Genetic Relation Algorithm And PageRanking For Enhancing Web Search

V.Vignesh
Assistant professor, Department of Computer Science
Veerammal Engineering College, Dindugul,
Tamil nadu, India

Dr.K.Krishnamoorthy
Professor, Sudharsan Engineering college, Pudukkottai,
Tamilnadu, India

***Abstract*—Due to rapid growth of the number of Web pages, web users encounter two main problems, namely: many of the retrieved documents are not related to the user query which is called low precision, and many of relevant documents have not been retrieved yet which is called low recall. Information Retrieval (IR) is an essential and useful technique for Web search. Because of its parallel mechanism with high-dimensional space, Genetic Algorithm (GA) has been adopted to solve many of optimization problems where IR is one of them. This paper proposes searching model which is based on GA to retrieve HTML documents. This model is called Information Retrieval Using Genetic Relation Algorithm (IRUGRA). It is composed of two main units. The first unit is the document indexing unit to index the HTML documents. The second unit is the GA mechanism which applies selection, crossover, and mutation operators to produce the final result, while specially designed fitness function is applied to evaluate the documents. IRUGRA is a promising technique in Web search domain that provides a high quality search results in terms of recall and precision.**

Keywords- Information Retrieval (IR), Genetic Algorithm (GA), IRUGRA

I. INTRODUCTION

The World-Wide Web provides users with access to an abundance of information. Users query particular information from the Web using web search engines, and these web search engines apply the information retrieval (IR) techniques to produce the needed information. Information Retrieval is primarily devoted to extracting relevant information in response to user query. The increasing amount of information on the web raises new and challenging problems for information retrieval which is denoted as web search problem. Recently, IR problems have gained a considerable importance, and most studies argue that IR can be seen as a standard optimization problem (Marghny and Ali, 2005; Petridis, Kazarlis and Bakirtzis, 1998; Deb, 1998). Therefore, many researches are directed towards the use of Genetic Algorithms (GAs) for developing such a system which has proved its simplicity and capability as a powerful search mechanism to solve many scientific and engineering problems (Minaei-Bidgoli and Punch, 2003; Asllaniand and Lari, 2007; Losee, 1996; Deb, 1998). As is clear from its title, the goal of this paper is to utilize the concept of GA with a significant improvement to produce what is called: "Information Retrieval Using the Genetic Relation Algorithm (IRUGRA) model".

II. THE DESIGN OF INFORMATION RETRIEVAL USING GENETIC RELATION ALGORITHM (IRUGRA)

IRUGRA consists of two main units. The main purpose of the first unit, namely indexing, is to extract the meaningful keywords from the documents and represent them in a way that makes the process of finding relevant documents efficient. GA is the second unit of IRUGRA and is utilized in this paper as a core of its behavior. This unit compares the user query with the indexed documents to retrieve the relevant set of documents and display them in a descending order according to a relevance measure.

More precisely, in order to obtain high quality results, additional units need to be combined with IRUGRA, namely, the query formatting unit and the ranking unit. These units are outlined as shown in Figure 1.

