Arabic WebPages Classification Based On Fuzzy Association

Aayat Shdaifat, Marwah ALian
[1, 2] Basic Sciences Department, Hashemite University
Zarqa, Jordan

Abstract

Information retrieval from web documents becomes a challenge according to the exponential growth in the number of web pages on the Internet and rapid changes in these pages. So, it is necessary to classify web pages into classes to provide their results for the applied tools to be used which makes information retrieval easier and helps in facilitating applications use on the internet. Web pages classification may be applied for various types of data that are available on the Internet like texts, images, audios and videos. Each type has different algorithms to classify and process. Unfortunately, the classification of Arabic Web pages considering their structure is more difficult. In this paper, the results of the classification of Arabic web pages which is obtained using fuzzy algorithm will be discussed.

Keywords: *Webpages, classification, fuzzy logic, Arabic language classification, stem level, word level.*

1. Introduction

With the daily increase of web pages numbers, the issue of classification gains more importance especially in information retrieval application. Since 1990, few researches have been done in the field of information retrieval and web documents classification for Arabic web pages which rapidly have been increased.
Webpage classification also called webpage categorization defined as the task to determine whether a webpage belongs to a single category or to many categories [1].
And classification of web documents content is important to many tasks in web information retrieval like maintaining web directories and focused crawling. Additional challenges are presented in web page classification field because of the uncontrolled nature of web content as compared to traditional text [2]. While text classification may be defined as the process of classifying text into a set of predefined set of classes based on its contents [3].

The importance of web classification appears through the use of many applications to the text and document classification like Automated indexing of documents which is one of the most known applications of document classification where documents are classified to a predefined set of classes using Meta data which can be used in classifying scientific articles like ACM, and in hospitals to classify patients' files. Also it can be used for WebPages using meta-data. Document organization is another application of document classification in which the vocabulary is controlled; it is used for newspapers in which past articles must be classified and makes access easier. [3].

The subject of this paper is concerned in classification of Arabic web pages by using Fuzzy association Theory. It is assumed that this method may be useful for documents of Arabic Web with respect to the special properties which it has; Arabic is a Semitic language that depends on the stem of a word; therefore, Arabic web pages need special preprocessing like dealing with the diacritics before classification process.

The rest of the paper is organized as follows. First, in section 2, we give a brief of related work for web pages classification. In section 3, we describe fuzzy classification algorithm. Subsequently, in section 4, we discuss the proposed model that depends on fuzzy association.
In section 5 we describe our experiment and implied dataset and show the experimental results. A comparison between fuzzy association and vector space mode is represented in section 6. Finally we conclude and outline future research direction in section 7.

2. Related Work

Some different works about classification of web pages have been done using Fuzzy Relationship but they are more in English, some in Arabic and other languages like Persian Language.

Many researches proposed methods and systems for classifying the web pages in English due to the various applications as in email filtering and spamming [4-10]. One of the researches that concerned in fuzzy association for English web documents is the work of Tsekouras, et al. [11] in which they propose a method for classifying web documents through fuzzy logic classification and data

clustering. First, they extract a number of words for each class. Then, they use an algorithm to partition the available words into a number of clusters and the center corresponded to a word. They use Hamming distance to calculate the dissimilarity measure between two words, and then they estimate the distances between this document and all cluster centers of each category. Finally, the category with smallest distances will be selected.

While the work of Yari, et al. **[10]** presents classification of Persian web documents using a model in Fuzzy Theory. The results of this research is obtained by a survey process in which the job sequences include the construction of software, dataset of Persian Web documents, testing and assessment of the documents. The output of this research is satisfactory since the accuracy of the classification is compared to the average accuracy of classification in other researches.

In the field of Arabic web classification, Al-Taani and Al-Awad [12] present analysis and comparison of six fuzzy similarity approaches applied to Arabic web pages classification. These approaches are Algebraic, Hamacher, MinMax, Special case fuzzy and Bounded Difference. In this work the clustering scheme is built and known for each category from training documents and the similarity between a test document and a category is measured using a fuzzy relation. The best performance achieved by Einstein measure, then the bounded measure, followed by AL -gebric measure.

However, in this study the operation of web documents classification is carried out in Arabic language by applying fuzzy association to Arabic web pages then testing the results using recall, precision and f-measure factors.

3. Classification using Fuzzy Associations

Fuzzy association is used to capture the relationships among different index terms. Each pair of words has an association value to distinguish it from the other. The fuzzy sets are sets in which elements have degrees of membership. Fuzzy Association algorithm is shown in Figure 1.

Fuzzy associative information retrieval (IR) is a mechanism that is formalized within the fuzzy set theory and based on the definition of fuzzy association. It picks up the association between the keywords to improve the retrieval results from traditional IR systems.

