over AL-Kabi's and WBC method.

4.5.1 Stem Expansion vs. Word based classification

The proposed system (SEC) outperform WBC for 11 out of 13 books in precision, while WBS achieve better precision for The Book of Grievance and the book of Knowledge as shown in Figure 4 . This result is predicted because using the stem expansion phase gives a large morphology variation for the words that can resulted in retrieving more documents that belong in reality to other classes.

The proposed algorithm (SEC) overcomes the side effect of expanding the query using stemming, by the weighting strategy adopted in the system, where stems in the expansion phase are giving the same weights of its original word in AL-Hadith.

SEC achieved precision value of 1 for The Book of Call to Praying and enhances the precision of 11 classes by 45% in average.

As shown in Figure 5 SEC achieves better F_measure values for all classes against WBS method and enhances the F_ measure by 49% in average.

4.5.2 Stem expansion vs. AL-Kabi classification

After implementing AL-Kabi's method for the same training -testing sets used to examine SEC, comparisons are conducted as shown in Figure 6 and 7.

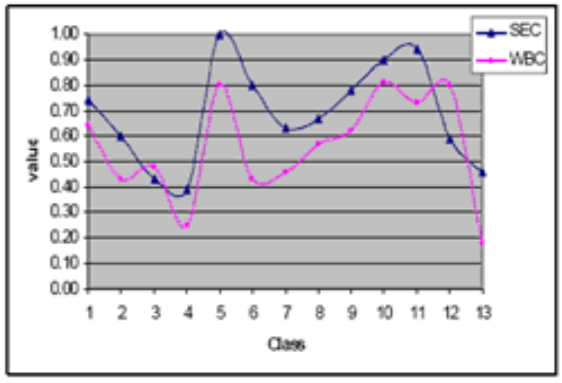

Figure 4. Precision comparison for stem expansion vs. word based classification.

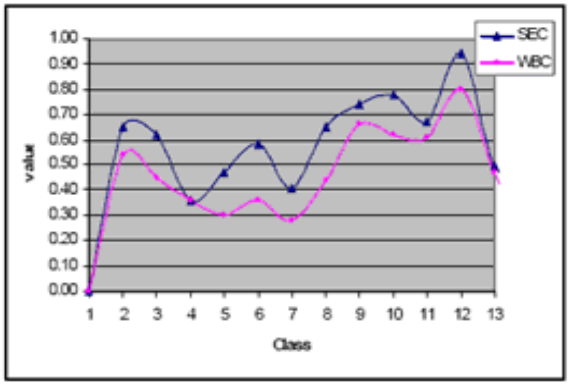

Figure 5. F_measure comparison for stem expansion vs. word based classification.

Figure 6 represents the precision graph of the two methods, where AL-Kabi's method achieved better precision for The Book of Knowledge and The Book of Grievance, while SEC out performed AL-Kabi's method in 11 out of 13 classes.

This behavior can be justified by the fact that those two books have small number of Ahadith and since the experiments are conducted on Ahadith in a closed domain (Sahih AL_Bukhari), author

believes that if those classes have larger number of Ahadith the superb of SEC will appear.
In addition AL-Kabi used the stem of words for term weighting in the preprocessing phase, which means that term frequency used in the training process will be the number of the stems occurrence in the training set. In contrary, in SEC the words occurrences is used for the weighting process in preprocessing phase while the stems are used in the expansion process for both training and testing sets.

