An Advanced Optimal Web Intelligent Model for Mining Web User Usage Behavior using Genetic Algorithm

Prof. V.V.R. Maheswara Rao¹, Dr. V. Valli Kumari² and Dr. K.V.S.V.N Raju²,
¹Shri Vishnu Engineering College for Women, Bhimavaram, India,
Email: mahesh_vvr@yahoo.com
²College of Engineering, Andhra University, Visakhapatnam, India,
Email: vallikumari@gmail.com, kvsvn.raju@gmail.com

Abstract—With the continued growth and proliferation of Web services and Web based information systems, the volumes of user data have reached astronomical proportions. Analyzing such data using Web Usage Mining can help to determine the visiting interests or needs of the web user. This type of analysis involves the automatic discovery of meaningful patterns, which represents fine grained navigational behavior of visitors from a large collection of semi structured web log data. Due to the non linear and complex nature of weblog all the existing conventional mining techniques are failed in the process of pattern discovery, result in only local optimal solutions. In order to get the global optimal solutions, Web intelligent models are required such as Genetic algorithms.

The present paper introduces the Advanced Optimal Web Intelligent Model with granular computing nature of Genetic Algorithms, IOG. The IOG model is designed on the semantic enhanced content data, which works more efficiently than that on normal data. The unique parameters of the IOG fitness function generates balanced weights, to represent the significance of characteristics, which yields the dissimilar characteristics of page vector. The evaluation function of IOG improves the page quality and reduces the execution time to a specified number of iterations as it considers practical measures like page, link and mean qualities. The genetic operators are intended in drawing the stickiness among the characteristics of a page vector. By integrating all the above the web intelligent model IOG, proceed towards intelligence and significantly improves the investigation of web user usage behavior.

Index Terms—Web Usage Mining, Genetic Algorithm, Pattern Discovery

I. INTRODUCTION

Over the last decade, with the continued increase in the usage of the WWW, web mining has been established as an important area of research. Whenever, the web users visit the WWW, they leave abundant information in web log, which is structurally complex, heterogeneous, high dimensional and incremental in nature. Analyzing such data can help to determine the browsing interest of web user. To do this, web usage mining focuses on investigating the potential knowledge from browsing patterns of the users and to find the correlation between the pages on analysis. The main goal of web usage mining is to Capture, Model and Analyze the web log data in such a way that it automatically discovers the usage behavior of web user.

The overall web usage mining process can be divided into Four interdependent stages as shown in Fig 1: Data collection, Pre Processing, Pattern Discovery and Pattern Analysis.

Fig.1.Web mining subtasks.

Due to the dynamic and complex nature of web log data, the existing systems find difficulty in handling the newly emerged problems during all the phases of web usage mining in particular, the pattern discovery. 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 soft computing methodologies seem to be a good candidate.

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 way for prediction and description of complex objective functions like Fitness and Evaluate in dynamic environment. Moreover, GA employs a set of operators that mimic the concept of survival of Fittest by regenerating recombination of the algorithm in response to a calculated difference between desired solution states.

The goal of present paper is to deploy Intelligent Optimal Genetic Algorithm IOG, in order to find the optimized solutions for investigating the web user usage behavior. The important task in any intelligent mining application is the preparation of a suitable target data set for which mining and statistical algorithms can be applied. For IOG, the present research frame work, Firstly, concentrating on modelling the web log data due to its complex nature. The main thrust of the data modelling is to generate the semantic enhanced content data. The integration of semantic content with IOG can potentially
provide a better understanding of the underlying relationships among the pages.

The IOG Fitness function is designed in such a way that, it generates dissimilar page vectors, each vector have similar characteristics. In addition, the IOG 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 IOG frame work collectively uses Fitness, Operators and Evaluation functions, helps to investigate the web user usage behavior significantly.

The present paper is organized as follows. The related work described in section 2. In next section 3, the overview of proposed work is introduced. In subsequent section 4, the experimental analysis is shown. In subsequent section 5, the conclusions are made. Finally in section 6 acknowledgements are mentioned.