FUZZY ASSOCAIATION ALGORITHM
INPUT: keyword indexed in document i
OUTPUT: classes of document i
BEGIN:

1. μ=0
2. L=1
3. **FOR** $\forall k\, a \epsilon\, d\, i$
4. **FOR** $\forall kb \in CK\, j$
5. L= L*(1-$r_{a,b}$)
6. **NEXT** b
7. N=1-L ;
8. **IF** N > μ **THEN** μ= N
9. **NEXT** a
10. **END** ALGORITHM

Procedure Correlation matrix
INPUT: a collection of all keywords appearing in training data
OUTPUT: Relationship between all key words in A
Begin:

FOR all i ϵ A

FOR all j ϵ A

$$r_{i,j} = \frac{n_{i,j}}{ni + nj - ni,j}$$

NEXT i
NEXT j

Where:
A is the collection of all keywords.
$n_{i,j}$ is the total number of documents where keyword i and j both appear.
n_i is the total number of documents where keyword i appears.
n is the total number of documents where keyword j appears.

Fig 1: Fuzzy Association Algorithm [13]

By providing the association between key words, additional documents that are not directly indexed by the keywords in the query request can also be retrieved [13].

4. The Proposed Model:

This study is concerning to applying fuzzy association to Arabic web pages, where data is collected and preprocessed before applying fuzzy association.

Data is collected from many Arabic websites, about 22399 documents, using Teleport pro tool, and the preprocessing stage includes five steps; converting from HTML to text, tokenization, morphological preprocessing, stop words removal, and light stemming.

Converting from HTML to text is a necessary step because it is hard to deal with pure HTML, next step is tokenization using one of the standard methods: separation of whitespace, alphabetic strings, and alphanumeric string. When this step is finished, ten bags of words will be resulted. The step of stop words removal; is the step of removing functional words that doesn't carry particular or useful meaning in information retrieval (IR), then applying light stemmer for data.

The classification and selection process is done by calculating the weight for all words in each class, then, selecting the word with the high weight for all classes. The word in each class must be distinct from the other classes. Noun and adjective are selected manually. If a word has a high weight in more than one class then it is ignored, such as " أزمة", which can be classified under News, Health and Economic classes. After that, a process of self-judgment and Knowledge for all words in each class is applied. Finally, assistance from experts in each field to approve a list of classified word is supported.

In this model, Fuzzy association algorithm is used, which is an algorithm that depends on the relation between terms (words) to classify Arabic web pages into eight classes; Social Science, News, Health, Education, Science, Economics and Art.

Fuzzy association algorithm is applied for stems and words; recall, precision, and F- measure. It is better to apply fuzzy association on stems than to apply it on words, since the stem expands the domain of the word that can be used.

Figure 2 depicts the flow diagram of the proposed model which contains the following steps: First; collecting classified web pages for 10 classes including sport, art, economy, computer, social society, health, science, news, education and government. The collection is used in system testing and training. Next step is; performing preprocessing task.

The third step is; applying fuzzy association algorithm for both training and testing data. Finally; the final results are checked. More details about these steps are given in the following sub sections.

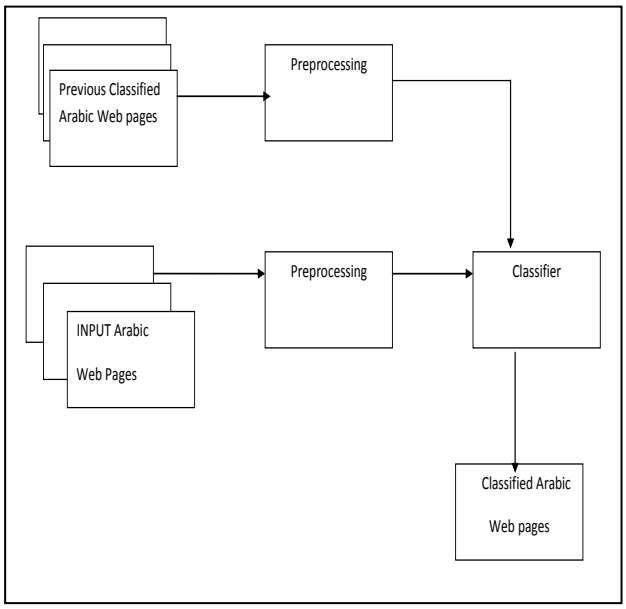

Fig 2: Web Pages Classification Using Fuzzy Association

4.1. Data Collection

Collecting data is a very important step for testing the proposed system. We used teleport pro as a tool for this purpose. Teleport Pro [14] is an all-purpose, high-speed tool for getting data from Internet; it provides ten simultaneous retrieval threads, accesses password-protected sites, filters files by size and type, searches for keywords, and much more.

