6.1 INTRODUCTION

In addition, the existence of non-linear data sources makes the investigation of web user usage behavior complex. The conventional methods are limited to optimization problems due to the absence of semantic certainty and presence of human intervention. In handling such data and overcome the limitations of conventional methodologies it is necessary to use a soft computing model, which is a consortium of methodologies, and can work synergistically with global processing capabilities.

The soft computing models are characterized by their ability for granular computation in avoiding the concept of approximation. Basically, soft computing models provide the foundation for computational intelligence systems and further outline the basis of future generation computing systems. These models are close resemblance to human like decision making and used for modeling highly non-linear data, where the pattern

Figure 6.1 Intelligent Optimal Genetic Model of CWUUBS
discovery, rule generation and learnability are typical. The Fuzzy Logic (FL), Artificial Neural Networks (ANN), Genetic Algorithms (GA) and various combinations of these techniques have made the soft computing paradigm. Among which, genetic algorithms, a biologically inspired technology and is more suitable for web data, which is intrinsically unlabelled, heterogeneous and dynamic.

Genetic algorithms are examples of evolutionary computing methods and are optimization type algorithms for both supervised and unsupervised techniques. A genetic algorithm is an elegantly simple, yet extremely powerful and adoptive model in resulting exact optimal solutions. The GA is more adequate since the implicit parallelism of GA can mine the large web data in less time and the stochastic ability of GA yields in optimum solution. To proceed towards web intelligence, reducing the need of human intervention, it is necessary to incorporate and embed artificial intelligence into web mining tools. To achieve the intelligence GAs seemed to be a good candidate. Moreover, the GAs minimizes the assumptions in making the real study of web user usage behavior. Thus, in order to find the integrated and optimized solution for investigating the web user usage behavior, it is essential to make use of an optimal web intelligent model with granular computing nature of genetic algorithms.

To achieve the optimized solution for investigating the web user usage behavior, the present chapter proposes an Intelligent Optimal Genetic Model (IOGM) as shown in figure 6.1, which is designed as an optimization tool, based on the concept of natural genetic systems. The goal of IOGM is to find the optimized solutions for investigating the web user usage behavior. The IOGM fitness function is designed in such a way that, it generates dissimilar page vectors, each vector having similar characteristics. In addition, the IOGM evaluation function is planned to identify the quality of optimized population averaged mean for page vectors by considering both link and page quality of each vector. The IOGM frame work collectively uses fitness, operators and evaluation functions, helps to investigate the web user usage behavior significantly.

The IOGM requires web user usage patterns as input that is generated by IFP-tree which is discussed in chapter 5. The IFP-Tree finds the association among the visited
pages in the identified category by applying incremental mining techniques. It utilizes the previous mining results and finds new patterns from the inserted or deleted part of the weblog. It also allows interactive mining with different supports.

The optimal web usage patterns evaluated by IOGM are the input for AKRS which will be presented in the chapter 7. The AKRS reviews the patterns and interprets the domain specific knowledge. In addition, it helps to represent the final knowledge comprehensively.

The remaining part of the chapter is organized as follows: Section 6.2 describes the motivations towards soft computing models. Section 6.3 presents the proposed work in detail. Section 6.4 shows the experimental analysis, finally section 6.5 gives the summary.

6.2 MOTIVATIONS TOWARDS SOFT COMPUTING MODELS

The investigation of web user usage behavior is real time problem which involves multiple conflicting measures of performance and desires optimization. These measures are not only computational intensive but it is also possible to loose the exact solution. To attain the optimal solution it is necessary to involve techniques having learning capability. This motivated towards the computing models which are robust, fast and close to reality. Additionally, the soft computing models strengthen by their ability to estimate the optimal stopping time of process. In particular, genetic algorithms emulate the concept of theory of evaluation in converging global optimal solution.

The following reasons are the other exacting motivation factors resulting in choosing the genetic algorithms in modeling IOGM proposed in this chapter.

