A Comparative Study on the Effect of used Crossover Operator on Performance of GA as a Web Page Classifier

Neda Sabouri
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

Hamid Haj Seyyed Javadi
Department of Mathematics and Computer Science, Shahed University, Tehran, Iran

Toktam Zhiani Asoudeh
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

ABSTRACT

By incredible and uncontrollable growth in amount of web pages on the World Wide Web, providing an infrastructure due to searching among them leads to appearance of topic specific crawling of the web. Process of focused crawlers is base on an automatic web page classification mechanism of belonging or not belonging the page to a particular topic. The Genetic algorithm (GA) is a common optimization and search technique, used as classifier of web pages. Crossover operation as one of the GA operators, by producing 2 children out of parents of past generation and determination of next generation through combining produced child, plays important roles in performance of this algorithm. Up to now many different crossover operators such as single-point, two-point and ring are presented. In this paper, we compare the effect of mentioned crossover operators on performance of GA algorithm as a web page classifier.

General Terms

Specific crawling, focused crawler, high performance

Keywords

Genetic algorithm, web mining, Crossover Operator, Classification

1. INTRODUCTION

The incredible increase in the amount of information on the World Wide Web converts it to an environment with potential to extract useful information and causes using tools such as web mining for having easy and quick access to pages related user's specific topic.

For this purpose, classification as one of web mining techniques has been used for determining belonging or not belonging the web page to user's specific topic automatically.

GA as an evolutionary algorithm, improves response of problem randomly but purposeful and is used as web pages classifier [1].

Application of GA to web page classification starts with Riberio et al [2] who has proposed a web page classifier of "if condition then class i" and applied fuzzy classification.

Pietramala et al [3] have introduced a GA, called Olex-GA, for the induction of rules of the form:

If a document like d, has at least one of retrieved words of positive pages whereas not include none of the terms retrieved from negative pages then d will belong to positive class else it will be placed in negative class.

Ozel et al [4] in 2010 , has developed a GA based automatic web page classification system which uses both html tags and terms belong to each tag as classification features and learns optimal classifier from the positive and negative web pages in the training dataset.

Because of significant role of the crossover operator in genetic algorithm performance [5], So far much research has been done in comparing the effects of different crossover operators.

Kellegoz and Toklu [6] have used GA for solving job scheduling problem and reducing waiting time and compared the performance of proposed algorithm with 11 different crossover operators such as single-point, two-point.

Kaya [7], in 2011 has presented two new crossover operators, sequential and random mixed and survived their performance in Solving problems in the concrete construction industry and compared with crossover operators such as single-point. In the same year He has proposed an other crossover called ring [8] and has been tested by a number of test functions with various level of difficulty and compared with single-point, two-point, heuristic crossover operators.

Chinassri [9] in 2012 for scheduling the university courses has used GA with 3 different crossover operators: cycle, order and partially matched.

In this study, the role of different crossover operators (single-point, two-point and ring), is used by GA for classification web pages and presented by Ozel et al [4], is investigated. So Tests are conducted on two different collections which consist of related course and project homepages from Cornell, Texas, Washington and Wisconsin universities in computer science as well as some irrelevant pages from them. The experiments indicated that using two-point crossover operator has higher accuracy and F-measure value and thus higher performance than the other two operators.

This paper is organized as follows:

In section 2 we explain the method of Feature extraction. Manner of document vector creation is provided in section 3. Structure of classifier, single-point, two-point and ring crossover operators are described in section 4. In section 5 used dataset and experimental results and discussion on the results are presented and finally conclusion of study is provided in section 6.

It is again emphasized that classification of web pages by GA is used in this study, is completely based on Ozel et al [4] presented algorithm.

2. Feature Extraction

After evaluation of all train web pages in learning process, since the increasing features, a lot of problems such as high execution time, are occurred and with respect to this issue of important words basically appear under Header, Anchor, Bold, Italic, Emphasize, Strong, Paragraph tags [2,11,12], only the words in above tags are extracted and by type classified into categories Title, Bold, Anchor, List, Paragraph and Header. <i>, , <b>, are located in Bold category, <h_1>, <h_2>,… in Header, <Title> in Title, <Li> in List , <a> in Anchor and <p> in paragraph categories.