Web pages are collected for training data that is used to train the system. Training data require a huge amount of web pages, to extract words and classify it into a set of ten classes. Also, these pages are needed to calculate the associated value between every two words (terms). About 22399 web pages have been collected for this study (1.68 GB). Table 1 shows the training data collected from different resource, and Table 2 shows the number of documents for each class used for the training data.

Table 1: Set of Training Data

Arabic Site Name	*URL*	*Class*	*Number of documents*
الجزيرة	www.aljazeera.net	News	1064
العربية	http://www.alarabiya.net	News	830
الاسواق	http://www.alaswaq.net/	Economic	2290
لافضل صحة	http://www.allbestHealth.com/	Health	2219
الجزيرة الرياضية	http://www.aljazeeraSport.net/HomePage	Sport	1817
الموسوعة العربية للكمبيوتر	http://www.c4arab.com/	Computer	2190
الموقع الرسمي لحكومة دبي	http://www.dubai.ae/	Government	346
حكومة دولة قطر	http://portal.www.gov.qa/	Government	1662
الملتقى الاول للتشكيليين العرب	http://www.fonon.net/	Art	1869
مجلة افق الثقافية	http://ofouq.com/today/index.php	Art	1042
الرأي	www.alrai.com	Social since	749
الدستور	http://www.addustour.com/Section.aspx?sec=1	Social since	1011
وزارة التربية و التعليم	www.moe.gov.jo	Education	1773
موقع ادارة النشاط العلمي	http://www.alme.gov.sa/	Education	830
علوم وتكنولوجيا	http://arabic.irib.ir/pages/Science/	Science	244
علوم و تكنولوجيا	http://scitech-ar.blogspot.com/	Science	42
شبكة علوم الاحياء	http://allbiology.net/	Science	51
شبكة العلوم العربية	http://olom.info/mgz/	Science	15
موقع الكون	http://www.alkoon.alnomrosi.net/	Science	285
افكار علمية	http://www.afkaaar.com/html/	Science	2085
Total			22399

Table 2: Number of documents of Each Class

NO.	*Class Name*	*NO. of documents per class*
1	Art	2911
2	Computer	2290
3	Economic	2219
4	Education	2603
5	Government	2008
6	Health	2190
7	News	1894
8	Science	2707
9	Social Science	1760
10	Sport	1817
Total		22399

4.2. Preprocessing

The preprocessing step includes converting HTML web pages to text files, tokenizing the text documents, removing stop word, indexing and stemming.

The step of converting from HTML to text; in this step we use the simple Html to text converter tool in order to get text from HTML pages because we can not deal with pure HTML web pages. The whole process is shown in Figure3.

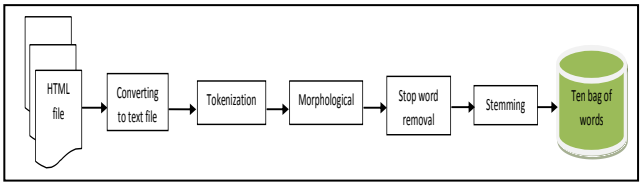

Fig. 3: Preprocesing stages

4.2.1 Tokenization

While tokenizing the documents is the first step in text classification in order to obtain the bag of words to represent feature vectors.

The standard methods used for tokenization include: separation of whitespace, alphabetic strings, and alphanumeric strings. In this work we use whitespace as the separator between words.

The tokens are saved in a table using SQL, in order to be used in other steps in this work.

4.2.2. Morphological Preprocessing

Also, morphological preprocessing is needed; that's because in written Arabic, diacritics are often omitted in text and a familiar reader with this language will not face any difficulty to read a text without vowels in a correct manner. In addition, the letters change its form according to their position in the word. Some of these letters make light modifications in writing which does not influence the meaning of the word. Regarding these entire features of Arabic language and to overcome the problem of variations of Arabic letters, we applied a normalization method on both testing and training data before indexing. [15] Then, replacing " أ " , " آ" and "إ " by " ا". Then, removing shada "ّ" and the diacritics.

4.2.3. Stop Words Removal

Similar to other languages, Arabic language contains functional words (stop words) which do not carry a particular and useful meaning for Information Retrieval (IR) [15]. In this study a list of stop words that includes 282 words is used as shown in Table 3.

