AN INTELLIGENT OPTIMAL GENETIC MODEL TO INVESTIGATE THE USER USAGE BEHAVIOUR ON WORLD WIDE WEB

V.V.R. Maheswara Rao[1] and Dr. V. Valli Kumari[2]

[1]Professor, Department of Computer Applications,
Shri Vishnu Engineering College for Women, Bhimavaram, Andhra Pradesh, India,
mahesh_vvr@yahoo.com

[2]Professor, Department of Computer Science & Systems Engineering,
College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India,
vallikumari@gmail.com

ABSTRACT

The unexpected wide spread use of WWW and dynamically increasing nature of the web creates new challenges in the web mining since the data in the web inherently unlabelled, incomplete, non linear, and heterogeneous. The investigation of user usage behaviour on WWW is real time problem which involves multiple conflicting measures of performance. These measures make not only computational intensive but also needs to the possibility of be unable to find the exact solution. Unfortunately, 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 that can work intelligently to attain optimal solution.

To achieve the optimized solution for investigating the web user usage behaviour, the authors in the present paper proposes an Intelligent Optimal Genetic Model, IOGM, which is designed as an optimization tool based on the concept of natural genetic systems. Initially, IOGM comprise a set of individual solutions or chromosomes called the initial population. Later, biologically inspired operators create a new and potentially better population. Finally, by the theory of evolution, survive only optimal individuals from the population and then generate the next biological population. This process is terminated as when an acceptable optimal set of visited patterns is found or after fixed time limit. Additionally, IOGM strengthen by its ability to estimate the optimal stopping time of process. The proposed soft computing model ensures the identifiable features like learning, adaptability, self-maintenance and self-improvement. To validate the proposed system, several experiments were conducted and results proven this are claimed in this paper.

KEYWORDS

Web usage mining, Genetic Algorithm, Optimal Solution, Selection, Crossover, Mutation

1. INTRODUCTION

The rapid advances in data generation, availability of automated tools in data collection and continued decline in data storage cost enable with high volumes of data. In addition, the data is non scalable, highly dimensional, heterogeneous and complex in its nature. This situation creates inevitably increasing challenges in extracting desired information. Thus, web mining evolves into a fertile area and got the focus by many researchers and business analysts. Web mining is a methodology that blends conventional techniques with incremental algorithms. Web mining is the

application of data mining techniques to discover and retrieve useful information and patterns from the WWW documents and services.

The web mining can be used to discover hidden patterns and relationships within the web data. The web mining task can be divided into three general categories, known as Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM) as shown in figure 1.

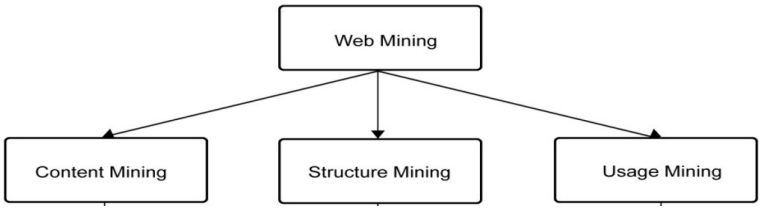

Figure 1. Types of web mining

Web content mining is the mining technique that can extract the knowledge from the content published on internet, usually as semi-structured (HTML), unstructured (Plain text) and structured (XML) documents. The content of a web page may be varied, like text, images, HTML, tables or forms. The ability to conduct web content mining allows results of search engines to maximize the flow of customer clicks to a web site, or particular web pages of the site, to be accessed numerous times in relevance to search queries. The main uses of this web mining are to gather, categorize, organize and provide the best possible information.

Web structure mining extracts the knowledge from the WWW and links between references in the web. Mining the structure of the web involves extracting knowledge from the interconnections of the hypertext documents in the WWW. This results in discovery of web communities, and also pages that are authoritative. The main purpose for structure mining is to extract previously unknown relationships between web pages.

Web usage mining is the process of automatic discovery and investigation of patterns in click streams and associated data collected or generated as a result of user interactions with web resources on web sites. The main goal of web usage mining is to capture, model and analyze the behavioural patterns and profiles of users interacting with web sites. The discovered patterns are usually represented as collection of pages, objects or resources that are frequently accessed by groups of users with common needs or interests. The primary data resources used in web usage mining are log files generated by web and application servers. The overall Web usage mining process can be divided into mainly three interdependent stages: Pre-processing, Pattern Discovery & Pattern Analysis as shown in figure 2.