- The use of heuristic techniques is well established in optimization problems where the objective function and the constraints do not have a simple mathematical formulation.
- To determine a good solution in small computing time, where the dimension of the problem significantly large.
The structure of problem is straightforward, representable by the data structure commonly use by GA. Moreover, the sophisticated genetic representation and latest developments in modeling the genetic operators craft the solution of investigation of web user usage behavior efficiently and effectively.

6.3 PROPOSED INTELEGENT OPTIMAL GENETIC MODEL (IOGM)

In investigating the web user usage behavior, the web miners have to take many intelligent decisions at each stage of web user usage behavior system CWUUBS. All such decisions have to be made sequentially at different levels within time. Since the weblog data is impacted by many external and explicit functions, therefore the data in the weblog becomes non linear and the problem space turns voluminous. Consequently, the programming techniques used for this type of the problem happen to be computationally expensive. To attain global solution and minimize the efforts required, the present chapter proposes an Intelligent Optimal Genetic Model (IOGM) which is a standard non linear optimization technique.

Optimization is the method of obtaining the global solution under any given circumstances. There are many optimization techniques in the literature

- Classical
- Linear programming
- Non Linear Programming without constraints
- Non linear programming with constraints
- Dynamic Programming
- Stochastic Programming
- Soft computing Techniques

Out of these a soft computing technique, genetic algorithm is used for IOGM, which is well suited to solve the problem of investigating the of web user usage behavior since the weblog data is characterized by both continuous and discontinuous functions. To get the optimum solution for investigating the web user usage behavior
it is necessary to process the potential patterns extracted by pattern discovery technique of web usage mining process.

The proposed work considers each step of genetic algorithm in the light of web mining. Towards this goal, the authors present the IOGM, which equally concentrates on all steps of genetic process and is more adaptable to incremental web log scenario. The architecture and the process of IOGM are as shown in figure 6.2.

![Figure 6.2 Architecture of Intelligent Optimal Genetic Model](image)

Figure 6.2 Architecture of Intelligent Optimal Genetic Model

The first and perhaps the most typical task is to identify the initial population by determining as a set of individual solutions from the list of visited patterns to initiate the process using IOGM encoding strategy. Subsequently, by the theory of evolution, only optimal individuals survive from the population and then generate the next biological population based on IOGM fitness function. In the next step, biologically inspired IOGM operators create a new and potentially better population. Later, IOGM sorts the visited patterns in desired order based on past performance combined from human thresholds. Finally, this process is terminated as and when an acceptable optimal set of visited patterns is found or after the lapse of a fixed time limit of IOGM end function.
6.3.1 IOGM Encoding Strategy

The encoding strategy is a process of representing the potential solution to a problem into a suitable form of viable individuals, the genetic algorithm can process. It is a crucial issue in genetic process as it plays a critical role to arrive at best performance of algorithm as robust as possible. Various encoding strategies have been introduced in the literature for effective implementation of genetic algorithms. IOGM adopts the Binary encoding strategy to determine initial population from the visited patterns generated by pattern discovery stage of web usage mining process.

The sample output generated by pattern discovery of web usage mining process is as shown in figure 6.3. This can be considered as input for IOGM.

<table>
<thead>
<tr>
<th>Ending with P5: {P5}, {P4, P5}, {P1, P4, P5}, {P3, P5}, {P1, P5}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ending with P4: {P4}, {P3, P4}, {P2, P3, P4}, {P1, P3, P4}, {P2, P4}, {P1, P4}</td>
</tr>
<tr>
<td>Ending with P3: {P3}, {P2, P3}, {P1, P2, P3}, {P1, P3}</td>
</tr>
<tr>
<td>Ending with P2: {P2}, {P1, P2}</td>
</tr>
<tr>
<td>Ending with P1: {P1}</td>
</tr>
</tbody>
</table>