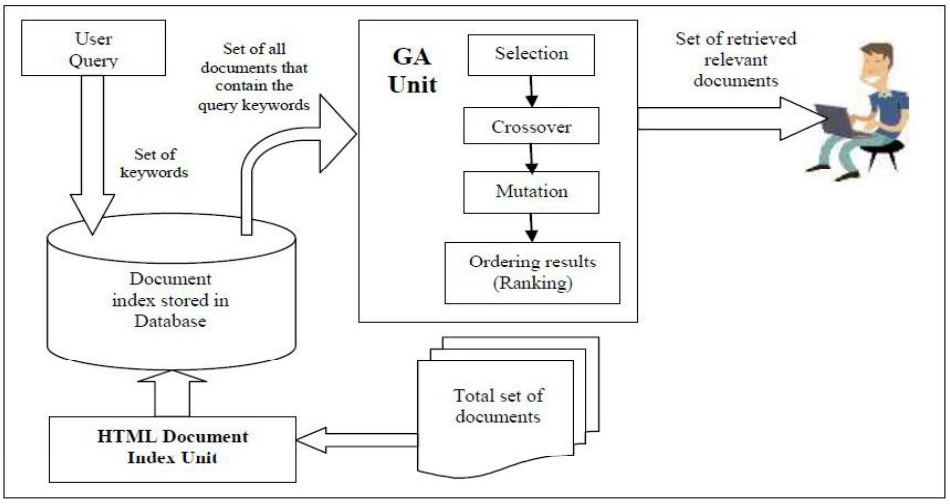

Figure 1: The units of IRUGRA

III. THE DESIGN OF HYBRID CROSSOVER

The proposed crossover operator chosen to be implemented in IRUGRA is a combination of reordering crossover (Vrajitoru, 2000), fusion crossover (Vrajitoru, 1998) and one-point crossover (Marghny and Ali, 2005). When genes within a chromosome are ordered based on their fitness value and the order is important, then the crossover applied to such chromosomes is called a reordering crossover. In fact, the order of genes in the proposed crossover to is important as it represents the ranked documents that will be displayed to the user. If one offspring is to be produced from the crossover process rather than two then it is called a fusion crossover. Combining these two techniques together and applying a one- point crossover on them forms the new crossover suggested in the GA unit of IRUGRA.

The cross point cp is selected randomly to perform a one-point crossover. In this example it is 3. Because the first gene of x has a greater fitness value than the first gene of y, x's genes along with the fitness values are considered as the first three genes of O. To complete the genes values of O, the other three genes are copied starting from the leftmost position of y. Then a competition between the genes in both x and y is done to complete the creation of O. Because the gene at position cp+1 in y has a greater value than that of x, then y's genes are copied into O (step C in Figure 2). Once all positions in the offspring are populated with genes, these genes are ordered from higher to lower based on their fitness value (step D in Figure 2). The algorithm of hybrid crossover is illustrated in Algorithm.

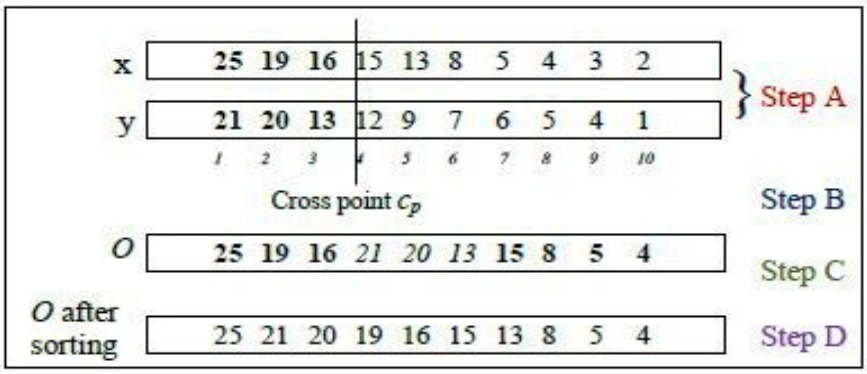

Figure 2: Illustration of the hybrid crossover process

Algorithm: The hybrid crossover operator

Prerequisite: Both parents are of same length and the genes are sorted with respect to their fitness value.
Select crosspoint *cp* randomly such that 0 < *cp* < parent length.
g_{max}= max gene(f(x_1), f(y_1)) --compare fitness value of first gene in both parents
parent 1= chromosome with g_{max}
Create offspring such that: *O*
= g1, g1 ≤ *cp*
= g2, g2 ≤ c_p, g2 *O* and length(*O*) ≤ length(parent1)
If length (*O*) < length(parent1)
begin
g`$_{max}$= max gene(f(x_{cp+1}), f(y_{cp+1}))
parent 1`= chromosome with g`$_{max}$
Copy genes from parent 1` to *O* such that genes are unique in *O*
end;
Order genes in *O* in descending order with respect to their fitness value.

IV. MUTATION

Mutation is the last genetic operator used in the GA unit of IRUGRA. In mutation, one or more genes are selected randomly to be replaced by other genes according to some criteria. It causes the individual genetic representation to be changed according to some probability pm ranging from 0.001 to 0.7. Because of its importance and effect on the generated chromosome, it is applied in this system with probability of 0.7.

An example of the mutation applied in this work is illustrated in Figure 3 where the numbers in this figure represent the fitness value of genes at these positions. The chromosome represented here is a continuation to the one shown in Figure 3. The position of mutation is selected randomly (position 7 in this example – Step B). The gene at this position is replaced by another gene selected randomly from the space such that it has a better fitness value or the same as the replaced one. In this example, the new value is 23 and it is better than the original one: 13 (Step C). This new value is unique within this chromosome; therefore it is exchanged with the original one. Then genes of this chromosome are re-ordered in descending order according to their fitness value to produce the new chromosome (Step D).