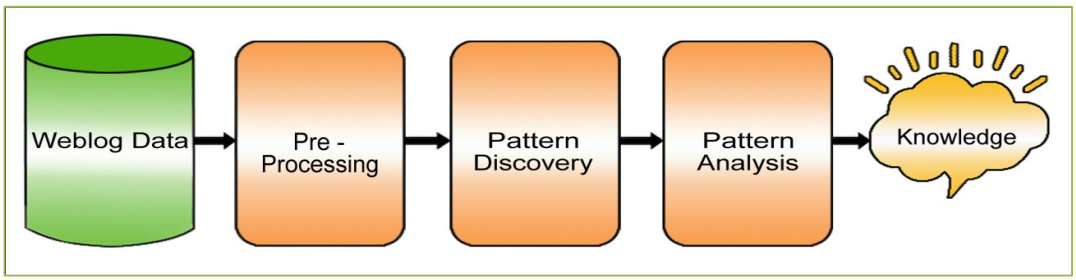

Figure 2. Stages of web usage mining

Pre-processing is Initial and very important stage in web usage mining applications in the creation of suitable target data set to which mining algorithms can be applied. The inputs for pre processing stage are the Web Server Logs, Referral Logs, Registration Files and Index Server Logs. In the pre processing stage, the click stream data is cleaned and portioned into a set of user transactions representing the activities of each user during different visits to the sites. The pre processing stage ultimately results in a set of user sessions, each corresponding to a delimited sequence of page views. However, in order to provide the most suitable data for further stages of web usage mining, Pre processing is an aspect of data mining whose importance should not be underestimated. Unless this phase is performed adequately, it is not possible for mining algorithms to provide reliable results.

Pattern discovery is the second stage of web usage mining process which can take the output generated by pre processing stage. The goal of pattern discovery is the stage of learning some general concepts from the pre processed data. In this phase, statistical, data base and machine learning techniques like classification, clustering and association rule mining are performed on the extracted information. This stage obtains possibly hidden patterns reflecting the typical behaviour of users, as well as summary statistics on web resources, sessions and users.

Pattern analysis is the final stage of usage mining which can extract interested patterns from the output of pattern discovery. The goal of pattern analysis is the task of understanding, visualizing, and interpreting the discovered patterns and statistics. The most common approaches for the pattern analysis are knowledge query mechanism such as SQL and OLAP. Any method used in the pattern analysis stage assumes the output of the previous phase that has already been structured. The output generated by pattern analysis is used as input to the various applications such as recommendation engines, visualization tools and web analytics. The types and levels of analysis, performed on the integrated usage data, depend on the ultimate goals of the analysts and the desired outcomes.

Due to the dynamic and unrealistic environment of web log data, the existing systems find difficulty in handling the newly emerged problems throughout all the phases of web usage mining especially, the pattern discovery. To proceed towards web intelligence, reducing the need of human intervention, it is necessary to integrate and entrench artificial intelligence into web mining tools. To achieve the intelligence, soft computing methodologies seem to be a good candidate.

The soft computing models are characterized by its 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 modelling highly non linear data, where the pattern discovery, rule generation and learnability are typical. The Fuzzy Logic, Artificial Neural Networks, Genetic Algorithms 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 of optimization type algorithms for both supervised and unsupervised techniques. The Genetic Algorithm is an elegant, 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. Moreover, the GAs minimize the assumptions in making the real study of web user usage behaviour. Thus, in order to find the optimized solution for investigating the web user usage behaviour, it is essential to make use of Genetic Algorithms with granular computing nature.

The goal of present paper is to deploy Intelligent Optimal Genetic Model IOGM, in order to find the optimized solutions for investigating the web user usage behaviour. IOGM is modelled on the principles of natural genetic systems, where the genetic information of each individual is encoded in structures called chromosomes. Each chromosome has an associated fitness value, which indicates its degree of goodness with respect to the solution it represents. Biologically inspired operators like selection, crossover and mutation are applied on the chromosomes in the population to yield potentially optimal solutions.

This paper is organized as follows. In section 2, related work is described. In next section 3, proposed work is presented in detail. In subsequent section 4, the experimental analysis of proposed work is shown. Finally in section 5 conclusions are mentioned.

2. RELATED WORK

The literature survey is conducted by the authors for the proposed work from 1999 to current year that comprises of major advances in the field of molecular biology, coupled with advances in genomic technologies have lead to an explosive growth in biological information generated by scientific community. It is an interdisciplinary field involving Biology, Computer Science, Mathematics and Statistics to analyze biological sequence data.

In the year 1999, M. Martin-Bautista and M. A. Vila [13] reviewed the genetic features in the data mining issues. They expressed that application of soft mining techniques to data mining and knowledge discovery is necessary in order to enhance the effectiveness of traditional data mining techniques. And they presented the features of genetic algorithms and its parameters.

In the next year 2000, H. Kargupta and S. Bandyopadhyay [7] explored the importance of linkage learning in genetic algorithms and other optimization algorithms. Linkage learning deals with the issue of intelligence acquired by the properties of representation. They developed the foundation to implement the genetic algorithms. They also identified the future directions of linkage learning in genetic algorithms.