4.3. Light Stemming

Light stemming is a process of stripping off a small set of prefixes and/or suffixes without dealing with infixes or recognized patterns. It also finds roots. The stemmer that is used can do the following: Striping off initial ('و'), removing definite Articles (ال، وال، بال، كال، فال، لل), and removing suffixes (ها، ان، ون، ين، يه، ية، ه، ة، ي).

4.4. Words Classification

In Arabic language there is no domain ontology as in English language, so it must be built. The algorithm depends on calculating the weight of all words in each class, selecting the high weight for all classes, applying self-judgment for all words in each class, and taking experts' approval. The steps of this algorithm are shown in details in Figure 4.

Table3: list of stop words

الآن	لي	لم	إلا	بِهِ	إليه	فهي	في
هنا	كنت	كان	الاَ	كَانَ	لأنها	فهو	من
ليس	أيهما	كل	الا	لاَ	لانها	فهم	على
لو	ايهما	تلك	وان	اِلَّا	لإنها	فهن	ان
يمكن	فلا	هو	وإن	اللَّهُ	لكل	لدى	أن
لماذا	فلابد	ولكن	وأن	وَلَا	كل	يتم	إن
بل	فلم	اي	عليه	بِمَا	انت	أحياناً	لو
بكل	فلماذا	أي	شيء	ذَٰلِكَ	أنت	احياناً	أو
عن	فلن	عندما	عنه	هُوَ	إنت	احيانا	عن
معه	فما	غير	والتي	مِّنَ	منها	أحيانا	آل
كيف	فماذا	وفي	له	مِنْ	انى	وكيف	ابداً
عند	فمن	بعد	مما	لَوْ	أنى	بلا	ابدا
أما	فحن	حتى	كم	اللَّهُ	ثم	منه	أبداً
إما	ففي	بل	بهذه	الَّتِي	د	منها	أبدا
اما	فهذا	بأن	به	إن	أ	منهن	لا
يكن	فهذه	هي	بها	لَهُ	م	منهم	لأ
وما	فهكذا	كما	لذا	فِيهَا	ا	آخر	الى
كي	فهل	بين	إذاً	مَن	حيث	اخر	إلى
فان	فهما	قبل	إذا	قُلْ	لان	أنن	الذي
فلن	آنٍ	خلال	انا	لَهُمْ	لأن	إنن	ما
فأن	فهمت	قد	لانه	إِذَا	لإن	انن	هذه
ماذا	فهنك	امام	لأنه	وَمَن	ومن	لأي	التي
ومع	أنى	أمام	فيها	شَيْءٍ	وهي	لاي	هذا
وقد	أتية	كانت	اليه	مَّا	وهو	لابد	مع
كذا	ألو	لكثر	لهم	اللَّهِ	انتي	لعل	وَمَا
تم	ابداً	أكثر	لهن	لَن	أنها	ولعل	إنَّ
كانوا	أبداً	وهو	لنا	اللَّهُ	إنها	تلك	إنني
هم	ولعلها	هل	لكم	قَالَ	انها	لهذا	أنني
هن	لها	ولا	لكا	مِنَ	أنهما	جداً	مَا
أنتم	عليها	ولأ	فقد	مِّن	إنهما	جدا	الَّذِينَ
انتم	ولم	انه	والذي	إِلَى	انهما	حينما	عَلَى
لنتما	بين	أنه	والتي	اللَّهِ	اهم	بدله	ثُمَّ
أنتما	و	إنه	فيما	وَلاَ	أهم	يأتي	اللَّهَ
انتن	في	فيه	بدُّه	لاَ	وهذا	يلتي	اِلاَّ
أنتن	مِن	وكان	بدُه	يَا	وإنما	انتا	إنتا

1. Calculate the weight of all words in each class. The weight was calculated used the following equation:
   Weight = tf * idf
   Where
   tf : term frequency
   idf: inverse documents frequency
   idf = log N/df
   Where
   N: total number of document.
   df: documents frequency.
2. Select the high weight for all classes. We used noun and adjective and we selected manually.
3. The word in each class must be distinct from the other classes.
4. If the words have high weight in more than one classes then we ignore it, such as " أزمة" , which can be classified under News ,Health and Economics classes.
5. Applying self-judgment and Knowledge for all words in each class.
6. Experts' approval.

Fig. 4: Algorithm for calculating the weight for each word.

Table 4 shows the number of distinct word for each class.

Table 4: Number of Distinct Word of Each Class

NO.	*Class Name*	*Number of words*
1	Art	153
2	Computer	50
3	Economic	263
4	Education	140
5	Government	78
6	Health	367
7	News	257
8	Science	201
9	Social Science	166
10	Sport	90

Two classes are ignored; Computer and Government classes; the first one is ignored for the few number of words that it has, and the second one because it is merged with the Social Science class due to the high similarity between the two classes. As a result of this step, we have eight classes rather than ten classes.