II. RELATED WORK

Over the past two decades, 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.

Genetic algorithms have been showed to be an effective tool to use in data mining and pattern Discovery [3], [4], [5], [6], [7]. Sankar K Pal, Varun Talwar [1] introduced several soft compute methodologies and deliberately expresses the relevance of genetic algorithms in the web log scenario. William F. Punch [2] used the genetic algorithm optimization technique through feature selection. In this method the error rate in data mining techniques in local optimization shows better performance rather than in global optimization.

The present IOG 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.

III. OVERVIEW OF THE PROPOSED WORK

The web usage mining is a task of applying data mining techniques to extract useful patterns from web log data in various stages as shown in Fig 2. These patterns can be used to investigate interesting characteristics of web users. Generally the web log data is complex and non-linearly growing in nature, it is necessary for web miners to utilize soft computing tools like Genetic Algorithms in order to get global optimal information intelligently. GA employs a set of operators that mimic the concept of survival of Fittest by regenerating recombination of the algorithm in response to a calculated difference between desired solution states over a period of time.

To investigate the usage behavior of the web user, one has to extract the potential patterns through the sessions, generated by the pre processing stage of web log. Directly the data can’t be given as an input to the genetic algorithm, therefore the pre processed sessions data has to be modeled which is suitable for genetic environment. By the data modelling semantic enhanced content data is generated, which improves performance of the GA significantly.

3.1. Web log Data Modeling for IOG:

The web log pre-processing results in a set of n pageviews, 
P = \{p_1, p_2, \ldots, p_n\}, and a set of m user transactions, 
T = \{t_1, t_2, \ldots, t_m\}, where each t_i in T is a subset of P. Pageviews are semantically meaningful entities to which mining tasks are applied. Each transaction t can be conceptually viewed as an l-length sequence of ordered pairs

\[
t = ((p_{i1}, w(p_{i1})), (p_{i2}, w(p_{i2})), \ldots, (p_{il}, w(p_{il})))
\]

where each \(p_{ij} = p_i\) for some \(j\) in \{1, 2, \ldots, n\}, and \(w(p_i)\) is the weight associated with pageviews \(p_i\) in transaction \(t\), representing its significance. In web usage mining tasks the weights are, either binary representing the existence or non-existence of a pageview in the transaction or they can be a function of the duration of the pageview in the user’s session. In the case of time durations, it should be noted that usually the time spent by a user on the last pageview in the session is not available. Hence for the proposed frame work weights are associated based on binary representation.
For association rule mining the ordering of pageviews in a transaction is not relevant, one can represent each user transaction as a vector over the n-dimensional space of pageviews. Given the transaction vector $t$ (bold face lower case letter represents a vector) as:

$$t = (w^t_{p_1}, w^t_{p_2}, ..., w^t_{p_n})$$

where each $w^t_{p_j} = w(p_j^t)$ for some $j$ in $\{1, 2, ..., n\}$, if $p_j$ appears in the transaction $t$, and $w^t_{p_j} = 0$ otherwise. Thus, conceptually, the set of all user transactions can be viewed as an $m \times n$ User-PageView Matrix, denoted by UPVM. An example of a hypothetical user-pageview matrix is depicted in Fig. 3. In this example, the weights for each pageview is the amount of time (e.g., in seconds) that a particular user spent on the pageview.

An association or sequential pattern mining techniques can be applied on UPVM as described in example to obtain patterns and in turn these patterns are used to find important relationships among pages based on the navigational patterns of users in the site.

As noted earlier, it is also possible to integrate other sources of knowledge, such as semantic information from the content of web pages with the web usage mining process. Generally, the characteristics from the content of web pages reflects behavior of web user. Each pageview $p$ can be represented as an $r$-dimensional characteristics vector, where $r$ is the total number of extracted characteristics from the site in a global dictionary. The vector, denoted by $p$, can be given by:

$$p = (f^p(w_{f_1}), f^p(w_{f_2}), ..., f^p(w_{f_n}))$$

where $f^p(w_{f_j})$ is the weight of the $j^{th}$ characteristic in pageview $p$, for $1 \leq j \leq r$. For the whole collection of pageviews in the site, one can represent an $n \times r$ PageView - Characteristic Matrix $PVCM = \{p_1, p_2, ..., p_n\}$.