S. Bandyopadhyay and S. K. Pal [22] presented in 2001 the analogy of concept of chromosome differentiation and variable string length in genetic algorithms. They also explained the integration of genetic algorithms to the classification problems. They demonstrated the effectiveness of GA with a variable string length classifier – VGACD.

During 2002, F. Picarougne, N. Monmarche, A. Oliver, and G. Venturini [4] presented a genetic search strategy for a search engine problem. They have defined search strategy which implements as directly as possible the concepts used in evolutionary algorithms. Their experiments have shown the relevance of genetic approach in search engine problems. They used a standard GA and they have directed towards advance algorithms with additional evolutionary techniques in their perspectives. Freitas, A.A [5] they discussed the use of evolutionary algorithms, particularly genetic algorithms in data mining and knowledge discovery. They also focused in the data mining task of classification. They felt that design evolutionary algorithms are a promising research direction in the knowledge discovery process as mentioned in their future research directions. H. K. Tsai, J. M. Yang, and C. Y. Kao. [8] demonstrated the usage of GA in finding the optimal global strategies by using clustering technique on biological datasets. Sankar K. Pal, Fellow [21] a deep survey of the existing literature on soft web mining along with commercially available systems. They also summarized the different types of web mining and its components. They illustrated the relevance of soft computing including GAs with examples. The extensive bibliography provided by them is evidence the relevance of usage of GAs in web mining. Sankar K. Pal, Fellow [30] clearly mentioned that the incorporation of computational intelligence techniques for mining the web is a future research direction.

Within the year 2003, B. Minaei-Bidgoli, William F. Punch [1] presented an approach for classifying the students in order to predict their final grade based on features extracted from log data of an education web based system. To minimize the prediction error rate they used genetic algorithms by weighting the features. They profoundly expressed applying the evolutionary algorithms for association rules to improve prediction accuracy as their future work. C. Lopez-Pujalte, V. P. G. Bote, and F. de Moya Anegon [3] evaluated the efficacy of genetic algorithm for various order based fitness functions that are designed as measures of goodness of the solutions. K. C. Wiese and E. Glen [12] presented a genetic algorithm to predict the secondary structure of RNA molecule where secondary structure is encoded as a permutation. More specifically, they focused at selection strategies, crossover operators and selection procedures. They also provided comparison study of several crossover operators.

All the range in 2004, B. Minaei-Bidgoli, G. Kortemeyer, and W. F. Punch [2] extended their previous work [1] and presented a new approach for predicting students' performance based on extracting the average of feature values for overall of the problem. Romao, W., Freitas, A. and Gimenes I [20] proposed a genetic algorithm designed specifically to discover of knowledge in form of prediction. A prototype was implemented and applied to a real world science and technology database. In addition, it works with natural linguistic terms which lead to the discovery of more comprehensible knowledge. S. Bandyopadhyay, S. K. Pal, and B. Aruna [23] developed a pattern classification methodology based on multi objective variable length GA to solve the problem of generating class boundaries. The usage of this classifier in other domains is a future work. S. J. Louis and J. McDonnell [24] provided a frame work for evaluating machine learning systems. S. Hill, J. Newell, and C. O'Riordan [25] proposed an inversion operator in a basic genetic algorithm and also compare the effectiveness. The results shown by them indicated that inversion operator is more useful in genetic algorithms with fitness scaling.

S. Y. Wang and K. Tai [28] implemented a bit array representation method for structural topology optimization using the genetic algorithm in 2005. An identical initialization method is also proposed to improve the GA performance.

In the subsequent year 2006, S. Y. Wang, K. Tai, and M. Y. Wang [29] presented a versatile, robust and enhanced genetic algorithm for structural topology optimization using problem specific knowledge. In their implementation process specifically pronounced the importance of choosing appropriate representation techniques, genetic operators and evaluation methods.

During 2008 and 2009, S. Ventura, C. Romero, A. Zafra, J. A. Delgado, and C. Hervas [27] designed a framework that can apply to maximize reusability and availability of evolutionary computation with a minimum effort in web mining. They also implemented a graphical user interface Gen Lab that allows users to configure an algorithm, to execute it interactively and to visualize the results obtained. The heavily demanding computational performance is an open problem as enmarked in their future research work.

The present IOGM frame work concentrating on assigning the weights to the characteristics of initial population, which minimizes the error rate in identifying the relation among the pages. Data analysis tools used earlier in web mining were mainly based on statistical techniques like regression and estimation. Recently GAs have been gaining the attention of researchers for solving certain web mining problems with the need to handle large data sets in computationally efficient manner.