5. Experiments and results

Two experiments on the selected data are made; the first one is applying fuzzy association using the stems of words, while the second one applying fuzzy association on words without stemming. In both experiments we use recall and precision as evaluation measures for information retrieval. More details are given in section 5.1 and 5.2.

Table 5 illustrates the number of test documents for each class. The size of used data is about 59.4 Mega Byte.

Table 5: Test Data for Each Class

NO.	*Class Name*	*Number of documents*
1	Art	786
2	Economic	762
3	Education	784
4	Health	723
5	News	627
6	Science	897
7	Social Science	839
8	Sport	697

5.1. Applying Fuzzy Association on the Stems of Words

We start with applying fuzzy association using stems of words, then we use recall and precision as evaluation measures for information retrieval, Where recall is defined as the ratio of correct documents by the system divided by the total number of correct predicted classified documents, while precision is the ratio of correct classified documents divided by the total number of system documents.

The results of recall and precision using stems are shown in Figure 5 and Figure 6 respectively. As shown in Figure 5 the highest recall is in the Health class because it has a large number of classified words. The similarity between Health and Science that is observed when analyzing data resulted in a low recall for Science. Also, there is a similarity between Social Science class and News class which is the reason for the low value of recall for Social Science class. Furthermore, the number of classified words used in News class is larger than that in Social Science class.

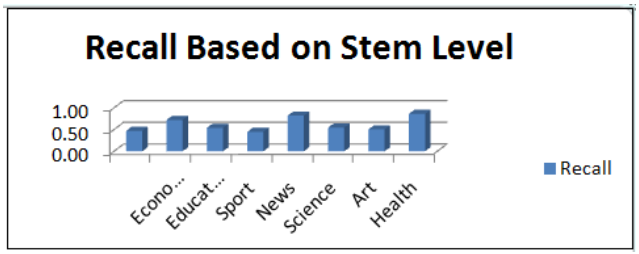

Fig.5 Recall based on Stem

Also, Figure 5 represents that the sport class has the lowest value of recall because it has the lowest number of classified words. However, it has a good precision because of the limited number of intersected used words in this class with other classes. The Education class has a low recall value, because of the number of words classified under this class were not commonly used.

From Figure 5 and Figure 6 we can see that Science class has high precision but low recall. This was mainly due to the high number of classified words and the similarity at this class with the Health class as discussed before. However, Art class has high precision and low recall because the number of words classified under the Art class is not commonly used in other classes. This indicates that most of the retrieved documents are relevant. Table 6 shows the precision and recall when applying fuzzy association based on stems of words.

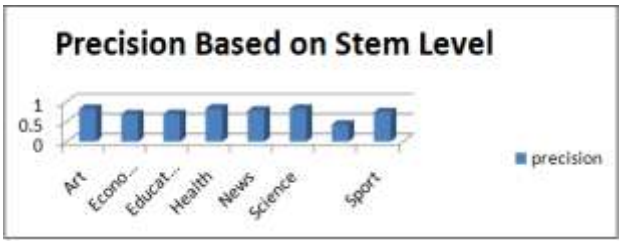

Fig. 6 Precision based on Stem Level

Table 6: Recall and Precision on Stem

Class	*Recall*	*Precision*
Social science	0.47	0.43
Economic	0.72	0.68
Education	0.54	0.69
Sport	0.45	0.74
News	0.82	0.78
Science	0.55	0.82
Art	0.50	0.82
Health	0.86	0.83

5.2. Fuzzy Associations Based on Words

The second experiment is to apply fuzzy association on words without stemming. The recall and precision at the word level is shown in Figures 7 and 8 respectively. As shown in Figure 7 the best recall value is in News class because of high number of classified words in this class. And by analyzing data it is found that there is a similarity between Science class and Health class which implies that science class has low recall value.

Fig. 7 Recall based on word level

From Figure 7 and 8 the sport class has high precision and low recall because it has the lowest number of classified words and it has limited number of intersected words with other classes. while Education class and Social Science class have higher precision values than recall values.

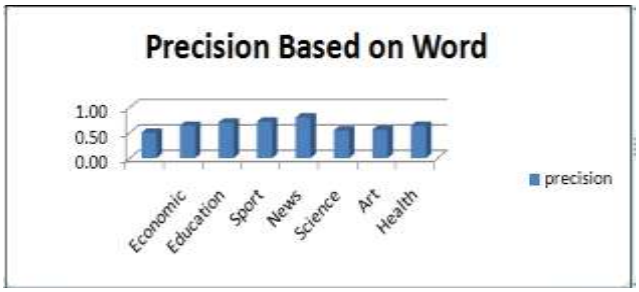

Fig. 8 Precision based on Word Level

Table 7 represents the recall and precision values for each class when applying fuzzy association on word level.