Figure 6.3 Sample sets of visited web pages

The encoding strategy adopted by IOGM creates various sets of chromosomes with genes as possible solution. The chromosome set is strings of 0’s & 1’s and coded as finite-length string over an alphabet of finite length. A commonly used principle for coding is known as principle of minimum alphabet [16]. Each chromosome refers to a coded possible solution. A set of such chromosomes in a generation is called an initial population, the length of which may be constant or may vary from one generation to another. An example of gene {P2, P3, P4} is encoded as a binary chromosome of length 8 and is shown in Figure 6.4. The presence of a web page is coded as 1, otherwise as 0. Evidently, $2^m$ different chromosomes are generated, where $m$ is length of string.
6.3.2 IOGM Evaluation Function

The evolution function quantifies the optimality of a solution so that a particular solution is ranked against all the other solutions. The function depicts the nearness of a given solution to the required outcome. It is an essential step in the overall process of genetic approach as it plays a key role to evaluate survival capacity. The IOGM employs a robust fitness function which is designed based on scores, ideal dimension and Euclidian distance of chromosomes.

In the Fitness function, Dc indicates number of genes in form of visiting patterns included in each chromosome and Nc indicates the number of chromosomes.

Thus the population \( N_p = D_c \cdot N_c \) genes \( (1) \)

The fitness function computed for each chromosome is expressed as a positive value and is to be maximized. It is composed of three terms namely Term1, Term2 and Term3. The first term is the sum of the length of the genes in chromosome \( C \),

\[
\text{Term 1 : } T_1(C) = \sum_{p_i \in C} \text{length}(G_i) \quad (2)
\]

Where length\( (G_i) \) is the original length given to gene \( G_i \). This term considers genes with only high positive value in a chromosome. This drawback is balanced by considering the minimum support in second term of the Fitness function. Let \( S \) be a minimum support.

\[
\text{Term 2 : } T_2(C) = \frac{N_p}{\text{abs}(|C| - S)} + 1 \quad (3)
\]

It reaches its maximum \( N_p \) when the dimension of \( C \) is exactly equal to the minimum support \( S \) and rapidly decreases when the number of genes contained in chromosome \( C \) is less than \( S \).
The chromosomes that are present in the initial population are associated by the highest possible variability values as far as the genes are concerned. The evaluation of the population alters assigned values in process of creating new population. Moreover, the fact that genes belonging to different chromosomes may not be guaranteed, as it depends on the number of genes \( D_c \). For this reason, fitness function third term is introduced, which measures directly the overall dissimilarity of the genes in the chromosomes. Let \( D(G_i, G_j) \) be the distance between gene \( G_i \) and \( G_j \), is the sum of the distance between the pairs of genes in chromosome \( C \) and the measures the total variability expressed by \( C \).

\[
\text{Term 3: } T_3(C) = \sum_{G_i \neq G_j} D(G_i, G_j) 
\]

The final form of the fitness function (FF) for chromosome \( C \) is derived from the equations 2, 3 and 4.

\[
\text{FF}(C) = \alpha \cdot T_1(C) + \beta \cdot T_2(C) + \gamma \cdot T_3(C) 
\]

Where \( \alpha \), \( \beta \) and \( \gamma \) are parameters that depend on the magnitude of the chromosome that represent the genes. In particular \( \alpha \), \( \beta \) and \( \gamma \) are chosen so the contributions given by \( T_1(C) \), \( T_2(C) \) and \( T_3(C) \) are balanced. This approach is used by the evolution function of IOGM for every chromosome \( C^* \), by means of the genetic operators.

\[
\text{FF}(C^*) = \max_{c=1, \ldots, N_c} \text{FF}(C) 
\]

### 6.3.3 IOGM Operators

The biologically inspired genetic operators are applied on population of chromosomes to generate potentially new offspring. This is an iterative and fundamental step in genetic approach so as to produce subsequent acceptable generation. The Selection, Crossover and Mutation are set of operators designated by IOGM which transforms individual chromosomes stochastically. Each chromosome has an associated value called fitness function that contributes in the generation of new population by genetic operators. At each generation, the IOGM utilizes the fitness function values to evaluate survival capacity of each chromosome. The IOGM operators create a new set of population that tries to improve on the current fitness function values by using old ones.
Selection: The selection process determines the number of times a particular individual chromosome is chosen for reproduction from initial population as a mating pool for IOGM further operations. The number of individual chromosomes obtain for the next generation is directly proportional to its fitness value, there by mimic the natural selection procedure. The IOGM deploys the stochastic universal selection to yield best offspring.

