World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:3, No:3, 2009

An Empirical Analysis of Arabic WebPages Classification using Fuzzy Operators

Ahmad T. Al-Taani and Noor Aldeen K. Al-Awad

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/15082/pdf

***Abstract*—**In this study, a fuzzy similarity approach for Arabic web pages classification is presented. The approach uses a fuzzy term-category relation by manipulating membership degree for the training data and the degree value for a test web page. Six measures are used and compared in this study. These measures include: Einstein, Algebraic, Hamacher, MinMax, Special case fuzzy and Bounded Difference approaches. These measures are applied and compared using 50 different Arabic web pages. Einstein measure was gave best performance among the other measures. An analysis of these measures and concluding remarks are drawn in this study.

***Keywords*—**Text classification, HTML documents, Web pages, Machine learning, Fuzzy logic, Arabic Web pages.

I. INTRODUCTION

WITH the rapid growth of the Internet, there is an increasing need to provide automated assistance to Web users for Web page classification. Such assistance is helpful in organizing the vast amount of information returned by search engines, or in constructing catalogues that organize Web documents into hierarchical collections [1]. Classification is expected to play an important role in future search services. For example, Chen et al. [2] showed that users prefer navigating through catalogues of pre-classified content. In order to meet such a strong need, we need automated Web-page classification techniques.

Web-page classification is harder than free text classification because of the noisy information founded in them such as advertisement represented through images, media sounds, navigation bars, and page formatting. So we need to summarize and benefit from these data and make them useful for end user who needs to manage and plan their work depending on a more accurate classification process. It is an essential matter to focus on the main subjects and significant content. As a result the critical task to deal with ambiguous web pages and their embedded structure through studying HTML language to remedy the process and then using some classification method such as machine learning, or fuzzy set theory [3].

The language may affect the whole process because of its complexity for dealing with words and phrases, which occurs frequently in Arabic language, in which this language has a little volume of spreading among the web in comparison to the other languages, and here are some factors that shows a clear picture about that:

1. A word may act to be different, depending on the context in which it will occur, so the word may share equally or nearly equal in different classes, so that it makes an ambiguous view, like (رسول→ رسل) , in which it may mean Mohammed (God's praise and peace upon him), messenger, emissary, plover, etc...

2. There are some cases in which words may have more than one root in the native language."سياج" it has two roots ("ساج", "سيج")

3. How to verify from the word structure itself if it starts with the present tense prefixes such as "تقوى", "يمين".

4. There is no indication about the origin of the word if it is a verb or noun; as the following example shows: (يسير) it may be interpreted as the present tense of the verb (سار) or it may be interpreted as the noun that means: (السهولة) simple or facile or uncomplicated.

5. And there are some idioms that occur frequently, and have no direct relevance to any of the categories such as "السوء الحظ" لحسن الحظ","بالإضافة إلى","بغض النظر", etc...

Recently much work has been done on Web-page classification [1] [4-16]. In these approaches different methods are proposed. These methods includes: Web summarization-based classification, fuzzy similarity, natural language parsing web page classification and clustering to find reliable list answers, text classification approach using supervised neural networks, machine learning methods, kNN model-based classifier, and fuzzy classifiers.

In this study, an analysis and comparison of six fuzzy similarity approaches applied to Arabic web pages classification is presented. 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.

This relation is called fuzzy term-category relation; where the set of membership degree of words to a particular category represents the cluster prototype of the learned model. Based

Ahmad T. Al-Taani is with the chairman of the department Computer Sciences, Yarmouk University, Irbid, Jordan. (Correspondence author: e-mail: ahmadta@yu.edu.jo.)

Noor Aldeen K. Al-Awad was graduated from the Department of Computer Sciences, Yarmouk University, Irbid, Jordan, in 2005 (e-mail: noor_kamel@yahoo.com.)

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/15082/pdf

on this relation, the similarity between a document and a category's cluster center is calculated using fuzzy conjunction and disjunction operators [4].