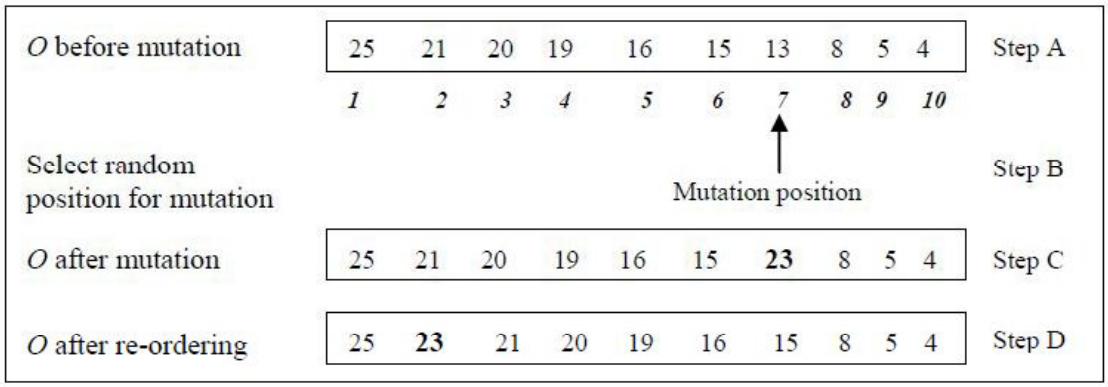

Figure 3: Illustration of the applied mutation in IRUGRA

V. EXPERIMENTS AND RESULTS OF IRUGRA

The measures used to evaluate the performance of IRUGRA. These measures are recall at rank N, precision at rank N and precision at recall M, where N is multiples of 10 and M is multiples of 10%. The storage space required to store the indexed documents using the enhanced inverted index (EII) is compared with the space required by the vector space model.

A. The Comparison between the Hybrid Crossover and Two-Point Crossover

The first experiment in the crossover comparisons is to study the first measure denoted as precision @ top N. In this experiment, the comparison will be done between the hybrid crossover technique and the two-point crossover, abbreviated as "2-point CO"

Figure 4 and Table 1 show the precision @ top N retrieved documents. It is shown that the GA unit of IRUGRA using hybrid crossover has much better performances than the 2- point crossover (referred to as "2-point CO" in this figure) for the reason mentioned above. Moreover, the hybrid crossover achieves 0.86 at the top 10 retrieved documents, while the 2-point crossover achieves only 0.34. In other words, hybrid crossover achieves an improvement of 152.84% at top 10 over the 2-point crossover.

Figure 4: Comparison of P@N between hybrid crossover and 2-point crossover techniques

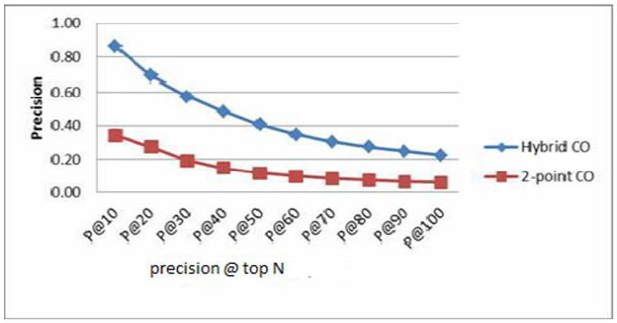

Table 1: The P@N enhancement percentage of hybrid crossover over the 2-point crossover techniques

Measure	Hybrid CO	2-point CO	% of improvement
P@10	0.86	0.34	152.84
P@20	0.69	0.27	152.52
P@30	0.57	0.19	199.22
P@40	0.48	0.15	233.22
P@50	0.41	0.12	250.46
P@60	0.35	0.1	259.51
P@70	0.30	0.08	267.68
P@80	0.27	0.07	276.05
P@90	0.25	0.06	283.58
P@100	0.22	0.06	285.64
Average	**0.44**	**0.14**	**236.07**

The second measure to be considered in evaluating this technique is the recall @ top N. Figure 5 shows that the recall @ top N retrieved for the hybrid crossover starts from 63% until it reaches 85% at R@60. That means this technique is capable of retrieving 85% of the total relevant documents at top 60 retrieved documents. However, the 2-point crossover technique starts by retrieving 31% of relevant documents at top 10, and as a whole it retrieves only 35% at top 100 retrieved documents. That implies hybrid crossover achieves enhancement of 104% at R@10 and drops to 82.32% at R@100. These results are shown in Table 2.

