Modern Applied Science; Vol. 9, No. 1; 2015
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Presenting a Fuzzy System for Identifying Persian Advertising Websites

Hamed Jelodar[1], Seyed Javad Mirabedini[2] & Ali Haroonabadi[2]

[1] Department of Computer Software, Bushehr branch, Islamic Azad University, Bushehr, Iran

[2] Department of Computer Software, Central Tehran Branch, Islamic Azad University, Tehran, Iran

Correspondence: Hamed Jelodar, Department of Computer Software, Bushehr branch, Islamic Azad University, Bushehr, Iran. E-mail: jelodarh@gmail.com

Received: October 7, 2014 Accepted: October 14, 2014 Online Published: December 2, 2014

doi:10.5539/mas.v9n1p129 URL: http://dx.doi.org/10.5539/mas.v9n1p129

Abstract

Online advertisement is a cheap and powerful tool which has targeted internet users. At the moment there is a multibillion dollar market for online advertising which is the main income of some popular websites. Some people use this cheap tool to achieve their personal goals i.e. they use incorrect advertisement techniques. For instance a user refers to Bing search engine and tries to search "Photoshop software download". The question is whether the returned results are related to the searched sentence or not. Unfortunately, there are some advertisement websites which use phony techniques (such as using fake key words) to attract users to their websites. Consequently, the user is not able to find his desired webpage and lose his time. In this paper Persian websites are investigated. A fuzzy system is proposed which performs identification and analysis of websites using two parameters; "url feature" and "number of important key words".

Keywords: fuzzy system, advertisement websites, advertisement web detection

1. Introduction

Online advertisement provides advertisers with a powerful mechanism to effectively target users. The advertisements might be customized based on browsing behavior of users, geographical location and personal interests. At the moment there is a multibillion dollar market for advertisement. It might be said that online advertisement is similar to commercials in TV channels. The users may encounter advertisement websites while surfing the net; however, they seek for efficient and correct searching and they are not really interested in such websites. In such circumstances, separating advertisement websites and actual websites will considerably increase efficient and correct search. Some researchers believe that using URL feature parameter facilitates identification of these websites; nevertheless, some others believe that image sizes and keywords must be also considered. Indentifying and separating disturbing advertisement websites and actual websites play significant role in improvement of search efficiency and time saving. This article consists of four sections.

1.1 Related Work

Some of web designers and network managers use illegal search engine optimization techniques for marketing in order to increase their website ranking and deceive search engines. Some Arabic web pages use content and link features to increase their rank in search engines. As Mr. Alkabi says, Arab users suffer from two main problems. First, the low percent of Arabic text in the web; second, Arabic web spam which cause search engines to find unrelated pages. Success of spamming techniques used in deceiving search engines threatens credit of the search engine. Moreover, they have other consequences such as wasting users' time and preventing them from obtaining required data. To address these mentioned issues they proposed an integrated system to decrease content web spam and Arabic link and to filter undesired Arabic web pages from search engines. It is called online Arabic web spam identification system. They detect these pages using size of images, size of webpage and number of internal links (AlKabi et al , 2014).

Since advertisement web pages are a type of spam web, this issue is also investigated in our research. There are numerous web pages including various types of content including text, audio, video and multimedia. These web pages have attracted more users to internet and search engines to exploit different interaction characteristics such as internet shopping, email, chat and media downloads. Vahesh et al have analyzed Arabic web spam in their article. They analyzed the spam behaviors using 10 common Arabic keywords. The keywords include “chat”, “game”, “Youtube”. “Facebook”, “University”, “association”, “music”, “photo”, “billiard”, “joy”. Furthermore, they collected a set of Arabic spam data which were mainly analyzed with content based characteristics. Afterward, they evaluated spam behaviors by popular Arabic keywords utilizing a tree classification algorithm. They succeeded to detect Arabic web spam with 90% precision (Wahsheh et al., 2012).