After that the calculated similarity represents the membership degree of document to the category, and each membership functions of fuzzy sets take values in [0,1] that is used for testing different test documents in order to come out with the weakness points as well as the strength points. It may be observed then that representing the document as a Boolean features vector [4] simplifies greatly the fuzzy similarity formula and reduces it to a major factor.

II. Methods

A. Overview of the proposed approach

A fuzzy similarity approach is used for Arabic web-pages classification. The proposed system is composed of five stages: Training, Noise Elimination, learning, classification, and testing stages as shown in Fig. 1. The clustering scheme is built and already known for each category from training documents and the similarity between a test document and a category is measured using a fuzzy relation.

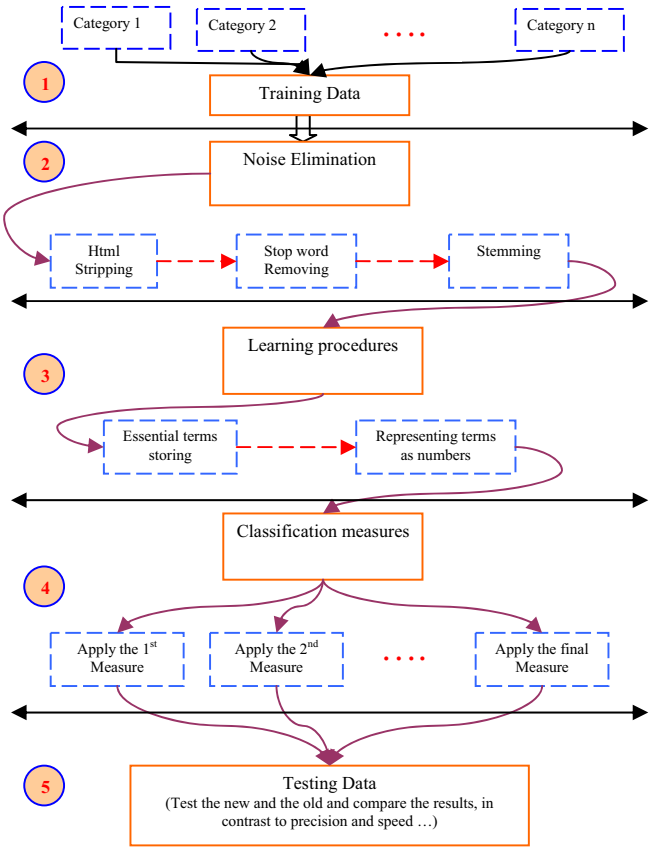


Fig. 1: The stages of the fuzzy similarity approach.

This relation is called fuzzy term-category relation, where the set of membership degree of words to a particular category represents the cluster prototype of the learned model. Based on this relation, the similarity between a document and a category's cluster center is calculated using fuzzy conjunction and disjunction operators. After that the calculated similarity represents the membership degree of the test document to the category, and each membership function takes a value between 0 and 1. This value is used for testing different test documents in order to come out with the weak points as well as the strength points. The document is *then* represented *(in one of the measures that is called Scfuzzy)*as a Boolean features vector, which greatly simplifies the fuzzy similarity formula and reduces it to a major factor [4].

Six well-known measures using fuzzy operators are used and applied to different Arabic web pages. These measures include: Einstein, Hamacher, bounded difference, Algebraic, MinMax, and Scfuzzy approaches. Then comparison between these measures is presented in this study.

B. Categories and their related terms

Text documents are represented as a set of categories: C = $\{c_1, c_2, \ldots, c_n\}$. The category of a document D: $c(D) \in C$, where c(D) is a categorization function whose domain is D and whose range is C [17].

To compute the similarity model of each text document, each document is passed to an HTML stripper, to a stop word eliminator, and to a stemmer. Then weights are computed for each term. These weights represent statistical similarity between documents.

Background knowledge is used in the classification process. Such background knowledge provided us with a corpus of text that contains information both about the importance of words to a category (in terms of their membership values in a large corpus), and the probability of words (what percentage that a test document will participate in the sameness process with the document). This gave us a large context in which to test the similarity of a training example with a new test example. Then this context is used in conjunction with the training examples to label a new web page.