Figure 5: Comparison of R@N between hybrid crossover and 2-point crossover technique.

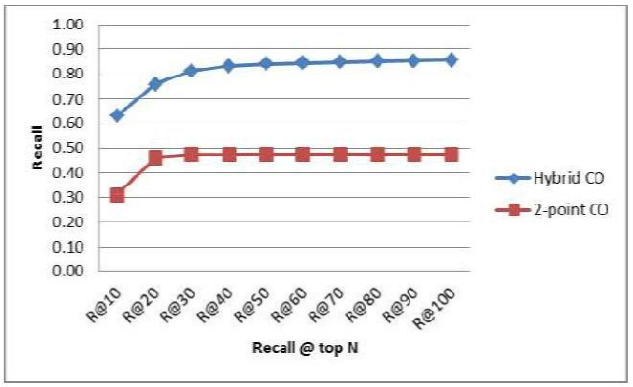

Table 2: The R@N enhancement percentage of hybrid crossover over the 2-point crossover techniques

Measure	Hybrid CO	2-point CO	% of improvement
R@10	0.63	0.31	104.50
R@20	0.76	0.46	63.55
R@30	0.81	0.47	71.09
R@40	0.83	0.48	75.25
R@50	0.84	0.48	77.09
R@60	0.85	0.48	77.72
R@70	0.85	0.48	78.37
R@80	0.85	0.48	79.02
R@90	0.85	0.48	79.67
R@100	0.86	0.48	80.32
Average	**0.81**	**0.46**	**78.66**

The third measure is the precision @ recall which evaluates the precision percentage when retrieving multiples of 10% of relevant documents. In other words, this measure evaluates the purity of the results from the irrelevant documents.

Figure 6 shows the performance of the proposed hybrid crossover over the 2-point crossover. By using the hybrid crossover, the GA unit of IRUGRA was able to achieve 99% of relevance when retrieving 30% of the total relevant documents. This percentage reduces to 87% when retrieving all the relevant documents. However, the two-point crossover has 50% of relevant documents when retrieving 30% of relevant documents, and this percentage dropped to 31% when retrieving all relevant documents. These scores show that the hybrid crossover managed to achieve an enhancement of 130.07% in the average over all 10-points shown in Table 3 for the precision @ recall measure.

Figure 6: Comparison of P@R between hybrid crossover and 2-point crossover technique

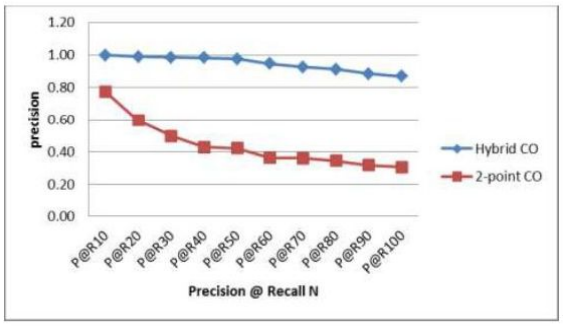

Table 3: The P@R enhancement percentage of hybrid crossover over the 2-point crossover techniques

Measure	Hybrid CO	2-point CO	% of improvement
P@R10	0.99	0.78	28.96
P@R20	0.99	0.60	66.07
P@R30	0.99	0.50	98.65
P@R40	0.98	0.43	127.68
P@R50	0.98	0.42	131.82
P@R60	0.95	0.36	162.26
P@R70	0.93	0.36	157.96
P@R80	0.91	0.34	164.13
P@R90	0.89	0.32	179.39
P@R100	0.87	0.31	183.77
Average	**0.95**	**0.44**	**130.07**

B. Comparing the Hybrid Crossover and Non-ordered Crossover

Another alternative technique for crossover is the one-point crossover applied to non- ordered chromosomes (abbreviated as Non-Ordered CO) to produce one offspring. What differentiates this technique from the hybrid crossover technique is that the genes within the chromosome are not ordered according to their fitness value. Thus, good genes (genes that have high fitness value) are scattered throughout the chromosome resulting in a chromosome having a mixture of good and bad genes distributed arbitrarily within the chromosome. Applying a one-point crossover on such a chromosome results in swapping these mixed genes from one side of the cross point to the other side without any noticeable improvement.