The number of advertisement web pages in the internet is increasing drastically; whereas, the major part of users are not willing to visit advertisement pages. Roud et al have analyzed seven giant advertisement companies to achieve automatic censor of ads. Other products avoid pop-up windows which are usually ads. In their definition advertisement means exciting people so that make them buy something. They considered various parameters to analyze advertisement websites. Image size is a reliable clue for detection of ads. Particularly, they considered 60*480, 150*500, 120*600 and 160*600 advertisement banners in the page borders. According to their research this criterion achieves 95% precision in detection of advertisements. Other criteria consist of words in the URL and captions of images. They used a number of keywords to detect text advertisement among which imperative words such as “buy”, “receive”, “sign in/join”, “click” might be mentioned (Rowe, 2002).

Mr. Stone and colleagues have checked the features of website advertising. According to the definition of publishers or sellers, publishers by hosting websites with advertisers will earn money. In general, publishers attract more and more viewers to their websites, will earn more money. Given that their definition from advertisers, they will pay to advertisement network to display their ads on websites, whenever their ads are shown they have to s pay to ad network and should give a percentage to their publishers. Other ad operations are costs based on click and costs based on action. Only gives money to the publisher when a user Click on an ordered advertisement .the cost based on action says that, only the publisher get money when a user clicks on an ad and perform certain actions, for example start to fill a form of a page. They also addressed the fake advertising website and according to surveys carried out manually, have achieved models and common feature between these ads websites (Stone et al., 2011).

1.2 Disturbing or Useful Online Advertisement

Nowadays, it rarely happens that a user does not face advertisement banners visiting each website. Web advertisement has become a billion dollar business. Comparing to traditional media, online advertisement is more proper and cost efficient choice. It is easy to create an account with an advertisement provider such as Double Click and rapidly send marketing messages to a huge population. Unfortunately, this useful feature has been abused. Hackers and other people have come to the conclusion that web advertisement is a low-cost and effective tool for false activities and cheating. Ads provide the advertisers with powerful mechanisms to target users. Ads might be customized with respect to user's browsing behavior, geographical location and personal interests. At the moment there is a multi-billion dollar market in online advertisement which provides income for many popular websites. The advantage of World Wide Web as a large digital library with precise and reliable data is reduced by large amount of ads in the websites; however, nobody is obliged to read and see the ads and the censorship might be imposed using automatic software. We aim at detecting Persian advertisement web pages. Ads are defined as information encourages a person to buy something. Obviously, some criteria are required to detect these websites. One of the most common criterion is keywords such as “advertisement”, “buy”, “shop”, “free”, “join”, “click”, “now” (Wang et al , 2011 ; Li et al ,2012 ; Krammer ,2008 ; Stone-Gross .et al , 2011) .

2. Methodology

2.1 Problem Definition

Internet websites include various topics. Some of the websites are dedicated to advertisement which are used for marketing. Nowadays, some of these websites use fake techniques (such as using keywords) to increase their visit and attract more users. For example a user searches “English to Persian Dictionary mobile software”; the question is whether the returned results are related to search item or not. Unfortunately the answer is no because some of the provided links are not related to the search topic. They are displayed as they have included some keywords in their website to deceive the search engine and attract the users. Consequently, performing an efficient search and returning correct results is the most prominent request of users. In this article Persian websites are investigated to separate advertisement web pages.

2.2 Proposed Approach

In this paper a fuzzy system based on two inputs is proposed. The results of our experiments demonstrate the acceptable performance of proposed method.

2.2.1 Fuzzy System

Fuzzy systems are able to utilize expert's knowledge to make decision and control a system. The most popular use of them is modeling complicated environments or any other situation where a clear model of the system is not available. It relies on some inputs and their corresponding outputs to make future decisions. It is difficult to know the reasons of test technique efficiency.

2.2.2 Input-Output Parameters of Fuzzy Systems

As mentioned before in this system URL features and number of important keywords (which are explained later) are used as inputs.

- Important advertisement keywords:

By important advertisement keywords, we mean the words which are widely used in advertisement websites. In the proposed method the websites are investigated to find the number of keywords which match these important ones. In this way, the fuzzy system would be able to calculate probability of advertisement website.

- URL features:

Advertisement websites usually have domains which include concepts such as buy, sell and so on. Our survey illustrated that the importance of advertisement domains is not the same; thus, different scoring techniques are utilized so that the importance of each keywords of the domain are considered in the calculations.