Table7: Recall and Precision on Word Level

Class	*Recall*	*Precision*
Social science	0.45	0.50
Economic	0.78	0.63
Education	0.63	0.70
Sport	0.42	0.71
News	0.82	0.80
Science	0.39	0.54
Art	0.56	0.56
Health	0.63	0.63

5.3. Stem Level and Word Level

The recall at stem level is higher than that at the word level in four classes; Health, Science, Sport and Social Science, but it is lower in three classes; Art, Economic and Education, and they are equal in one class which is News class. Figure 9 represents the recall at stem level and word level. These results can be explained as follows; the number of classified words that are used in the stem level is greater than the number of words that are used in the word level. For example, the word "المريض " is stemmed to "مريض " , then all words that have this stem are used to classify related web pages, such as:
" المريض، المريضين، المريضات، المريضون، مريضون، مريضات، مريض، فالمريض، فالمريضات، .. etc."

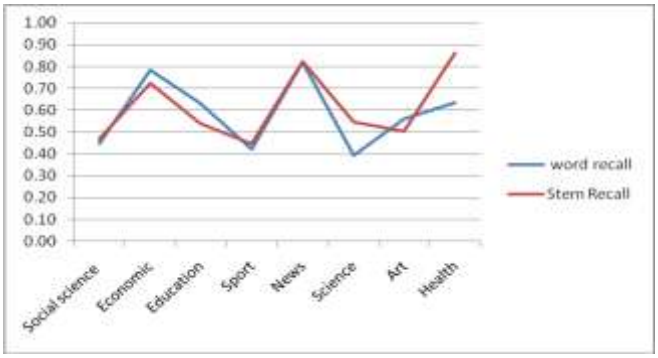

Fig. 9 Recalls on Stem and Word Level

The precision at stem level is better in five class; Economic, Sport, Science, Art and Health class, while the remaining three classes; Social Science, Education and News class, get better precision at the word level. Word level shows better precision because of the use of light stemmer in this system which deals with prefixes and suffixes.

The stem of a word gives an abstract meaning, so using stems of words sometimes produces difficulty in classification process and gives unexpected results. This is because the stem may appear in different documents classified under many classes. For example, the word "تأكله ", and "تأكله ", both have different meaning even they have the same stem " اكل"، since we have replaced the " آ " and the "أ " by "ا " in the preprocessing steps. Also the diacritics that are used in Arabic language affect the meaning of the word. For example, the word " مُناخ " differs from the word "مِناخ" , and the words " بِر " , " بُر " and "بَر " have different meaning when the diacritics are ignored. Figure 10 shows the precision for both word level and stem level.

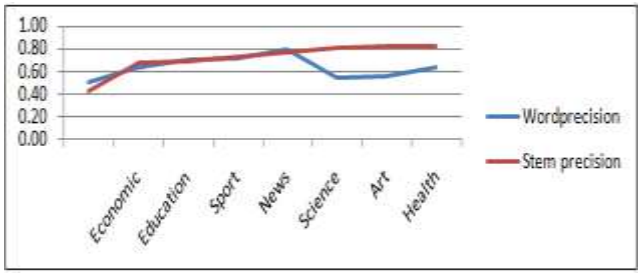

Fig.10 Precision on Stem and Word Level

F-measuer is used as an information retrieval measure which combines recall and precision with equal weight and calculated as in formula 1:

F-measure = $(2\times p\times r)/(p+r)$ (1)

Where p is the precision and r is the recall. [16]

F-measure is shown in Table 8, which results from applying fuzzy assocation in each class for both levels; stem and word levels.

Table 8: F-measure at Word Level and Stem Level

Class Name	*Word F-measure*	*Stem f-measure*
Social science	0.47	0.45
Economic	0.70	0.70
Education	0.66	0.61
Sport	0.53	0.56
News	0.81	0.80
Science	0.46	0.65
Art	0.56	0.62
Health	0.63	0.84

The F-measure using the stems of words is better than that of using the words alone. This was not the case for Art, Education and News classes, since there recall and precision values are higher at the word level than the recall and precision at the stem level. Figure 11 shows the F-measure curves for both the word level and the stem level.