Afterwards in order to reduce the number of features, ampersand and stop words are removed and by using porter's algorithm the word will be replaced by its stemmed [13].

Finally (Tag, Term) pairs will be considered as features of algorithm. If the tags are nested, their shared words will be considered as distinct features [4].

3. Creation of Training Pages Vectors

Provided classification algorithm is used for determining belonging or not belonging the page to a particular topic.

Related and unrelated pages to the topic, respectively, are called positive and negative.

Each of pages in positive and negative classes is classified into two parts: training and testing pages.

After evaluation of positive and negative pages and extraction of their features, for each page, document vector is constructed and represented as [4]:

$$\vec{D} = (d_{1N1},\ldots,d_{1Nn},d_{2N1},\ldots,d_{2Nn},\ldots,d_{MN1},\ldots,d_{MNn})$$

Where d_{ij} denotes the counting occurrences of term i in tag j.

Then each document vector is normalized by dividing sum of all values in the vector to its maximum value. So the values in a document vector lie between 0 and 1 .

4. Genetic Algorithm

Genetic algorithm consists of 4 main parts [4]:

(i) Determination of chromosome structure, (ii) Generation of initial population, (iii) Evaluation of a population, reproduction (crossover, mutation, determination of new generation) and (iv) repeating 3 and 4 steps number of generation (gen-size) times to achieve an optima classifier.

4.1 Determination of Chromosome Structure

Each chromosome as a point in the search space and a possible solution for problem is representing as [4]:

$$\vec{C} = (c_{1N1},\ldots, c_{1Nn}, c_{2N1},\ldots,c_{2Nn},\ldots,c_{MN1},\ldots, c_{MNn})$$

Where c_{ij}, real number in the range [0,1], denotes the weight of term i in tag j.

4.2 Generation of initial population

To start this algorithm, we should randomly generate pop-size (number of chromosomes in the population) chromosomes [4].

4.3 Determination of fitness function

Since in classification based on GA the goal is to find the best chromosome, with respect to the fitness value, and using it as classifier, a function must be provided to determine the fitness of each existing chromosome.

Used fitness function acts based on cosine similarity [14] and contingency matrix [2, 15] (given in Table 1) [4].

		Actual class	
		positive	**negative**
Predicted class	**Positive**	True positive	False positive
	Negative	False negative	True negative

Table 1. Contingency matrix for binary classification system

The cosine similarity between a chromosome vector ($\vec{c}$) and a document vector ($\vec{d}$), represents the degree of similarity between them and is shown as sim($\vec{c}$, $\vec{d}$) .

$$\mathrm{Sim}(C,D) = \frac{\vec{C}\,.\,\vec{D}}{|\vec{C}|\times|\vec{D}|} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} c_{ij} \times d_{ij}}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} c_{ij}^2}\times\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} d_{ij}^2}} \quad (1)$$

$\vec{c}$ and $\vec{d}$ are normalized , so since $c_{ij} \geq 0$ and $d_{ij} \geq 0$, sim($\vec{c}$, $\vec{d}$) varies from 0 to 1.

To evaluate fitness of a chromosome, the sim values of the chromosome and all documents in training data set are computed. A threshold value is determined by taking the difference between the average and the standard deviation of computed sims.

The fitness evaluation of a chromosome is performs in four steps [4]:

1) sim of that chromosome to each of web page document in training dataset is computed.

2) The difference between the average and standard deviation of computed sims is taken as a threshold value.

It should be mentioned that a number of threshold values were used and is observed that the chosen threshold equation performs well for this GA.

3) If the sim between a web page and the chromosome is greater than obtained threshold value, the web page is classified as a positive document otherwise it's labeled as negative.