In the desired fuzzy system the performance of mentioned factors in determining advertisement websites is assessed based on the inputs. It must be noticed that there are other factors affecting identification of advertisement websites; nonetheless, here merely two factors are considered in the fuzzy system. The aforementioned fuzzy system has one output which demonstrates probability of being an advertisement website based on different states of the inputs.

To evaluate performance of test technique FIS tool in MATLAB software is utilized. It is depicted in figure 1.

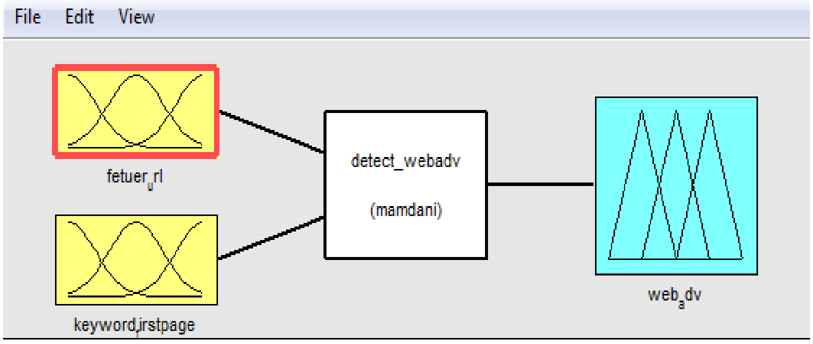


Figure 1. Generic model of fuzzy expert system

This system includes two input fields associated with effective factors in ads detection. For each factor three classes of linguistic variables are considered; low, normal and high. There is also one output field showing the probability of being an advertisement website. The output is divided into five classes to which five different linguistic variables (very low, low, normal, high, very high) are assigned. Figures 2 and 3 demonstrate membership functions for input and output parameters. All membership functions are triangle ones. The generic structure of fuzzy system and its complete characteristics are shown in figure 4.Used fuzzy rules has been shown in table 1.

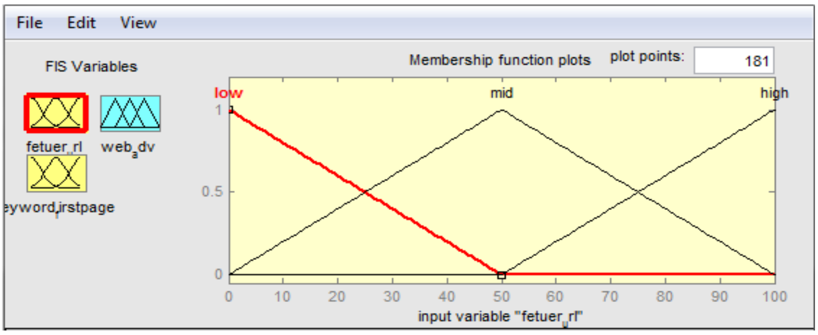


Figure 2. Membership function of URL features

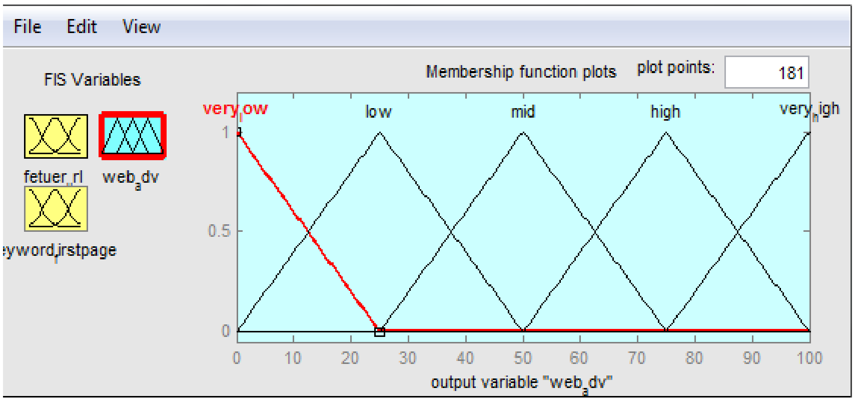
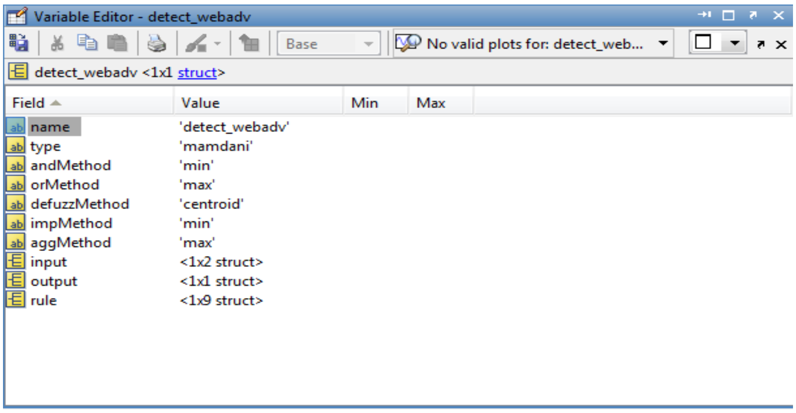