The key idea of stochastic universal selection gives preference to better individual chromosomes by permitting them to pass on their genes to the next generation and disallow the entry of worst fit individual chromosome into the next generations. It principally works at the level of chromosomes with no bias. The goodness of each individual chromosome depends on its associated fitness function value.

The methodology of stochastic universal selection of IOGM functions is to place the population $N_p$ with equidistant markers on the wheel, which has as many slots as the population size $N_p$. The each $N_p$ individual chromosomes picks at random by spinning the wheel and a single random number marker 1 in the range $[0, 1/N_p]$ is generated. The $N_p$ individual chromosomes are then selected by generating the $N_p$ markers, starting with marker 1 and spaced by $1/N_p$, and decide the individual chromosome whose fitness function value spans the location of the markers. The function $ET(i)$ is the Estimated Trials of individual chromosome $i$, $LBET(i)$, $UBET(i)$ are the Lower and Upper Bound functions of $ET(i)$ respectively. The individual chromosome achieves the assured least number of spins by choosing minimum times of $LBET(i)$ and within the $UBET(i)$.

![Figure 6.5 Example of stochastic universal selection](image-url)
This phenomenon continues until the number of individual chromosomes identical to the number of markers that lie within the matching slot as shown in figure 6.5.

**Crossover:** The crossover plays a vital role in the design and implementation of any robust genetic models. The key focus of crossover creates new chromosome which is better than its parents, by taking the best characteristics from each of the parents. Towards this, it exchanges the information between randomly mated pair of fixed-length chromosomes by recombining parts of their mating pool to produce new survival offspring. This operation, carry out probabilistically, combines parts of two parent chromosomes to generate new offspring. The IOGM chooses single point crossover technique as it generates successful new offspring.

The single point crossover technique of IOGM, initially pairs all the chromosomes at random in the population obtained by stochastic universal selection procedure. The example pair of chromosomes depicted with two different colors is as shown in figure 6.6(a).

![Figure 6.6(a) Sample pair of chromosomes](image)

Later, it selects the crossover point K randomly between 1 and L - 1, where L is the length of the chromosome. This crossover point occurs between two bits and divides each individual chromosome into two parts. The crossover point is represented with a thin vertical line across pair of chromosomes as shown in figure 6.6(b).

![Figure 6.6(b) Single point crossover operation before crossover](image)
Finally, it performs crossover operation on a pair of chromosomes at the crossover point. Then, the parts of two parents after the crossover point are exchanged to form new offspring as shown in figure 6.6(c). The single point crossover technique of IOGM repeats with a good number of trails to arrive at a feasible offspring which push forward to the mutation process in obtaining the optimal solution.

**Mutation:** Mutation is a process which alters the genetic structure of the chromosome randomly. Its central aim is to submit an application of genetic diversity from one generation of population of chromosomes to the next. The optimal solution resides in a chromosome which is not presented in the selection population. Thus, the previous genetic operators are unable to reach the global optimum. In such situations, only mutation helps in generating new population with which optimal solution can be attained. IOGM make use of bit by bit mutation to arrive at optimal population.

The bit by bit mutation of IOGM alters every gene value in a chromosome from its initial state with a low probability. It inverts the values of chosen gene. The value of 0 is replaced with 1 and vice versa. This results as new chromosome with entirely new gene values. An example of bit by bit mutation is shown in figure 6.7.
Mutation is an important part of IOGM, helps to prevent the population from stagnating at any local solution of the solution space. With the new mutated gene values, the IOGM is capable to arrive at better optimal solution than previously possible.