4) Fitness value of the chromosome is determined by [2]:

$$\text{Fitness} = TPR \times TNR \quad (2)$$

Where True-Positive-Rate and True-Negative-Rate values are computed as (based on contingency matrix):

$$\text{TPR} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

$$\text{TNR} = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad (4)$$

According to Table 1, True Positives (True Negatives) mean the positive (negative) documents that were correctly labeled by the classifier and false positives (False Negatives) are the negative (positive) documents that were incorrectly labeled.

4.4 Reproduction

In this step, selection, crossover and mutation operators are used.

4.4.1 Selection

By this operator, some of existing chromosomes are chosen for reproduction. Among the different selection methods, a biased roulette wheel is used [4]. Base on this operator

(1) For each chromosome i :

1) probability p_i, is computed by :

$$P_i = \frac{Fi}{\sum_{i=1}^{pop-size} Fi} \quad (5)$$

Where F_i is the fitness value of chromosome i.

2) Cumulative probability C_i, is obtained by

$$C_i = \sum_{i=1}^{j} F_j \quad (6)$$

(2) A random number r, uniformly distributed in [0,1] is generated pop-size times and each time the kith chromosome is selected if $C_{k-1} < r \leq C_k$.

if $0 \leq r \leq C_1$, C_1 else if $C_{pop\text{-}size} \leq r \leq 1$ then $C_{pop\text{-}size}$ will be chosen.

4.4.2 Crossover

Crossover is a genetic operator that combines two chromosomes (parents) to produce a new chromosome, hoping to reach more appropriate chromosomes.

There are a number of different crossover operators can be used in GA.

In this study, single-point, two-point and ring crossover operators are investigated. For this purpose, first the two parental chromosomes are selected randomly among the chosen chromosomes at the selection step. A random number r, uniformly distributed in [0,1], is generated to determine whether combination should be occurred or not.

4.4.2.1 Single-point crossover

A random integer number (p) in (0, m×n) is generated, if $r \leq p(c)$ then selected chromosomes are spilt at p point and combined with each other. Otherwise no crossover is performed and the randomly chosen parents are taken as children without any change [4,6,7].

4.4.2.2 Two-point Crossover

Two random integer number p, q in (0, m×n) are generated. If $r \leq p(c)$ then selected chromosomes are split at p and q points and elements between them will be exchanged. Otherwise no crossover is performed and the randomly chosen parents are taken as children without any change [8].

4.4.2.3 Ring Crossover

A random integer number p in (0, 2× m×n) is generated. If $r \leq p(c)$, Firstly, two selected chromosomes are combined with a form of ring, then the first child is created by cutting the obtained ring of r point to the length of m×n and the other one is created by reversing the remaining genes. Otherwise no crossover is performed and the randomly chosen parents are taken as children without any change [9].

After applying each above crossover operation, the randomly chosen parent chromosomes are excluded from the chromosome list generated in the selection step and the crossover operation continues until the list becomes empty.

4.4.3 Mutation

After the crossover operation, all of resulting chromosomes are subjected to the mutation operation such that for each gene c_{ij} in the chromosome C, a random number r from the interval [0,1] is generated and the gene is mutated as follows [4]:

$$c'_{ij} = \begin{cases} \text{Random value in } [0,1] & \text{If } r \leq P(M) \\ c_{ij} & \text{Otherwise} \end{cases} \quad (7)$$

Where c′ij is the new value of the gene c_{ij} after the mutation, P(M) is the mutation probability whose value is determined experimentally in test step.

4.4.4 Evaluation the Fitness of resulting chromosomes

In this step fitness of resulting chromosomes (as new generation) are evaluated [4].

4.4.5 Determination of the new generation

All the chromosomes in new and previous generation are sorted according to their fitness values and the best pop-size chromosomes are selected as the next generation. Therefore the best chromosomes found in each generation are kept without changing through the solution process [4].

4.5 Classifier selection

The above processes are repeated gen-size (predefined parameter) times and the best chromosome, with respect to the fitness value, is returned to be used as the classifier [4].