Figure 3. Output membership function


Variable Editor - detect_webadv
Base
No valid plots for: detect_web...

detect_webadv <1x1 struct>

Field	Value	Min	Max
name	'detect_webadv'		
type	'mamdani'		
andMethod	'min'		
orMethod	'max'		
defuzzMethod	'centroid'		
impMethod	'min'		
aggMethod	'max'		
input	<1x2 struct>		
output	<1x1 struct>		
rule	<1x9 struct>		

Figure 4. Generic structure of fuzzy system

Table 1. Used fuzzy rules in this system

	URL=LOW	URL=MID	URL=HIGH
KEY WORD = LOW	Very low	Very low	Low
KEY WORD = MID	low	low	high
KEY WORD = HIGH	mid	mid	Very high

3. Simulation and Results

FIS in MATLAB software is exploited to implement fuzzy system. It is considered for simulating a set of input parameters. The desired parameters consist of:

❖ Advertisement keywords investigated in this paper include 9 keywords: buy, advertising, free, free ads, Toman, Rial, winner and gift. They constitute the input of the system.

- ❖ Advertisement domains include advertisement keywords. To show the importance of each keyword different scores are assigned to them which are presented in table **2**.

Table 2. Scores assigned to advertisement keywords

No.	Keyword	score
1	agahi	90
2	niaz	85
3	shop	60
4	bazar	40
5	forush	90
6	tabligh	75
7	market	70
8	kala	30
9	payam	30

In this experiment the results of Bing search engine are utilized. 50 questions with different topics are sent to the search engine. Then, some of returned links are analyzed. Table 3 includes utilized questions.

Table 3. The used questions

ADVERTISEMENT	NON-ADVERTISEMENT	NUMBER OF RESUTLS	SEARCH ENGINE	SEARCH TEXT (PERSIAN)	QUESTION
15	**14**	**29**	**Bing**	"دانلود ویدیوهای "PluralSight	**Q1**
3	**13**	**16**	**Bing**	دانلود فایل تقویم 1393"	**Q2**
6	**24**	**30**	**Bing**	"دانلود رایگان فیلم آموزش قالی بافی"	**Q3**
4	**14**	**18**	**Bing**	"دانلود نرم افزار فوتوشاپ	**Q4**
4	**23**	**27**	**Bing**	دانلود oxford elementary "dictionary	**Q5**
…	**…**	**…**	**…**	**…**	**…**
…	**…**	**…**	**…**	**…**	**…**
4	**26**	**30**	**Bing**	"دانلود رایگان نرم افزار حقوق و دستمزد"	**Q48**
2	**10**	**12**	**Bing**	دانلود آموزش تعمیر خودرو پراید"	**Q49**
2	**20**	**22**	**Bing**	"دانلود رایگان کتاب آموزش پرورش قارچ"	**Q50**

The results obtained from 3 experiments (Q1, Q2, Q3) out of 50 performed experiments are presented here. It is noteworthy that in these experiments websites which are located in high and very high range of the fuzzy output are considered as advertisement websites.

- Results of Q1 based on the question “downloading Plural Sight Videos”:

As can be seen in table 2, in Q1 15 links among 29 returned links are introduced as advertisement websites. Now the 29 introduced links are analyzed using the proposed fuzzy system.