3. PROPOSED INTELLIGENT OPTIMAL GENETIC MODEL: IOGM

While investigating the web user usage behaviour, the web miners have to take many intelligent decisions at each stage of web user usage mining. All such decisions have to be made sequentially at different levels within time. Since the web log data impacted by many external and explicit functions, therefore the data in the web log becomes non linear and the problem space turn into so dynamic. Consequently, the programming techniques used for this type of the problem happen to 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 with & without constraints
- Dynamic Programming
- Stochastic Programming
- Soft computing Techniques

Out of which a the soft computing technique, Genetic Algorithm used for IOGM, which is well suited to solve the problem of investigation of web user usage behaviour since the weblog data characterized by both continuous and discontinuous functions. To get the optimum solution for investigating the web user usage behaviour it is necessary to be processed 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 more adaptable to incremental web log scenario. The architecture and the process of IOGM are as shown in Figure 3.

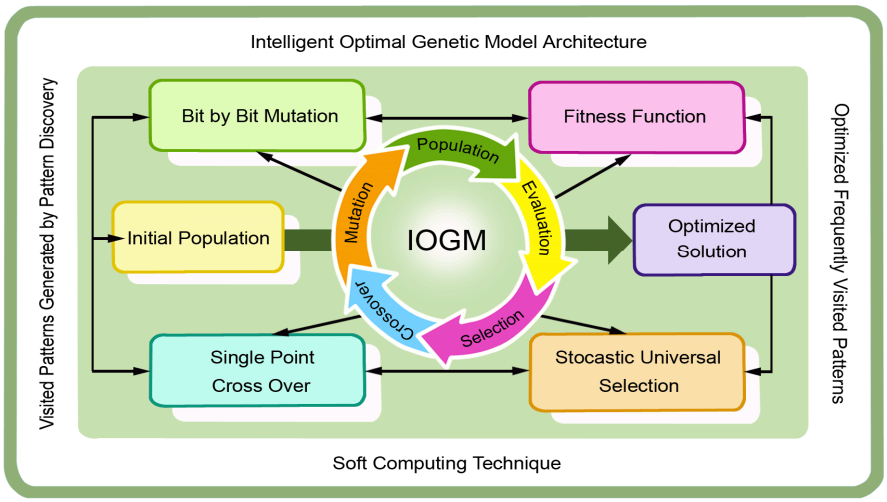

Figure 3. Architecture of IOGM

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 fixed time limit of IOGM end function.

3.1 IOGM encoding strategy – Initial population:

The encoding strategy is a process of representing the potential solution to a problem into a suitable form of viable individuals so that the genetic algorithm can process. It is 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 4. This can be considered as input for IOGM.

Ending with P5: {P5}, {P4, P5}, {P1, P4, P5}, {P3, P5}, {P1, P5}
Ending with P4: {P4}, {P3, P4}, {P2, P3, P4}, {P1, P3, P4}, {P2, P4}, {P1, P2, P4}, {P1, P4}
Ending with P3: {P3}, {P2, P3}, {P1, P2, P3}, {P1, P3}
Ending with P2: {P2}, {P1, P2}
Ending with P1: {P1}

Figure 4. Sample 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 5. 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.

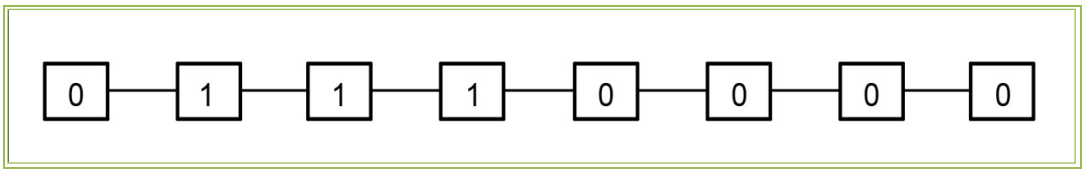

Figure 5. Sample Binary chromosome

3.2 IOGM Evaluation function – Fitness 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, D_c indicates number of genes in form of visiting patterns included in each chromosome and N_c indicates the number of chromosomes.

Thus the population $Np = Dc.Nc$ 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,

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

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 m*inimum support*.

$$\text{Term 2 :} \quad T_2(C) = \frac{Np}{\text{abs}(|C| - S)} + 1 \qquad 3$$

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.

$$\text{Term3:} \quad T_3(C) = \sum_{G_i, G_j \in C, G_i \neq G_j} D(G_i, G_j) \qquad 4$$

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

$$FF(C) = \alpha . T_1(C) + \beta . T_2(C) + \gamma . T_3(C) \qquad 5$$

Where α, β and γ are parameters that depend on the magnitude of the chromosome that represent the genes. In particular α, β and γ are chosen so the contributions given by T1(C), T2(C) and T3(C) are balanced. This approach is used by the evolution function of IOGM for every chromosome C*, by means of the genetic operators.

$$FF(C*) = \max_{c=1,\ldots,Nc} FF(C) \qquad 6$$

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 permit 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 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 marker1 in the range [0, $1/N_p$] is generated. The Np individual chromosomes are then selected by generating the N_p markers, starting with marker1 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 spins by choosing minimum times of LBET(i) and within the UBET(i). 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.