Since stem expansion is used for both the training and testing sets in the proposed system, more

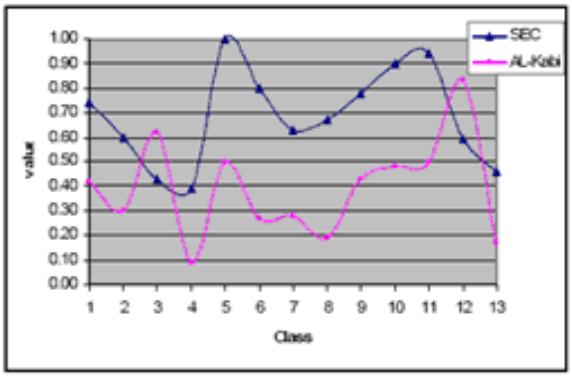

Figure 6. Precision Comparison for Stem expansion vs. AL-Kabi's classification

non-related documents are presented to be retrieved, which justifying AL-Kabi's method achieves better precision for two classes out of thirteen classes .
As Shown in Figure 7, SEC method outperformed AL-Kabi F_Measure in all the 13 classes.

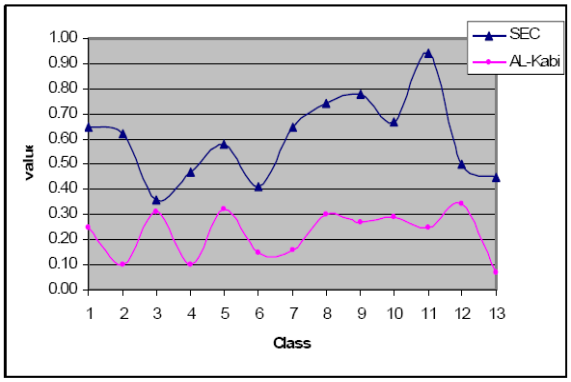

Figure 7. F_Measure comparison for stem expansion vs. Al-kabi's classification.

5. Conclusions

Arabic language is considered as one of the languages that will never distinguish and few researches were made on Arabic corpus linguistics. However, it is the official language of twenty Middle Eastern and African countries and is the religious language of all Muslims, regardless of their origin.

A classification method called Stem Expansion is proposed in this study, in order to discover knowledge from AL-Hadith by assigning each Hadith to one book (class) of predefined classes. SEC is considered as supervised classification method.

In this study a corpus containing 1321 Hadith from thirteen books from Sahih AL-Bukhari is selected and each Hadith is assigned to one class. Sahih AL-Bukhari is used as the base for deciding the correctness of classification results.

The results of the proposed system (SEC) are compared with the results of two methods; one proposed by AL-Kabi and the other is word based classification technique (WBC). The comparison shows that SEC was better against WBC and AL-Kabi in recall for all classes while WBC and AL-Kabi achieve better precision for only two out of thirteen classes, and SEC achieves better F_Measure for all the thirteen classes against the other two methods (WBC and AL-Kabi).

The results show that SEC performed better in classifying AL_Hadith against existing classifications methods (WBC and AL-Kabi) according to the most reliable measurements (recall, precision, and F_Measure) in text classification field.

Acknowledgements:
Grateful thanks, gratitude and sincerest appreciation to Dr. Azzam T. Sleit and Dr. Bassam H. Hammo for their guidance.

Corresponding Author:
Khitam M.Jbara.
Department of Computer science.
The University Of Jordan.
P.O. Box : 710481 Amman 11171 Jordan.
Amman,Jordan.
E-mail: ktjlc2000@yahoo.com

References

1. AL-Kabi, M. N., AL-Sinjilawi, S. I. (2007). "A Comparative Study of the Efficiency of Different Measures to Classify Arabic Text", University of Sharjah Journal of Pure and Applied Sciences, 4(2), pp. 13-26.
2. AL-Kabi, M. N., and AL-Shalabi (2005),"AL-Hadith Text Classifier", Journal of applied sciences, 5(3), pp.548-587.
3. Al-Serhan, H., Al Shalabi, R. and Kannan, G. (2003). "New Approach for Extracting Arabic roots". Proceedings of the 2003 Arab conference on Information Technology (ACIT'2003), pp 42-59.
4. Bellot P., Crestan E., El-Bèze M., Gillard L., de Loupy C. (2003),"Coupling Named Entity