C. Fuzzy conjunction/disjunction based Algorithm

Each document d has one category c; and as a result each category has one or more document. A set of n documents and their related categories are represented as an ordered pair where D = $\{\langle d_1, c(d_1)\rangle, \langle d_2, c(d_2)\rangle, \ldots, \langle d_n, c(d_n)\rangle\}$. The resulted documents that have many terms are stored with their relevant categories, as each row represents (term, document no., category no.). Then each term is counted for each document and is represented as the term and it's frequency as follows: p = $\{\langle t_1, f(t_1)\rangle, \langle t_2, f(t_2)\rangle, \ldots, \langle t_n, f(t_n)\rangle\}$. $f(t_i)$ is the frequency of the term t_i in the document or web page p.

Now the frequencies for each term are summed up for all the documents of a category to give the repetition of terms among their relevant categories. The membership value for each term is obtained by using Formula 1.

$$M(t_i, c_j) = \frac{\sum f(t_i)\{(c(P_k) = c_i), (f(t_i) \in p_K), (p_K \in D)\}}{\sum f(t_i)\{(f(ti) \in p_K), (p_K \in D)\}} \quad (1)$$

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/15082/pdf

The membership value in fuzzy set theory denotes the degree of relevance of term t_i to category c_j, and the terms that consisted out from the n text document and their crisp classification, and so there are multiple categories with their membership degree values related to each term [4]. The binary classification vectors are applied in order to compare them to the existing fuzzy ones. For example, the membership value of some term t_i to the category c_j will be one if all the occurrences of the term t_i happen in only one category.

There are some consequences about the documents that are classified in many categories; and this may result implicitly into moderately convergence between each degree of voting or distribution for the term being examined. Tables I - III show an example of the relevant processes.

TABLE I
TERMS, FREQUENCIES, AND CATEGORIES.

Doc	Term			Cat
	حذق	لطف	هدم	
d1				c1
d2				c2
d3				c3
d4		1		c5
d5		1		c6
d6			1	c8
d7				c7
d8		1	2	c5
d9	1		1	c8
d10			1	c5

TABLE II
TERM-CATEGORY FREQUENCY.

Term	Categories		
	c5	c6	c8
حذق	0	0	1
لطف	2	1	0
هدم	3	0	2

TABLE III
M (TI,CJ):- THE MEMBERSHIP VALUE OF TERM-CATEGORY FREQUENCY.

Term	Categories		
	c5	c6	c8
حذق	0	0	1
لطف	0.6666667	0.3333333	0
هدم	0.6	0	0.4

According to table III, term 1 (حذق) will be classified to category c8, term 2 (لطف) will be classified to category c5, and term (هدم) will be classified to category c5.

D. Fuzzy-based similarity approach

After computing the membership values of the individual terms in each category, then we need to measure the likelihood for a given test web page to be classified into the existing categories of the training datasets.

The test web page can be classified correctly if each of its terms is participated in the process of comparison or similarity. Let a test web page w = {⟨t_1, Deg (t_1)⟩, ⟨t_2, Deg (t_2)⟩... ⟨t_n, Deg (t_n)⟩}, where Deg (ti) represents the membership degree for a (ti) to be associated with a test web page and computed by formula 2:

$$Deg\ (t_i) = \frac{f(t_i)}{Max\ (f(t_i))} = \frac{f(t_i)}{f(t_m)} \quad (2)$$

Then the similarity between w and a category c_j is given by formula 3:

$$sim\ (w, c_j) = \frac{\sum_{t \in w} M(t, c_j) \otimes Deg\ (t)}{\sum_{t \in w} M(t, c_j) \oplus Deg\ (t)} \quad (3)$$

Where ⊗ and ⊕ denote the fuzzy conjunction and disjunction operators, respectively. These operators are replaced with an equation from table 4 and they are represented as t-norm, t-conorm respectively as [4] [18].