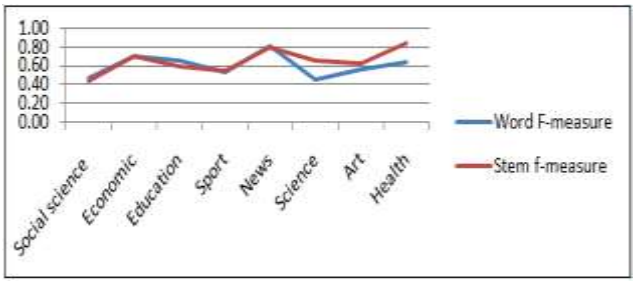

Fig. 11 F-measure at the Word Level and Stem level

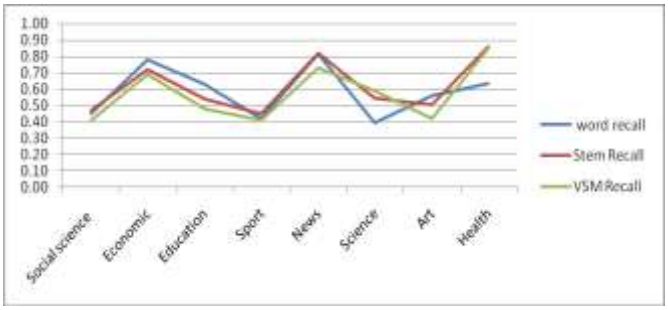

Fig. 12 Comparisons between the Three Types of Data Using the Recall measure

6. Fuzzy Association and VSM

Vector Space Model (VSM) is the most widely used model in Information Retrieval (IR). It can be defined as an algebraic model that represents the documents in the vector space.

When comparing fuzzy associations with the VSM, The fuzzy association gives the highest recall and F-measure values, while it gives the highest precision in the most other classes.

The recall in the proposed system with fuzzy association gives values between 45% in Social Science class to 84% in Health class; while the vector space model gives 82% in the Social Science class to 48% in the News class. In spite of these good results, we were troubled with the huge size that we need to execute the system; a computer with minimum of 4 Giga RAM is needed and about 20 GB is the resulted database.

Figure 12 shows the recall of the three types; using the fuzzy association technique applied on stems, words, and VSM. The recall of science class is best in VSM than other because of similarity between science and Health. While Figure 13 shows the precision using the fuzzy association technique applied on stems, words and VSM, where the VSM is best in Social science and education that depend on number of classified word and number of testing document in these classes and similarity between social science and news.

Figure 14 shows the F-measure of the fuzzy association technique applied on stem level, word level, and the VSM, also the figure shows that the VSM is best in social science and science because of similarity between Social science and news and between science and health.

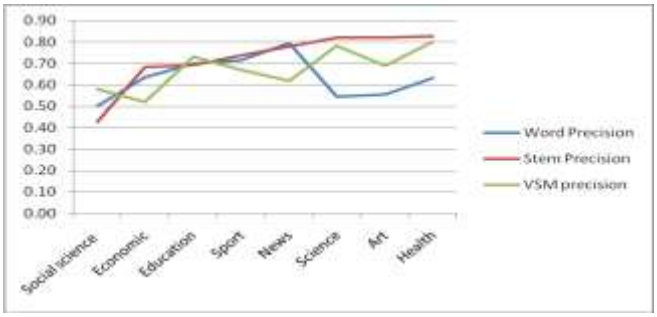

Fig. 13 Comparisons between the Three Types of Data Using Precision

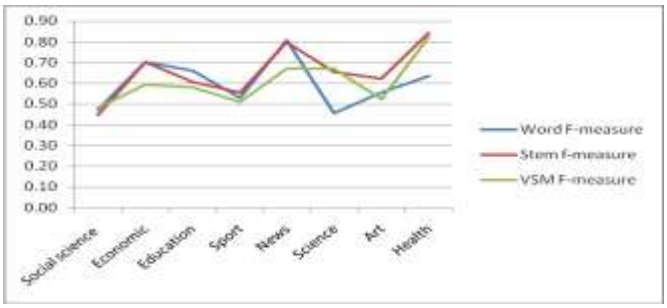

Fig. 14 Comparisons between the Three Types of Data Using the F-measure

7. Conclusion and future work

This research presents a study to increase the performance of web page retrieval using fuzzy association. It is concerned in Arabic web pages.

1.68GB of Arabic web pages is collected as training data while for testing data more than 59.4MB of Arabic web pages is collected.

After preprocessing, words are classified into eight classes then a correlation matrix is calculated depending on the relation between terms in the corpus, which gives the meaning of the association. Then, fuzzy calculations are applied on stemmed words and on original words without stemming. Finally in evaluation, recall, precision and F-measure are used.

The results of this study indicates that the best recall when applying fuzzy association on stems is 86% while the best recall when applying fuzzy association on words is 82%. Furthermore, the best precision when applying fuzzy association on stem is 83%, but the best precision when applying fuzzy association on words is 80%.