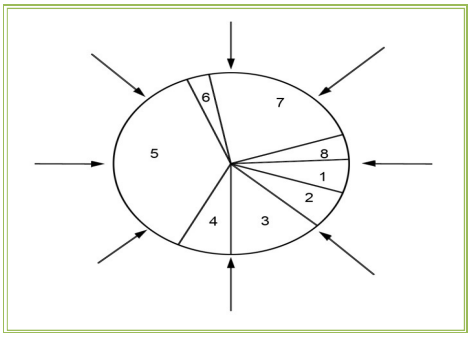

Figure 6. Stochastic universal selection

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 colours is as shown in figure 7(a).

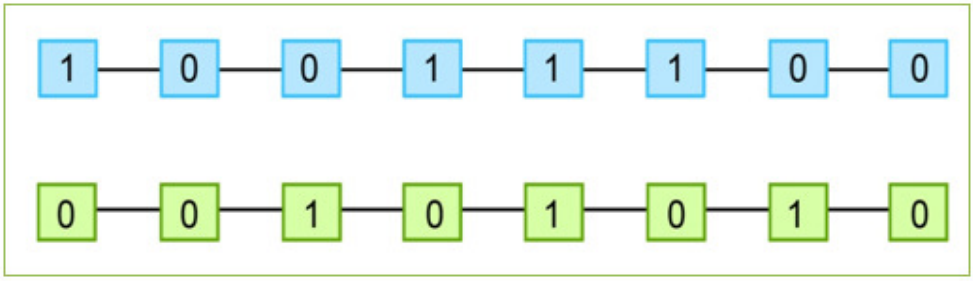

Figure 7(a) Pair of chromosomes

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 7(b).

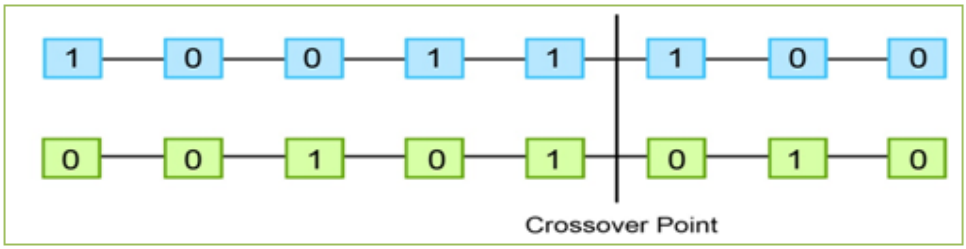

Figure 7(b) single point crossover operation before crossover

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 7(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.

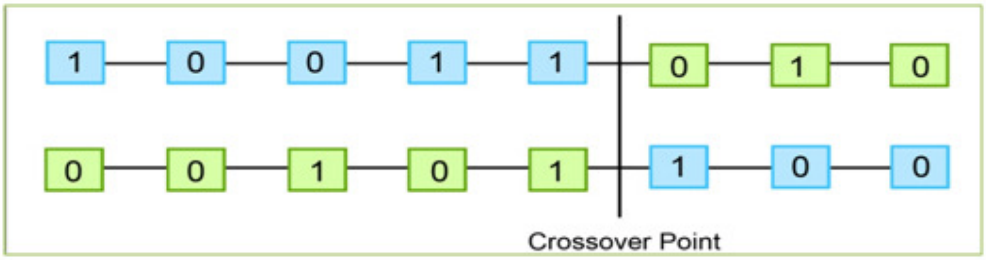

Figure 7(c) single point crossover operation after crossover

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 8.

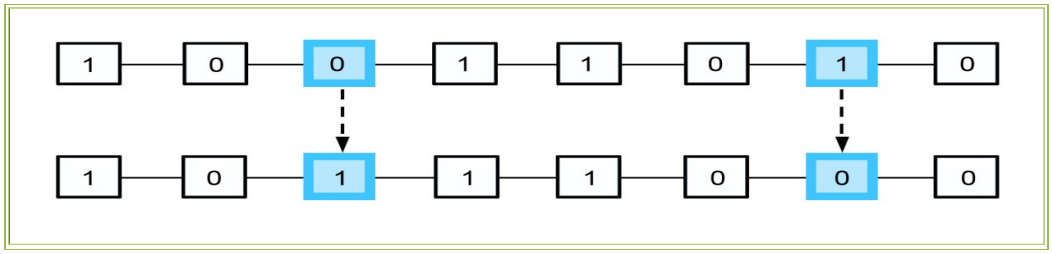

Figure 8. Process of bit-by-bit mutation

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.