Recognition, Vector-Space Model and Knowledge Bases for TREC-11 Question Answering Track", In Proceedings of The Eleventh Text retrieval Conference (TREC 2002), NIST Special Publication.
5. Duwairi, R. M. (2006). "Machine Learning for Arabic Text Categorization". Journal of the American Society for Information Science and Technology, 57(8), pp.1005-1010.
6. El-Halees, A. (2006). "Mining Arabic Association Rules for Text Classification", Proceedings of the First International Conference on Mathematical Sciences, AL-Azhar University of Gaza, Palestine, 15-17 July.
7. El-Kourdi, M., Bensaid, A., and Rachidi, T. (2004). "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm", Workshop on Computational Approaches to Arabic Script-Based language (COLING-2004), University of Geneva, Geneve, Switzerland, pp. 51-58 .
8. Greengrass Ed. (2000), . Information Retrieval: A Survey, 2000.
9. Hammo, B., Abu-Salem, H., Lytinen, S., and Evens, M. (2002). "A Question Answering System to Support the Arabic Language", Workshop on Computational Approaches to Semitic Languages, Philadelphia, PA, July, pp.55-65.
10. Hammo, B., Sleit, A., El- Haj, M. (2008), "Enhancing Retrieval Effectiveness of Diacritisized Arabic Passages Using Stemmer and Thesaurus", The 19th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS2008).
11. Isa, D., Lee, L.H., Kallimani, V.P., RajKumar, R. (2008),"Text Document pre-processing with the Bayes formula for Classification using the Support Vector Machine", IEEE Transactions on Knowledge and Data engineering, Volume 20, Issue 9, pp. 1264 - 1272 .
12. Kanaan, G., AL-Kabi, M. N., and AL-Shalabi, R. ,2005 ,"Statistical Classifier of the Holy Quran Verses (Fatiha and YaSeen Chapters)", Journal of Applied Science, 5(3), pp.580-583.
13. Khreisat, L. (2006). "Arabic Text classification Using N-gram Frequency Statistics A Comparative study", Proceedings of the 2006 International Conference on Data Mining (DMIN 2006), Las vegas, USA, pp. 78-82.
14. Kroeze, J., Matthee, M., and Bothma T. (2003),"Differentiating Data- and Text-Mining Featureinology",University of Pretoria, Proceedings of SAICSIT 2003, pp. 93 –101
15. Lam, S. L. and Lee, D. L. (1999). "Feature Reduction for Neural Network Based Text Categorization", Proceedings of the Sixth IEEE International Conference on Database Advanced Systems for Advanced Application, Hsinchu, Taiwan, pp.195-202.
16. Liu, B., Li, X., Lee, W. S., and Philip, S. Yu3 (2004), "Text Classification by Labeling Words", in Proceedings of Nineteeth National Conference on Artificial Intelligence (AAAI-2004), San Jose, USA, pp.425-430.
17. Lukui S., Jun Z., Enhai L. Pilian H. (2007),"Text Classification Based on Nonlinear Dimensionality Reduction Techniques and Support Vector Machines", Third International Conference on Natural Computation (ICNC 2007), pp. 674-677.
18. Saleem A.,Martha E. (2004), "Event Extraction and Classification for Arabic Information Retrieval Systems", IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004).
19. Yu-ping, Q., Qing A., Xiu-kun, W., Xiang-na, L. (2007), "Study on Classification Algorithm of Multi-subject Text", ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, pp. 435-438.
20. Zhang J. (2004)."The Optimality of Naïve Bayes ", Proceedings of the 17th International Florida Artificial Intelligence Research Society Conferences, Florida,USA , pp.562-567.
21. Baarah, A.H (2007). "Applying Text Mining Algorithms For Knowledge Discovery On The Holy Quran Text".Master's Degree thesis (not published), The University of Jordan, Amman, Jordan.
22. Larkey, L. S., Ballesteros, L., and Connell, M. E. (2002), "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis", Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Developing in Information Retrieval, Tempore, Finland, August 11-15, 275-282.

24/8/2010