The t-norm value represents the mathematical value given in Table IV; which contains the membership value M (t_i, c_j) for each term in each of the categories. This is a representation of all documents in each category, and so we deal now with the higher level of abstraction that contains the categories level and not the documents level. The second part is the t-conorm (Deg(t_i)), in which its value is determined by measuring the relevance of each term in the test web page document to the term with the highest occurrence frequency. There should be one term (t_m) in which its degree Deg(t_m) equals 1, and the other terms which their frequencies less than the maximum is proportional to the maximum frequency of (t_m). Table V shows an example of the process of the degree membership values Deg(t_i).

E. A Special case of fuzzy similarity

The proposed method for representing the degree membership for the test web page document may contain more computation, which is considered to be time consuming and take more space for storage. The test document may be represented as Boolean features vector [4], in which the degree function (Deg(t_i) =1) for all the terms in the document. In this case, the similarity is calculated using formula 4:

$$sim(w, c_j) = \left(\frac{1}{Count(t_i)}\right) \sum_{t \in w} M(t, c_j) \quad (4)$$

$M(t,c_j)$ is the membership value in the training data, and Count(t_i) is the total number of terms in the test web page. We verified the correctness of formula 4 by applying it to the t-

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/15082/pdf

norm, and t-conorm of Table 4. The results were always give t-norm (x, y) = X and t-conorm (x, y) = 1.

TABLE IV
PAIRS OF T-NORM AND T-CONORM.

t-norm (x, y)	t-conorm (x, y)
Einstein Product: $\frac{x \cdot y}{2-[x+y-(x \cdot y)]}$	**Einstein Sum:** $\frac{x+y}{[1+x \cdot y]}$
Algebraic Product: $x \cdot y$	**Algebraic Sum:** $x+y-x \cdot y$
Hamacher Product: $\frac{x \cdot y}{[x+y-(x \cdot y)]}$	**Hamacher Sum:** $\frac{x+y-2x \cdot y}{1-(x \cdot y)}$
Minimum: $\min(x,y)$	**Maximum:** $\max(x,y)$
Bounded Difference: $\max(0,x+y-1)$	**Bounded Difference:** $\min(1,x+y)$

TABLE V
DEGREE MEMBERSHIP VALUE OF THE WEB PAGE TEST DOCUMENT.

Terms (t_i)	Term_Frequency	Deg(t_i)
نفس - tm	22	1
مرض	14	14/22 = 0.636
خطر	12	12/22 = 0.545
كثر	6	6/22 = 0.273
نبغ	1	1/22 = 0.045

F. The Classification Task

Depending on the value of sim(w,c_j), we repeated the same calculation for the next category with the same document and the next category till the last one. Then the category of the test web page p is the one that represents its contents by selecting the largest value of the similarity outputs, i.e. Cat(p) = MAX(sim(w,c_1), sim(w,c_2), . . . sim(w,c_n)).

III. EXPERIMENTAL RESULTS

First, the training data was collected from different sources, and then the different stages of the six fuzzy similarity approaches were applied to these data. The empirical analysis was performed on 50 standard well-known Arabic web pages, 5 web pages for each category. These categories are: Autobiography (1), Children's Stories (2), Economics (3), Health and Medicine (4), Interviews (5), Religion (6), Science (7), Short Stories (8), Sociology (9), and Tourist and Travel (10). The six fuzzy operators were analyzed according to the number of documents classified correctly and the CPU time taken by each operator to classify the documents.

Fig. 2 and Fig. 3 below show the results after analyzing the six methods according to the number of documents classified correctly. From these figures we can conclude that the algorithms performed differently for all categories depending on their precision, the category itself, or the whole test data. Einstein measure gave best performance among the other methods and then the Bounded measure followed by Algebraic measure and ScFussy (Special case fuzzy).

From Fig. 3, we can conclude that the algorithms perform differently for all categories depending on their precision, the category itself, or the whole test data set. Einstein measure was gave best performance among the other measures and then the Bounded measure followed by Algebraic measure.