Although non-ordered crossover techniques is much better than the two-point crossover, it is still not able to beat the proposed hybrid crossover. Referring to Figure 7 of P@N measure, it is shown that this technique starts at a precision of 0.86 at the top 10 retrieved documents and ends with a precision of 0.22 at the top 100 retrieved documents, as compared with 0.58 and 0.13 for the same points for the hybrid crossover technique. That means the second technique is enhanced from 48.12% to 70.62%. These scores are illustrated in Table 4.

Figure 7: Comparison of P@N between hybrid crossover and the non-ordered crossover techniques.

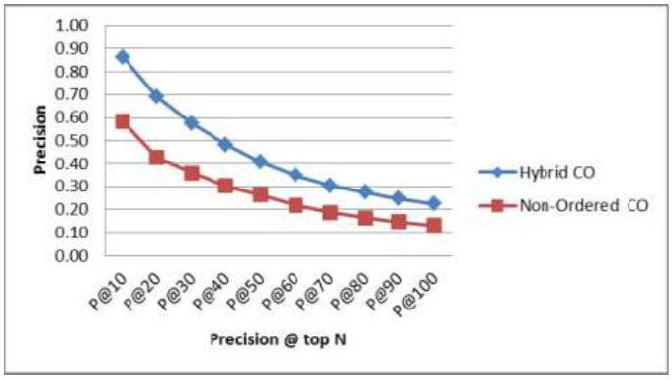

Table 4: The P@N enhancement percentage of hybrid crossover over the non-ordered crossover techniques

Measure	Hybrid CO	Non-Ordered CO	% of improvement
P@10	0.86	0.58	48.12
P@20	0.69	0.43	61.95
P@30	0.57	0.36	59.99
P@40	0.48	0.3	59.81
P@50	0.41	0.26	55.05
P@60	0.35	0.22	59.14
P@70	0.3	0.19	62.56
P@80	0.27	0.16	66.79
P@90	0.25	0.15	69.58
P@100	0.22	0.13	70.62
Average	**0.44**	**0.28**	**61.36**

When comparing this technique with the hybrid crossover technique in terms of R@N measure as illustrated in Figure 8, it is noticed that the non-ordered crossover performance ranges between 39% at R@ top 10 and 51% at R@ top 100. This means that this technique lags behind hybrid crossover technique by 60.33% to 68.11%. Table 5 lists the scores for each point of the scale of R@N measure.

Figure 8: Comparison of R@N between hybrid crossover and the non-ordered crossover techniques

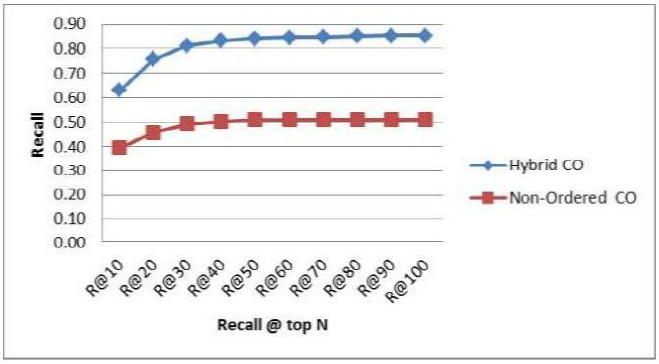

Table 5: The R@N enhancement percentage of hybrid crossover over the non-ordered crossover techniques.

Measure	Hybrid CO	Non-Ordered CO	% of improvement
R@10	0.63	0.39	60.33
R@20	0.76	0.46	65.50
R@30	0.81	0.49	65.10
R@40	0.83	0.50	65.47
R@50	0.84	0.51	65.09
R@60	0.85	0.51	65.68
R@70	0.85	0.51	66.29
R@80	0.85	0.51	66.90
R@90	0.85	0.51	67.50
R@100	0.86	0.51	68.11
Average	**0.81**	**0.49**	**65.60**

The last measure to be compared between the non-ordered crossover technique and the hybrid crossover is the precision @ recall measure. The results are shown in Figure 9. The performance of the former technique ranges between 0.92 at P@R10 and 0.43 at P@R100, compared with hybrid crossover which ranges from 1 at P@R10 to 0.87% at P@R100. As demonstrated in Table 6, it is found that the proposed technique enhanced the performance by 102.89% at P@R100.