5. Experimental results

In this section different crossover operators used by obtained classifier on test dataset and their effects on classification performance are tested.

5.1 Dataset

Used datasets consists of related course and project homepages from Cornell, Texas, Washington and Wisconsin universities in computer science as well as some irrelevant pages from them. These pages were obtained from the webkb [10] project (well-known and freeware datasets).

The goal is to classify above homepages to positive (relevant pages to above universities) and negative (irrelevant pages). So all pages are classified into 2 classes: Test and train where Number of pages in each class and count of features for each dataset, respectively, are given in Table 2 and Table 3.

Table 2. Number of documents in the datasets

Dataset	Class	Test	train	Total
Course	Course(positive)	72	158	1051
	Not course (negative)	246	575	
Project	Project (positive)	17	69	504
	Not project (negative)	104	314	

Table 3. Number of features for each dataset

Tag	Number of features	
	Course	**Project**
Title	502	268
A	6700	1857
P	2675	489
Header	2695	2714
Bold	641	479
Li	10856	5120
total	24069	10927

5.2 Determination of GA Parameters

Genetic algorithm parameters were determined experimentally such that gen-size = 500, pop-size = 30, P(C) = 0.8 and P(M) = 0.1 were good choices for this classifier. For each experiment, the learning phase was repeated 10 times and the average of results were reported.

5.3 Crossover operators comparisons

In this study for measuring and comparing the effect of different crossover operators on performance of the classifier, two most commonly values, accuracy and F-measure are used.

For this purpose , first statistical method $\chi 2$ is applied to determine the weight of each term in training homepages of each dataset (all positive and negative pages with N Number) where denotes the dependence of feature F to positive homepages class c_j [16].

$$\chi^2(F, Cj) = N \times \frac{(AD-CB)^2}{((A+C)\times(B+D)\times(A+B)\times(C+D))} \quad (8)$$

Where A, B, C, D values are computed base on Table 4.

Table 4. A, B, C, D variables calculation

	Documents contains F feature	**Documents without F feature**
Pages in c_j class	A	C
Other pages	B	D

5.3.1 Accuracy Comparison

The accuracy of the learned classifier with different crossover operators is computed as [2,4]:

$$\text{accuracy} = \frac{True\ Positives + True\ Negatives}{Positives + Negatives} \quad (9)$$

Results of course and project dataset are presented in Table 5, Figure 1 and Table 6, Figure 2, respectively and the highest accuracy value for each classifier in each dataset is written in boldface.

According to Table 5, for all feature weight thresholds (FWT), accuracy of classifiers with two-point and ring are higher than single-point crossover.

Table 5. Effect of FWT on accuracy for course dataset

Course dataset			
FWT	**Ring**	**Two-point**	**Single-point**
25	90%	91%	90%
40	92%	92%	92%
58	94%	**95%**	**93%**
70	93%	**95%**	**93%**
85	**95%**	94%	92%
105	94%	91%	90%
150	88%	90%	88%

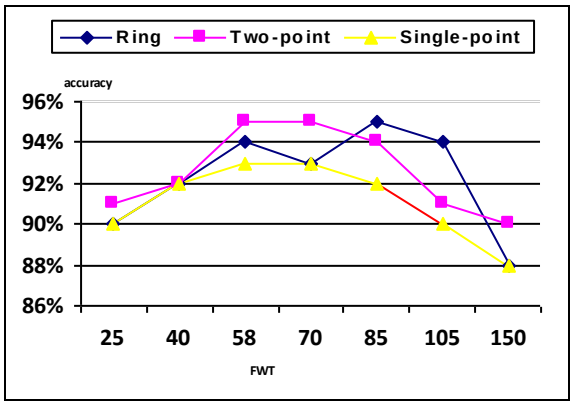

Fig 1: Accuracy values for course dataset

On the other hand due to the Table 6, accuracy of classifier with two-point crossover is highest.