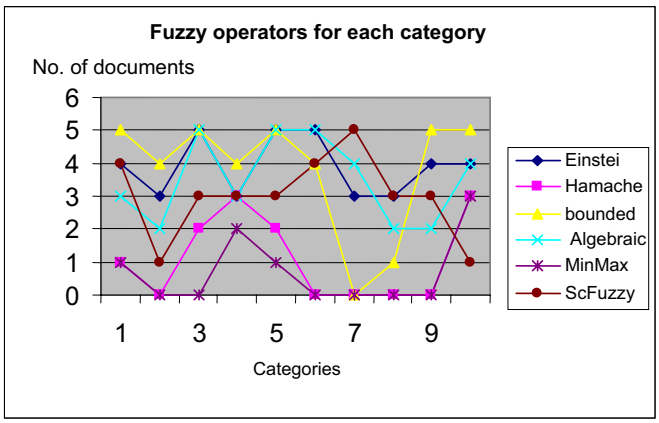


Fig. 2: Measures Performance

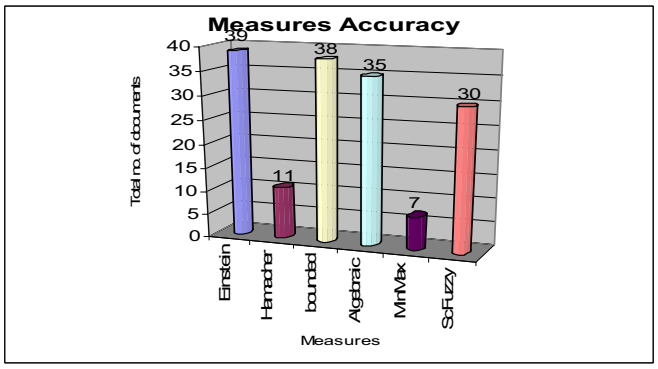


Fig. 3: Measures Accuracy

In order to investigate the effectiveness of the proposed approaches, we also analyzed these methods in terms of the CPU time. Fig. 4 shows the results after analyzing the six methods according to the average time taken by each method to classify the whole documents (time average for every category). Fig. 5 shows the CPU time analysis time average for every approach, i.e. time taken by each method to classify the whole documents (all categories). From these figures we can conclude that the six operators were ordered according to the CPU time as follows (best first): MinMax, Algebric, Bounded, Hamcher, ScFuzzy and Einestien.

Finally, we concluded after the completion of this research that if the number of documents classified correctly is the only criterion to be taken into account then Einstein, Bounded, and

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/15082/pdf

Algebraic methods are the best among the other methods. If the CPU time is important then MinMax, Algebric, and Bounded are the best among the others.