3.4 IOGM End function:

The IOGM results the best chromosomes as the process goes to the 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 methodology of the end function that works by considering population size N_p, fitness function value FF, length of gene Length(Gi) , crossover probability P_c and mutation probability μ_m . It interrupts IOGM process when no significant or sufficient improvement found in two or more consecutive generations. In some cases, IOGM uses *time* threshold set by user, as a criteria for ending the process.

3.5 IOGM Algorithm:

Step 01:	Define the Evaluation function F.
Step 02:	t ← 0 (iteration No =0, pop size =0)
Step 03:	Initialize P (t).
Step 04:	Evaluate P(t) (page from modeled data).
Step 05:	Generate an offspring page O.
Step 06:	t ← t+1 (new population).
Step 07:	Select P (t) from P (t-1).
Step 08:	Crossover P (t).
Step 09:	Mutation P(t)
Step 09:	Evaluate P (t).
Step 10:	Go To 5 (while no termination (no of iterations)).
Step 11:	End Condition P (t) Gives the output to the user.

The IOGM 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 discovered by pattern discovery technique. The biologically inspired operators adjust the probabilities of selection and crossover so that IOGM exhibits self learning capability to carryout in producing 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 behaviour intelligently.

4. EXPERIMENTAL ANALYSIS

The proposed IOGM is experimented over a period of six 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 the pattern discovery technique.

A) The IOGM compared with the standard web mining technique in terms of performance. The experimental results indicate that noticeable improvement of IOGM performance over the standard web mining technique as shown in figure 9.

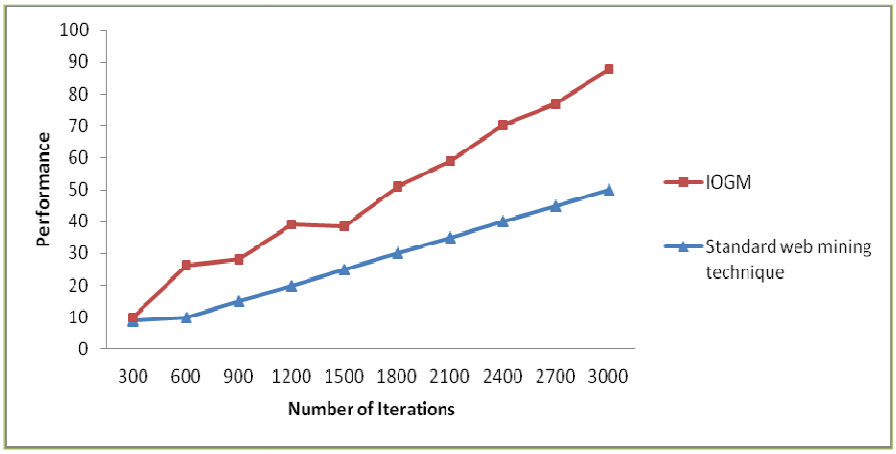

Figure 9. Comparison of performance of IOGM Vs standard web mining technique

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 10.

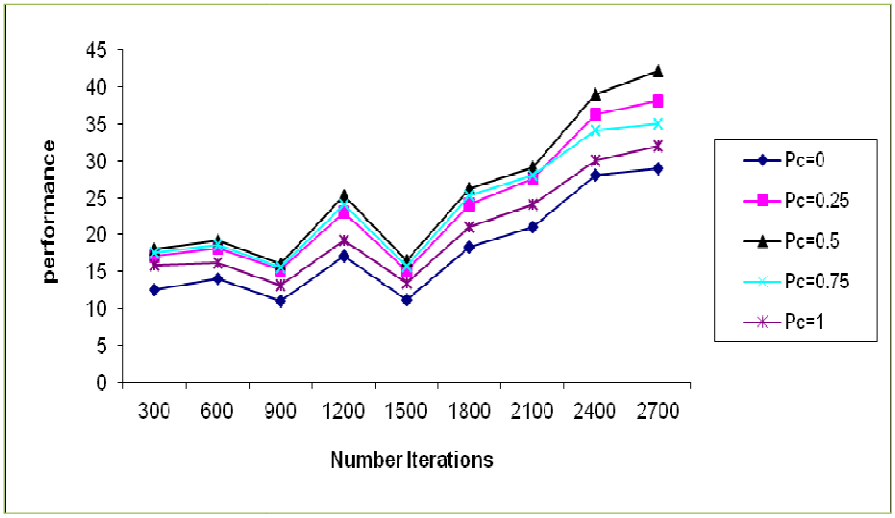

Figure 10. Comparison of performance at different crossover probabilities

C) In addition, the standard analysis algorithms are applied on the collective output generated by IOGM, the web user usage behaviour is identified as shown in Fig 11.