Table 6. Effect of FWT on accuracy for project dataset

Project dataset			
FWT	**Ring**	**Two-point**	**Single-point**
13	74%	78%	74%
15	82%	85%	83%
18	89%	**93%**	89%
20	87%	90%	86%
23	**90%**	92%	**90%**
25	88%	91%	88%

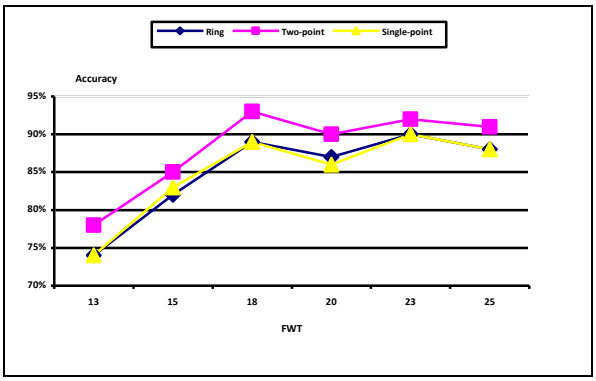

Fig 2: Accuracy values for project dataset

Based on above experimental results, it can be concluded that for both datasets, two-point crossover operator is leading to higher accuracy than the other two crossover operators.

The best FWT depends on used dataset and ratio of negative and positive pages in training dataset.

5.3.2 F-measure comparison

The F-measure as another commonly used value for measuring performance of classifiers is defined as [2,4]:

$$F\text{-measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (10)$$

Where precision and recall are computed as :

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \qquad (11)$$

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \qquad (12)$$

The F-measure value for course and project dataset are shown in Table 7, Figure 3 and Table 8, Figure 4 respectively and the highest value for each classifier in each dataset are written in boldface.

According to Table 7, maximum F-measure of classifier with single-point, two-point and ring crossover operators are respectively 86%, 89% and 88%.

In project dataset, according to Table 8, are 65%, 68% and 58%.

Base on presented results, it's specified that the classifier with two-point crossover has higher performance than the other two crossover operators.

Table 7. Effect of FWT on F-measure of course dataset

Course dataset			
FWT	**Ring**	**Two-point**	**Single-point**
25	79%	80%	79%
40	83%	82%	82%
58	87%	**89%**	**86%**
70	85%	88%	85%
85	**88%**	88%	85%
105	87%	81%	81%
150	73%	77%	75%

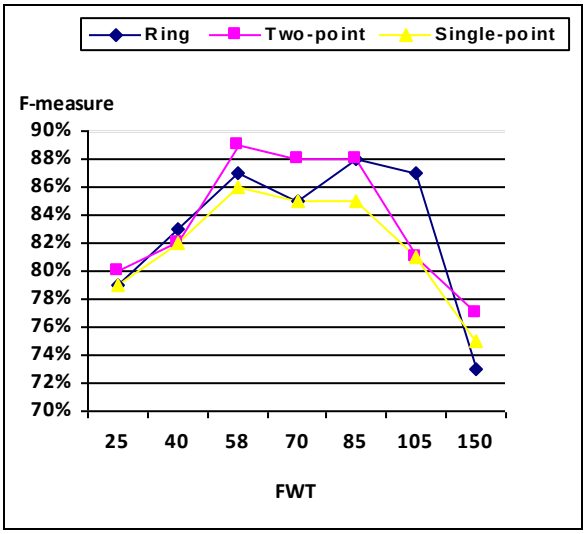

Fig 3: F-measure values for course dataset

Table 8. Effect of FWT on F-measure of project dataset

Project dataset			
FWT	**Ring**	**Two-point**	**Single-point**
13	46%	47%	43%
15	52%	55%	50%
18	**58%**	**68%**	**65%**
20	50%	53%	45%
23	54%	59%	54%
25	29%	32%	29%