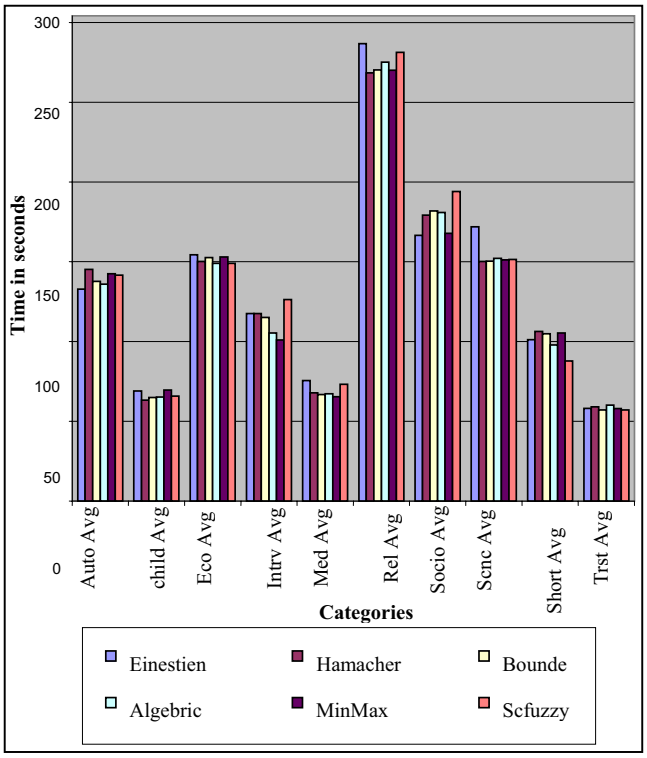


Fig. 4: Time average for every category

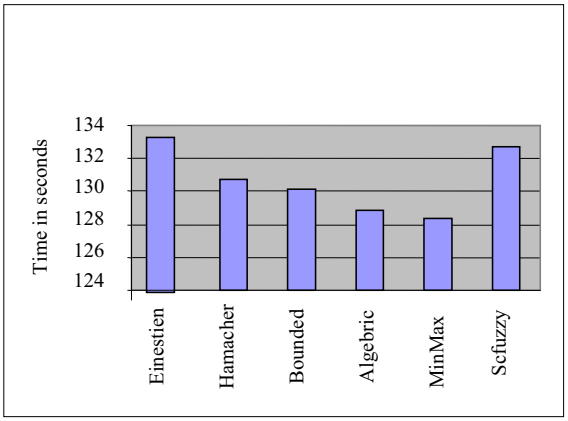


Fig. 5: Time average for every approach

IV. Conclusions And Future Work

We have presented in this paper a fuzzy similarity approach for Arabic web page classification. The approach used fuzzy term-category relation by manipulating membership degree for the training data and the degree value for a test web page. We used and compared six measures in this study. These measures are: Einstein, Hamacher, bounded difference, Algebraic, MinMax, and Special case fuzzy (Scfuzzy). The best performance is achieved by the Einstein measure then the Bounded measure followed by Algebraic measure.

The training data is first collected from different sources, and then normalized by passing it through the noise elimination module. The approach also includes the HTML stripping, stop word removing, and stemming. The learning process began by representing terms as numbers to reduce their representation. The final step in the process was to apply the six measures to the web pages.

Future work will consider the use of hyperlinks embedded in each web page to some depth and find out the synonyms of their text terms, i.e. classifying pages depending on their hyperlinks, in which each web page is categorized based on the group of web pages that it refers to, and recursively get the category label with the most proposed one.

References

[1] Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, Wei-Ying Ma, "Web-page Classification through Summarization", *Proceedings of the ACM SIGIR 04,* July 25–29, 2004, Sheffield, South York Shire, UK.

[2] H. Chen, S. T. Dumais, "Bringing order to the Web: Automatically categorizing search results", *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'00)*, ACM pp. 145-152, 2000.

[3] Michie, D., Spiegelhalter, D.J., Taylor, C.C., Machine Learning, Neural and Statistical Classification, Ellis Horwood, London, 1994.

[4] D. H. Widyantoro, J. Yen, "A Fuzzy Similarity Approach in Text Classification Task", *Proceedings of Ninth IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2000)*, pp. 653-658, San Antonio, Texas, May 2000.

[5] Ahmad T. A-Taani, Noor Aldeen K Al-Awad, "A Comparative Study of Web-pages Classification Methods Using Fuzzy Operators Applied to Arabic Web-pages",PWASET, vol. 7, pp. 33-35, 2005.

[6] Hui Yang, Tat-Seng Chua, "Effectiveness of Web Page Classification on Finding List Answers", *Proceedings of the ACM SIGIR 04*, July 25–29, 2004, Sheffield, South York Shire, UK.

[7] Stephanie W. Haas, Erika S. Grams, "Page and Link Classifications: Connecting Diverse Resources", *Proceedings of the ACM*, pp. 99-107, Digital Libraries 1998, Pittsburgh PA USA.

[8] Michelangelo Ceci, Donato Malerba, "Hierarchical Classification of HTML Documents with WebClassII", In: F. Sebastiani (Ed.): ECIR 2003, LNCS 2633, pp. 57-72, 2003.

[9] Rongbo Du, Rei Safavi-Naini and Willy Susilo, "Web Filtering Using Text Classification", *Proceedings of the 11th IEEE International Conference on Network (ICON 2003)*, pp. 325-330, 2003.

[10] Lawrence Kai Shih, David R. Karger, "Using URLs and Table Layout for Web Classification Tasks", *Proceedings of the WWW2004*, May 17–22, 2004, pp. 193-202, New York, USA.