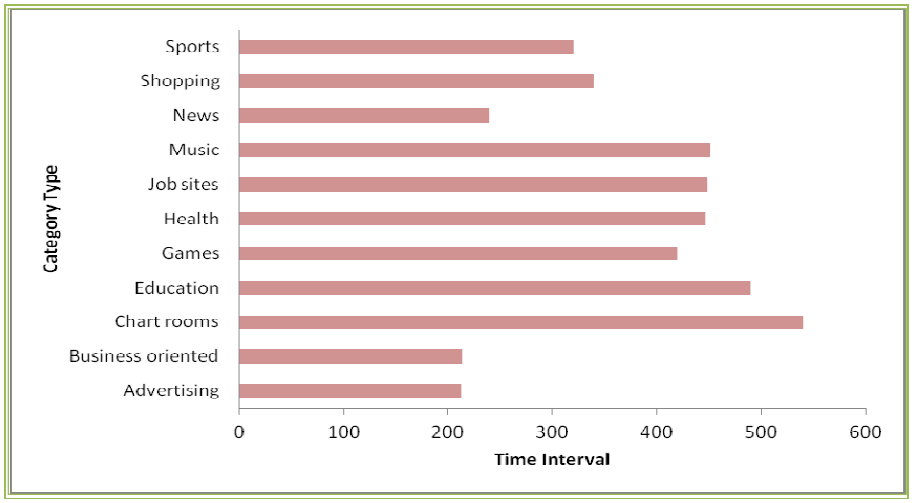

Figure 11. Web user usage behaviour

5. CONCLUSIONS

The present model has proven the relevance of soft computing methodologies in the identification of the desired patterns. The results are evident that the proposed Intelligent Optimal Genetic Model has a promising future to arrive at optimal solution intelligently in the web usage mining. The encoding strategy, evaluation function, fitness function, operators and end function of proposed IOGM works together and achieve sustained adaptability to characterize web user usage behaviour in unpredictable external environment. The nature of biological diversity of IOGM prevents the population from stagnating at any local solution. Moreover, the stochastic process of IOGM, assures the optimal solution always. Hence the proposed IOGM helps to investigate user usage behaviour on WWW in any application domain.

ACKNOWLEDGEMENTS

The authors record their acknowledgements to the authorities of Shri Vishnu Engineering College for Women, Bhimavaram; Andhra University, Visakhapatnam and Acharya Nagarjuna University, Guntur for their constant support and cooperation.

REFERENCES

[1] B. Minaei-Bidgoli, William F. Punch “Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System”, http://www.lon-capa.org, 2003.

[2] B. Minaei-Bidgoli, G. Kortemeyer, and W. F. Punch. “Optimizing classification ensembles via a genetic algorithm for a Web-based educational system”, In Proceedings of Joint International Association for Pattern Recognition (IAPR) Workshops on Syntactical and Structural Pattern Recognition (SSPR 2004) and Statistical Pattern Recognition (SPR 2004), 2004.

[3] C. Lopez-Pujalte, V. P. G. Bote, and F. de Moya Anegon., “Order-based fitness functions for genetic algorithms applied to relevance feedback”, Journal of the American Society for Information Science and Technology, 54(2), pp: :152-160, 2003.

[4] F. Picarougne, N. Monmarche, A. Oliver, and G. Venturini., “Web mining with a genetic based algorithm”, In Eleventh International World Wide Web Conference, Honolulu, Hawaii, pp: 7-11, 2002.

[5] Freitas, A.A., “A survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery”, See: www.pgia.pucpr.br/~alex/papers. A chapter of: A. Ghosh and S. Tsutsui. (Eds.) “Advances in Evolutionary Computation”. Springer-Verlag, 2002.