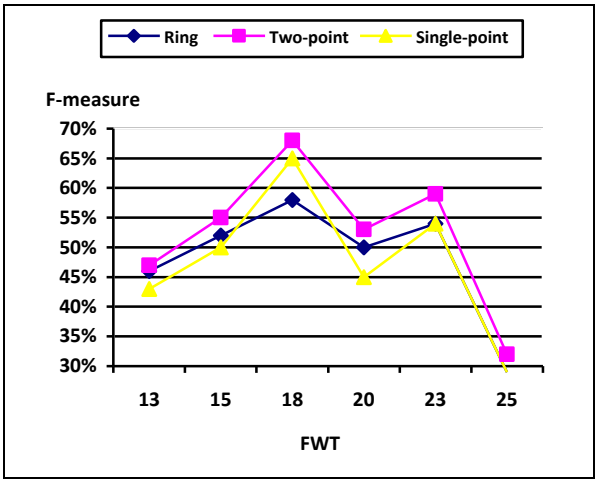

Fig 4: F-measure values for project dataset

6. Conclusion

In this paper we have study on the effect of used crossover operator on performance of GA as a web page classifier. So we have used single-point, two-point and ring crossover operators and evaluated the performance of them. According to the results, it was found that used classifier with two-point crossover, not only will lead to higher accuracy than the two other operators, but also has the highest rate of F-measure.

Meanwhile, the running time of used classifier is the same for 3 mentioned crossover operators.

7. References

[1] Holland, J. 1975. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence", Cambridge, MA: MIT Press (1992).

[2] Ribeiro, A., Fresno, V., Garcia-alegre, M., Guinea, D. and Carlos, J. 2003. Web page classification, A soft computing approach, Lecture Notes in Artificial Intelligence, 2663, pp.103-112.

[3] Pietramala, A., Policicchio, V., Rullo, P. and Sidhu, I. 2008. A Genetic Algorithm for Text Classification Rule Induction, Lecture Notes in Artificial Intelligence, 5212, pp. 188-203.

[4] Özel, S. 2010. A Web Page Classification System Based on a Genetic Algorithm Using Tagged-Terms as Features", Expert Systems with Applications.

[5] Reeves, C. and Rowe, J. 2003. Genetic Algorithms Principles and Perspectives, Kluwer Academic Publishers. Dordrecht.

[6] Kellegoz, T., Toklu, B. and Wilson, J. 2008. Comparing efficiencies of genetic crossover operators for one machine total weighted tardiness problem, Applied Mathematics and Computation , Vol 199, pp. 590–598.

[7] Kaya, M. 2011. The effects of two new crossover operators on genetic algorithm performance, Applied Soft Computing, 11, pp. 881–890.

[8] Kaya, Y., Uyar, M. and Tekin, R. 2011. A Novel Crossover Operator for Genetic Algorithms: Ring Crossover, Computing Research Repository Journal, Vol .abs/1105.0.

[9] Chinnasri, W. 2012. Performance comparison of Genetic Algorithm's crossover operators on University Course Timetabling Problem , Computing Technology and Information Management, vol.2, pp. 781-786.

[10] Craven, M., Dipasquo, D., Freitag, D., Mccallum, A., Mitchell, T., Nigam, K. and Slattery, S. 1998. Learning to Extract Symbolic Knowledge from the World Wide Web, The 15th national conference on artificial intelligence, pp. 509-516.

[11] Kim, S. and Zhang, B. 2003. Genetic mining of html structures for effective web document retrieval, Applied Intelligence, 18, pp. 243-256.

[12] Trotman, A. 2005. Choosing document structure weights, Information Processing and management, 41(2), pp. 243-264.

[13] Porter, M. 1980. An algorithm for suffix stripping", Program, 14(3), pp. 130-137.

[14] Salton, Wong and Yang. 1975. A vector space model for automatic indexing, Communications of the ACM ,18(11),pp. 613-620.

[15] Kamber, M. and Han, J. 2008. Data mining: Concepts and Techniques (2nd ed), San Francisco, Morgan Kaufman publisher, ISBN 1-55860-901-6.

[16] Yang, Y. and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization, the Fourteenth International Conference on Machine Learning (ICML-97), 412–420.