[11] Eric J. Glover1, Kostas Tsioutsiouliklis, Steve Lawrence, David M. Pennock, Gary W. Flake, "Using Web Structure for Classifying and Describing Web Pages", *Proceedings of the* WWW2002, May 7–11, 2002, pp. 562-569, Honolulu, Hawaii, USA.

[12] Gongde Guo, Hui Wang, David A. Bell, Yaxin Bi, Kieran Greer, "An kNN Model-Based Approach and Its Application in Text Categorization, *Proceedings of the 5th International Conference (CICLing 2004)*, Seoul, Korea, February 15-21, 2004, pp. 559-570.

[13] Anders Ardö, DTV, Lyngby, Denmark Traugott Koch, NetLab, Lund, Sweden, "Automatic classification applied to the full-text Internet documents in a robot-generated subject index*", Proceedings of the 23rd International Online Information Meeting*, London, 7-9 Dec 1999, pp. 239-246.

[14] Aijun An, Yanhui Huang, Xiangji Huang, Nick Cercone, "Feature Selection with Rough Sets for Web Page Classification", In: Transactions on Rough Sets II: Rough Sets and Fuzzy Sets, James F. Peters, Andrzej Skowron, Didier Dubois, Jerzy W. Grzymała-Busse, Masahiro Inuiguchi, Lech Polkowski (Editors), 2004.

[15] J. A. Roubos, M. Setnes, J. Abonyi, "Learning fuzzy classification rules from data", In: Developments in Soft Computing, John R, Birkenhead R., (Editors), Springer-Verlag, Berlin/Heidelberg, pp.108-115, 2001.

[16] Heiner Stuckenschmidt, Jens Hartmann,Frank van Harmelen, "Learning Structural Classification Rules for Web page Categorization", *Proceedings of FLAIRS 2002, special track on Semantic Web, S. Haller , G. Simmons (Editors)*..

[17] Sarah Zelikovitz, Haym Hirsh, "Improving Short-Text Classification using Unlabeled Background Knowledge to Assess Document Similarity", *Proceedings of the Seventeenth International Conference on Machine* Learning (ICML-2000), Morgan Kaufmann Publishers.

[18] Włodzisław Duch, "Similarity-based methods: a general framework for classification, approximation and association", Control and Cybernetics, vol. 29 (2000), Grudzia¸dzka, Toru´n, Poland.



Ahmad T. Al-Taani was born in Jordan in 1962. He received his Bachelor of Science degree in computer science in 1985 from Yarmouk University, Jordan. He received his Master of Science degree in software engineering from National University, San Diego, California, U.S.A., in 1988. He received his Doctor of Philosophy degree in computer vision from University of Dundee, U.K., in 1994.

Currently (2008), he is the chairman of the department of computer sciences at Yarmouk University. He works with the department of computer sciences at Yarmouk University, Jordan, since 1995 and he chaired the department from 1998 to 2000. He was with the department of computer systems, Hail Community College, King Fahd University of Petroleum and Minerals, Saudi Arabia, from September 2000 to June 2003 and he chaired the department for two years. In 1991 he got a three years joint scholarship from Yarmouk University and the British Council to study for the PhD. In 1999 he got a research fellowship from the DAAD and Yarmouk University for three months at University of Kiel in Germany. He has publications in the national and international journals and conference proceedings. He has also supervised a number of graduate students. His research interests are in artificial intelligence applications such as in the fields of Image Processing, Arabic Character Recognition, Direct Machine Translation, Arabic Web Page Classification, and Heuristic Search.

Dr. Al-Taani is a member in World Academy of Science, Engineering and Technology and in the Jordanian Society for Mathematical Sciences. He is a member of the Editorial Boards for many refereed journals. He is also a referee for many different national and international conferences.



Noor Aldeen K. Alawad was born in Jordan in 1981. He received his Bachelor of Science degree in computer science from Yarmouk University, Jordan in 2003. He received his Master of Information Technology degree in computer science from Yarmouk University, Jordan in 2005. His research interests include: fuzzy logic, Arabic web page classification, Arabic natural language understanding, data mining and web mining.

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/15082/pdf