[6] H. Kargupta. "The gene expression messy genetic algorithm", In Proceedings of the IEEE International Conference on Evolutionary Computation, pp: 631- 636, 1996.
[7] H. Kargupta and S. Bandyopadhyay. "A perspective on the foundation and evolution of the linkage learning genetic algorithms", The Journal of Computer Methods in Applied Mechanics and Engineering, Special issue on Genetic Algorithms, 186, pp: 266-294, 2000.
[8] H. K. Tsai, J. M. Yang, and C. Y. Kao., "Applying genetic algorithms to finding the optimal Gene order in displaying the microarray data", In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), pp: 610-617, 2002.
[9] H. K. Tsai, J. M. Yang, Y. F. Tsai, and C. Y. Kao., "An evolutionary approach for gene expression patterns", IEEE Transactions on Information Technology in Biomedicine, 8(2), pp: 69-78, 2004.
[10] Hyunchul Ahn, Kyoung-jae Kim, "Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach, Applied Soft Computing, Volume 9, Issue 2, pp: 599–607, 2009.
[11] J. Han and K. C. Chang. "Data mining for Web intelligence", IEEE Computer, pp: 54-60, November 2002.
[12] K. C. Wiese and E. Glen., "A permutation-based genetic algorithm for the RNA folding problem: A critical look at selection strategies, crossover operators, and representation issues", Biosystems, 72(1-2), pp: 29-41, 2003.
[13] M. Martin-Bautista and M. A. Vila, "A survey of genetic feature selection in mining issues," in Proc. Congr. Evol. Comput. (CEC99), pp: 13–23, 1999.
[14] Mehmet Kaya, "Automated extraction of extended structured motifs using multi-objective genetic algorithm" Expert Systems with Applications, Volume 37, Issue 3, pp: 2421-2426, 2010.
[15] M. F. Bramlette, "Initialization, mutation and selection methods in genetic algorithms for function optimization" In Proceedings of the 4th International Conference on Genetic Algorithms, Morgan Kaufmann, San Mateo, pp: 100-107, 1991.
[16] M. H. Marghny and A. F. Ali, "Web mining based on genetic algorithm" In Proceedings of ICGST International Conference on Artificial Intelligence and Machine Learning, 2005.
[17] Mitsuo, G., Runwei, C., "Genetic algorithm & Engineering Optimization", John Wiley & Sons 2000.
[18] Mu-Jung Huang, Hwa-Shan Huang and Mu-Yen Chen, "Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach", Expert Systems with Applications, Volume 33, Issue 3, pp: 551-564, 2007.
[19] Pei, M., Goodman, E.D., and Punch, W.F. "Pattern Discovery from Data Using Genetic Algorithms", Proceeding of 1st Pacific-Asia Conference Knowledge Discovery & Data Mining (PAKDD-97), 1997.
[20] Romao, W., Freitas, A. and Gimenes, I. "Discovering interesting knowledge from a science and technology database with a genetic algorithm", Applied Soft Computing 4(2), pp: 121–137, 2004.
[21] Sankar K. Pal, Fellow, IEEE, Varun Talwar "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions" IEEE transactions on neural networks, vol. 13, no. 5, 2002.
[22] S. Bandyopadhyay and S. K. Pal. "Pixel classification using variable string genetic algorithms with chromosome differentiation", IEEE Transactions on Geoscience and Remote Sensing, 39(2), pp: 303-308, 2001.
[23] S. Bandyopadhyay, S. K. Pal, and B. Aruna. "Multi-objective GAs, quantitative indices and pattern classification", IEEE Transactions Systems, Man and Cybernetics - B, 34(5), pp: 2088-2099, 2004.
[24] S. J. Louis and J. McDonnell, "Learning with case injected genetic algorithms", IEEE Transactions on Evolutionary Computation, 8(4), pp: 3160-328, 2004.
[25] S. Hill, J. Newell, and C. O'Riordan, "Analysing the effects of combining fitness scaling and inversion in genetic algorithms". In Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04), pp: 380-387, 2004.
[26] Shital C. Shah, Andrew Kusiak, "Data mining and genetic algorithm based gene/SNP selection", Elsevier, Artificial Intelligence in Medicine 31, pp: 183-196, 2004.
[27] S. Ventura, C. Romero, A. Zafra, J. A. Delgado, and C. Herv´as, "A java framework for evolutionary computation soft computing." Soft Computing, vol. 4, no. 12, pp: 381–392, 2008.
[28] S. Y. Wang and K. Tai. "Structural topology design optimization using genetic algorithms with a bit-array representation", Computer Methods in Applied Mechanics and Engineering, 194, 3749-3770, 2005.
[29] S. Y. Wang, K. Tai, and M. Y. Wang. "An enhanced genetic algorithm for structural topology optimization", International Journal for Numerical Methods in Engineering, 65, pp: 18-44, 2006.

[30] Wang, X, A Abraham and KA Smith, "Soft computing paradigms for web access pattern analysis". In Proc. of the 1st International Conference on Fuzzy Systems and Knowledge Discovery, pp: 631 – 635, 2002.

[31] W. Fan, E. A. Fox, P. Pathak, and H. Wu, "The effects of fitness functions on genetic programming-based ranking discovery for web search", Journal of the American Society for Information Science and Technology, 55(7), pp: 628 - 636, 2004.

Authors

Prof. V.V.R. Maheswara Rao received his Master of Computer Applications degree from Osmania University, Hyderabad, India. He is working as Professor in the Dept of Computer Applications at SVECW, Bhimavaram, AP, India. He is currently pursuing his Ph.D. in Computer Science Engineering at Acharya Nagarjuna University, Guntur, India. His Research interests include Webmining, Artificial Intelligence. He is member of CSI, CI and ISTE.

Dr. V. Valli Kumari holds a PhD degree in Computer Science and Systems Engineering from Andhra University Engineering College, Visakhapatnam and is presently working as Professor in the same department. Her research interests include Security and privacy issues in Data Engineering, Network Security and E-Commerce. She is a member of IEEE and ACM.