Table 4. Output of the first experiment

Input			Result
index	Feature URL	Keyword	
1	0	0	8.0000
2	0	4	8.0714
3	0	21	9.1021
4	85	94	88.7952
5	0	4	8.0714
6	0	21	9.1021
7	0	43	8.1798
8	70	190	84.3226
9	70	87	77.0000
10	90	121	90.1021
11	0	5	8.1045
12	0	51	9.0875
13	0	34	8.7059
…	. . .	. . .	. . .
25	75	121	82.222
26	0	8	8.222
27	0	8	8.222
28	85	165	95.6543
29	90	89	87229

The results revealed that fuzzy system makes mistake in only 4 links among 29 links. In other words, the accuracy rate of the fuzzy system is 86%.

- Q2 experiment,”1393 calendar file download”:

As can be seen in table 2, in Q2 3 links among 16 returned links are introduced as advertisement websites. Now the 16 introduced links are analyzed using the proposed fuzzy system.

Table 5. The output of second experiment

Input			Result
index	Feature URL	Keyword	
1	0	21	9.1021
2	0	6	2.0714
3	20	48	23.0247
4	0	39	9.7952
5	10	58	13.7567
6	0	43	9.1021
7	85	143	86.1798
8	70	190	92.3226
9	30	87	42.0000
10	0	121	48.1021
11	0	0	8.0000
12	70	351	89.0875
13	0	34	8.7059
…	. . .	. . .	. . .
16	0	8	8.222

The experiment results demonstrate that the fuzzy system has no error in all 16 links and determined 3 advertisement websites correctly i.e. the accuracy rate is 100%.

- Q3 experiment “salary software free download”

As can be seen in table 2, in Q3 4 links among 30 returned links are introduced as advertisement websites. Now the 30 introduced links are analyzed using the proposed fuzzy system.

Table 6. The output of third experiment

Input			Result
index	Feature URL	Keyword	
1	*0*	*39*	*19.7952*
2	*0*	*43*	*22.1021*
3	*0*	*21*	*9.1021*
4	*90*	*240*	*97.7952*
5	*0*	*6*	*2.0714*
6	*0*	*121*	*48.1021*
7	*0*	*43*	*22.3453*
8	*85*	*283*	*94.3226*
9	*0*	*87*	*49.0750*
10	*0*	*121*	*49.1021*
11	*70*	*95*	*91.1045*
12	*70*	*120*	*89.0875*
…	. . .	. . .	. . .
30	*0*	*5*	*8.1045*

The experiment results illustrate that the fuzzy system has no error in all 30 links and determined 4 advertisement websites correctly i.e. the accuracy rate is 100%. in the following figures the effects of each input on advertisement website determination are shown.