Figure 9: Comparison of P@R between hybrid crossover and the non-ordered crossover techniques

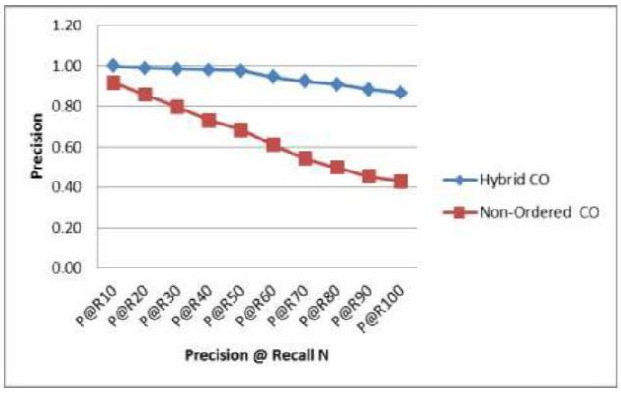

Table 6: The P@R enhancement percentage of hybrid crossover over the non-ordered crossover techniques.

Measure	Hybrid CO	Non-Ordered CO	% of improvement
P@R10	1	0.92	8.69
P@R20	0.99	0.86	15.29
P@R30	0.99	0.8	24.5
P@R40	0.98	0.73	34.01
P@R50	0.98	0.68	43.32
P@R60	0.95	0.61	55.35
P@R70	0.93	0.54	70.73
P@R80	0.91	0.5	82.58
P@R90	0.89	0.45	96.97
P@R100	0.87	0.43	102.89
Average	**0.95**	**0.65**	**53.43**

VI. CONCLUSION:

A new crossover technique is presented called hybrid crossover. The proposed techniques are compared with the existing ones in terms of the three measures, precision at top N (P@N), recall at top N (R@N) and precision at recall (P@R). Each operator is examined using these three measures and the results are presented graphically and numerically. The hybrid crossover is compared with 2-point cross over and non-ordered crossover. In all cases, hybrid crossover of IRUGRA performance is good than other existing techniques.

REFERENCES

[1] Zhang, X., Wei, K., and Meng, X., (2014), A XML query results ranking approachbased on probabilistic information retrieval model, *9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2012,* pp. 915 – 919. IEEE Conference Publications

[2] Salton, G., and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science,* vol. 41, no. 4, pp. 288-29

[3] Saini, M. Sharma, D. Gupta, P.K . (2014), Enhancing information retrieval efficiency using semantic-based-combined-similarity-measure. *International Conference on Image Information Processing (ICIIP),* pp. 1 - 4. IEEE Conference Publications

[4] Pohl, S., Zobel, J. and Moffat, A (2010), Extended Boolean retrieval for systematic biomedical reviews, *Proceedings of the Thirty-Third Australasian Conference on Computer Science,* vol. 102, pp. 117-126

[5] Green, J.J. (2004). Google PageRank and related technologies, [online]. Available at: http://www.lazworld.com/whitepapers/PageRank-Technologies.pdf [Accessed 22/9/2012]

[6] Guezouli, L. and Kadache, A. (2012), Information retrieval model based on neural networks using neighbourhood, *International Conference on Information Technology and e-Services (ICITeS),* pp. 1 – 5. IEEE Conference Publications

[7] Hemalatha, M. and Sathya Srinivas, D. (2009). Hybrid neural network model for web document clustering, *Second International Conference on the Applications of Digital Information and Web Technologies, ICADIWT '09,* pp.531 - 538. IEEE Conference Publications.

[8] Callen, B. (2005). Search engine optimization made easy [Online]. Available at: http://www.seoelite.com [Accessed 16/4/2007]

[9] Vrajitoru, D. (1998). Crossover improvement for the genetic algorithm in information retrieval. *Information Processing and Management* , vol. 34, no. 4, pp. 405-415

[10] Sivanandam , S. N., and Deepa, S. N. (2008). *Introduction to Genetic Algorithms.* New York: Springer Berlin Heidelberg..

[11] Picarougne, F., Monmarch´e, N., Oliver, A., and Venturini, G. (2002). Geniminer: Webmining with a genetic based algorithm. *Proceedings of the IADIS International Conference WWW/Internet,* pp. 263–270. Lisbon, Portugal