6.3.4 IOGM End Function:
The IOGM results in the best chromosomes as the process is repeated for infinite number of iterations. The process needs to be stopped after finite number of iterations. To take the decision on number of iterations IOGM design an end function that stops the process at a finite value n, where n is the stopping time of the process. The objective of end function determines the value of n, with which the process of IOGM achieves optimal solution. Moreover, in stochastic process of IOGM, the value must be probabilistic since no finite stopping time can assure the optimal solution always.

The end function works by considering population size Np, Fitness Function value FF, length of gene Length(Gj), crossover probability Pc and mutation probability p_m. It interrupts IOGM process when no significant or sufficient improvement is found in two or more consecutive generations. In some cases, IOGM uses time threshold set by user, as a criteria for terminating the process.

6.3.5 IOGM Algorithm:
The IOGM is implemented with the theories and techniques of genetic approach for web usage mining which derives optimal visiting patterns. In this process, encoding strategy along with evolution function tuned the length of the chromosome and the population size of patterns is discovered by pattern discovery technique. The biologically inspired operators adjust the probabilities of selection and crossover so that IOGM exhibits self learning capability to produce survival patterns. The genetic diversity of mutation refines the patterns replacement strategy. The termination criterion fixes the number of iterations that reduces the computational cost of IOGM in deriving optimal patterns. These patterns help in investigating the web user usage behavior intelligently.

The IOGM algorithm is constructively designed based on concept of genetic algorithm as follows,
Algorithm of IOGM:

<table>
<thead>
<tr>
<th>Algorithm : Intelligent Optimal Genetic Model - IOGM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 01: Define the Evaluation Function F.</td>
</tr>
<tr>
<td>Step 02: $t \leftarrow 0$ (iteration No =0, pop size =0)</td>
</tr>
<tr>
<td>Step 03: Initialize $P(t)$.</td>
</tr>
<tr>
<td>Step 04: Evaluate $P(t)$ (page from modeled data).</td>
</tr>
<tr>
<td>Step 05: Generate an offspring page $O$.</td>
</tr>
<tr>
<td>Step 06: $t \leftarrow t+1$ (new population).</td>
</tr>
<tr>
<td>Step 07: Select $P(t)$ from $P(t-1)$.</td>
</tr>
<tr>
<td>Step 08: Crossover $P(t)$.</td>
</tr>
<tr>
<td>Step 09: Mutation $P(t)$</td>
</tr>
<tr>
<td>Step 09: Evaluate $P(t)$.</td>
</tr>
<tr>
<td>Step 10: go to 5 (while no termination (no of iterations)).</td>
</tr>
<tr>
<td>Step 11: End Function $P(t)$ Gives the output to the user.</td>
</tr>
</tbody>
</table>

6.4. EXPERIMENTAL ANALYSIS

The proposed IOGM is experimented over a period of twelve months server side weblog data under standard execution environment. For the IOGM algorithm number of visited patterns is given as input to start the process. These patterns are generated by a pattern discovery technique.

A) The IOGM is compared with the standard web mining technique in terms of performance. The experimental results indicate a noticeable improvement of IOGM performance over the standard web mining technique as shown in figure 6.8.
B) The IOGM experimented for different cross over probabilities (Pc=0, 0.25, 0.5, 0.75, 1) on different number of iterations. It shows high performance at the average mean Pc as shown in Figure 6.9.
6.5. SUMMARY

In summary, the present chapter discusses the optimization problem where in various optimization techniques are studied. The relevance of soft computing model, especially genetic algorithm, in the nonlinear weblog scenario along with inspired motivation is discussed. The IOGM architecture with complete details is presented. The concept and methodology of each step of IOGM is explained and illustrated with an example. The working of IOGM algorithm is demonstrated.

Several experiments are conducted to evaluate the optimization of IOGM under standard execution environment. The encoding strategy, evolution function, fitness function and end function of IOGM algorithm are discussed in this chapter, and are put together to validate the sustainability of associated patterns. Moreover, the proposed IOGM ensures the identifiable features like learning, adaptability, self-maintenance and self-improvement. Hence the proposed IOGM helps to investigate user usage behavior on WWW in any application domain. The next organic part of CWUUBS is applied on these optimal patterns.