As a future work we are looking for using another stemmer that is not light in the preprocessing step. This is supposed to increase the recall and precision. Also, one of future steps is using more data in the training step and increasing the number of words in each class. Also, we are looking for building Arabic domain ontology using Arabic dictionaries and experts help.

References

[1] B.choi,a z.yao, “web page classification”, studies in fuzziness and soft computing 180, springer, pp.221-274, 2005.
[2] Xiaoguang Qi and Brian D. Davison," Web page classification: Features and algorithms", ACM Computing Surveys (CSUR), Volume 41 Issue 2, Article No. 12 ,2009.
[3] Syiam M., Fayed Z., and Habib M., "An Intelligent System for Arabic Text Categorization", International Journal on Intelligent Cooperative Information System, Vol. 6(1), pp.1-19, 2006.
[4] Dumais S., and Chert H, "Hierarchical Classification of Web Content", Preceeding of the 23ed Annual International ACM SIG1R Conference on Research and Development in Information Retrieval ,ACM, pp. 256- 263, 2000.
[5]Haruechaiyasak C., Shyu M., Chen S. and Li X., "Web Document Classification Based on Fuzzy Association", 26th Annual International Computer Software and Applications Conference, pp. 487-492, 2002.
[6]Aixin, S. and Ee-Peng L., Wee-Keong, " Web Classification Using Support Vector Machine", Preceedings of the WIDM’02, pp. 96-99, 2002.
[7]Shen D., Chen Z., Yang Q., Zeng H., Zhang B., Lu Y. and Ma W," Web-page Classification through Summarization". Preceeding of the ACM SIGIR 04, pp. 25–29, 2004.
[8]Holden N. and Freitas A.," Web Page Classification with an Ant Colony Algorithm", Springer-Verlag Berlin Heidelberg, pp.1092–1102, 2004.
[9]Chen R. and Hsieh C.," Web page classification based on a support vector machine using a weighted vote schema", Expert Systems with Applications, Elsevier, pp. 427–435,2006.
[10] Yari A., Abbasi A. and Moemen BellahS., "Presentingfuzzy relation to classify the Persian Web documents", Intelligent Computing and Intelligent Systems (ICIS), IEEE International Conference , Vol 2, pp. 220-223,2010.
[11] Tsekouras G. E., Anagnostopoulos C., Gavalas D. and Dafni E. ,"Classification of Web Documents using Fuzzy Logic Categorical Data Clustering", Artificial Intelligence and Innovations: from Theory to Applications,IFIP The International Federation for Information Processing Volume 247, pp. 93-100, 2007.
[12] Al-Taani, A., and Al-Awad N.K, A," Comparative Study of Web-pages Classification Methods using Fuzzy Operators Applied to Arabic Web-pages", Proceeding of the World Acadmey of Science,Engineering and Technology, pp. 33-35, 2005.
[13] Haruechaiyasak C., Shyu M., Chen S.and Li X, "Web Document Classification Based on Fuzzy Association", 26th Annual International Computer Software and Applications Conference, pp. 487-492,2002.
[14] Downloaded from http://teleport-pro.en.softonic.com.
[15] Y. Kadri and J. Y. Nie, "Effective Stemming for Arabic Information Retrieval", International conference at the British Computer Society, pp. 68-74, 2006.
[16] Yiming yang and xin liu .A," reexamination of text categorization method", proceedings of the ACM SIGIR Conferences on Research and Development in information retrieval, 1999.

Mrs. Aayat M.Shdaifat got Bs.c degree in Computer Information System from Prince Al-Hussain bin Abdullah II Faculty for Information Technology at Hashemite University in 2006 Then she got M.Sc degree in computer Information System, from King Abdullah II School for Information Technology at University of Jordan in 2009. Since 2011 she is working in Hashemite university as an instructor at Faculty of Science.

Ms. Marwah Alian got the B.Sc degree in Computer Science from Hashimite University in 1999. After graduation, she worked as programmer then as a teacher in many high schools in Jordan then she received the M.Sc degree in Computer Science from University of Jordan in 2007 with a thesis titled as: The Shortest Adaptive Learning Path in eLearning Systems.
After graduation she became a member in the technical team of a project called Science Education Enhancement and Development (Seed) supported by Japan International Cooperation Agency (JICA). In 2008 she became a member of Computer Science Department in Science and Information Technology Faculty at Isra University. Since 2011 she became an instructor at Science Faculty in Hashemite University.
Ms.Marwah has interests in researches in the area of elearning Systems, Adaptive Learning, Mobile Learning, data mining and social networks impact. And she has made a number of researches in these areas of research.