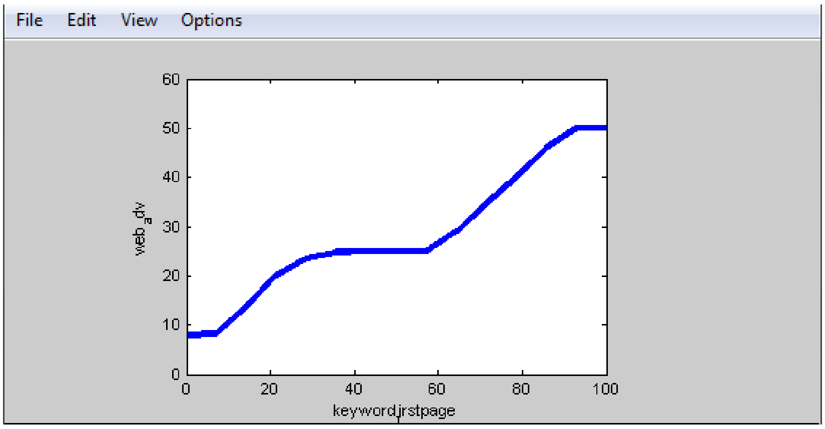


Figure 5. The effect of advertisement keywords input

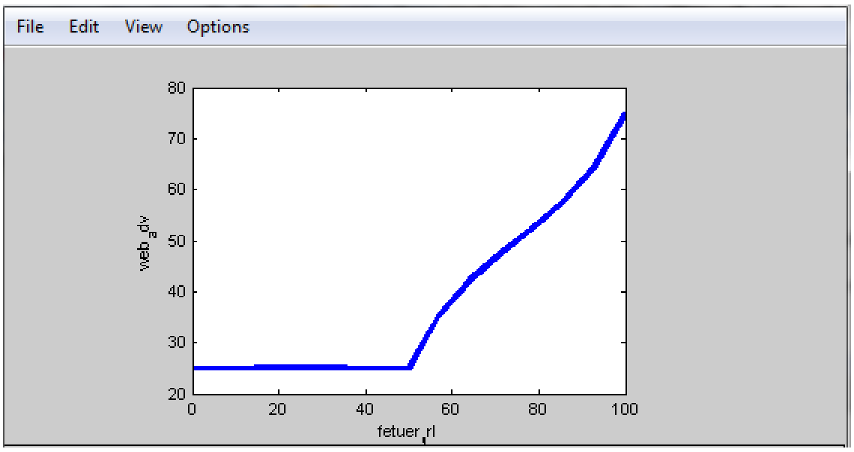


Figure 6. The effect of website domain input

4. Discussion and Conclusion

The target of Internet advertisers is attracting the users to the advertiser's website or the introduction of a brand by placing promotional content and links on Web sites. Internet advertising such as commercials is in television and Most Internet users when suddenly encounter with these ads will become upset. It seems that Displaying the advertisement on the user's computer, can be seen as a cost that user pays for deals content. Therefore by blocking these ads, the bandwidth and cost for users will be saved. The paper focused on detection of Persian advertisement websites. An approach was proposed to optimize the search. A fuzzy system was proposed for this purpose and a series of experiments were conducted to prove its performance. The results revealed that our designed system has acceptable performance and it is practically efficient. In the future one problem of above article will be study that it could be suitable idea for future research. Insufficient number of input parameters which it means which it means in this article we use from two parameters however, we could also use effective parameters in order to identify advertising website For example parameters such as image descriptions, internal link and more. If there will be more number of input parameters of fuzzy system we can have a more accurate diagnosis

References

AlKabi, M. N., Wahsheh, H. A., & Alsmadi, I. M. (2014). OLAWSDS: An Online Arabic Web Spam Detection System. *International Journal of Advanced Computer Science & Applications, 5*(2), 105-110. http://dx.doi.org/10.14569/ijacsa.2014.050216

Krammer, V. (2008, October). An effective defense against intrusive web advertising. In Privacy, Security and Trust, 2008. PST'08. *Sixth Annual Conference on*, 3-14. http://dx.doi.org/10.1109/PST.2008.10

Li, Z., Zhang, K., Xie, Y., Yu, F., & Wang, X. (2012, October). Knowing your enemy: understanding and detecting malicious web advertising. *In Proceedings of the 2012 ACM conference on Computer and Communications Security,* 674-686. http://dx.doi.org/10.1145/2382196.2382267

Rowe, N. C., Coffman, J., Degirmenci, Y., Hall, S., Lee, S., & Williams, C. (2002, July). Automatic removal of advertising from web-page display. In *JCDL, 2*, 406-406. http://dx.doi.org/10.1145/544220.544354

Stone-Gross, B., Stevens, R., Zarras, A., Kemmerer, R., Kruegel, C., & Vigna, G. (2011, November). Understanding fraudulent activities in online ad exchanges. *In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference,* 279-294. http://dx.doi.org/10.1145/2068816.2068843

Wahsheh, H., Alsmadi, I., & Al-Kabi, M. (2012). Analyzing the popular words to evaluate spam in arabic web pages. *IJJ: The Research Bulletin of JORDAN ACM–ISWSA, 2*(2), 22-26.

Wang, Y., Burgener, D., Kuzmanovic, A., & Maciá-Fernández, G. (2011, June). Understanding the network and user-targeting properties of web advertising networks. In Distributed Computing Systems (ICDCS). *2011 31st International Conference on*, 613-622. http://dx.doi.org/10.1109/ICDCS.2011.10

Copyrights