The integration process, involve the transformation of user transactions in UPVM into “content-enhanced” transactions containing the semantic characteristics of the pageviews. The goal of such a transformation is to represent each user session as a vector of characteristics rather than as a vector over pageviews. In this way, a user’s session reflects not only the pages visited, but also the significance of various characteristics that are relevant to the user’s interaction. While, in practice, there are several ways to accomplish this transformation, the most direct approach involves mapping each pageview in a transaction to one or more content characteristics. The range of this mapping can be representing the set of characteristics. Conceptually, the transformation can be viewed as the multiplication of UPVM with PVCM. The result is a new matrix $TCM = \{t_1, t_2, ..., t_n\}$, where each $t_i$ is a $r$-dimensional vector over the set of characteristics. Thus, a user transaction can be represented as a content characteristics vector, reflecting the user’s interests.

As an example of content-enhanced transactions consider Fig.4 which shows a hypothetical matrix of user sessions (UPVM) as well as an index for the corresponding Web site conceptually represented as a term-pageview matrix (TPVM). Note that the transpose of this TPVM is the pageview-characteristic matrix (PVCM). The UPVM simply reflects the pages visited by users in various sessions. On the other hand, the TPVM represents the concepts that appear in each page. For simplicity the weights are assumed with binary values.

The corresponding Characteristics-Enhanced Transaction Matrix CETM (derived by multiplying the UPVM and the transpose of the TPVM) is depicted in Fig. 5. The resulting data is very much suitable for the proposed IOG as initial population.
3.2. Intelligent Optimal Genetic Algorithm IOG:

GAs are local optimal algorithms that start from an initial collection of pages (a population) representing possible solution to the problem. Each page characteristics are called a chromosome and has an associated value called fitness function (ff) that contributes in the generation of new population by means of genetic operators. Every position in a chromosome is called a gene, and its value is called the allelic value. This value may vary on an assigned allelic alphabet. Often, the allelic alphabet is $\{0, 1\}$. At each generation, the algorithm uses the fitness function values to evaluate the survival capacity of each page characteristics by using simple operators to create a new set of artificial creators (a new population) that tries to improve on the current ff values by using pieces of old ones. Evaluation is interrupted when no significant improvement of the fitness function can be obtained.

3.2.1. Proposed IOG Genetic Algorithm:

01: Define the Evaluation function F.
02: $t \leftarrow 0$ (iteration No =0, pop size =0)
03: Initialize P (t).
04: Evaluate P(t) (page from modeled data).
05: Generate an offspring page O.
06: $t \leftarrow t+1$ (new population).
07: Select P (t) from P (t-1).
08: Crossover P (t).
09: Mutation P(t).
10: Go To 5 (while no termination (no of iterations)).
11: Sort P (t) (sort the pages given to the user in descending order according to their support).
12: End Condition P (t) Gives the output to the user.

3.2.2. IOG – Fitness Function:

In the Fitness function, $dc$ indicates number of characteristics included in each chromosome and $nc$ indicates the number of chromosomes. The population with thus contain $np=dc.nc$ phases. The fitness function computed for each chromosome is expressed as a positive value that is higher for “better” chromosomes and is thus to be maximized. It is composed of three terms. The first term is the sum of the score of the pages in chromosome C, i.e., $t_1(C) = \sum_{p_i \in C} \text{score}(p_i)$ where score($p_i$) is the original score given to page $p_i$ as previously described. This term considers the positive effect of having as many pages with as high a score as possible in a chromosome but also rewards chromosomes with many pages. This drawback is balanced with second term of the Fitness function. Let ID be such an ideal dimension; the ratio $t_2(C) = \frac{np}{\text{abs}(|C| - ID) + 1}$ constitutes the second term of the fitness function. It reaches its maximum np when the dimension of C is exactly equal to the ideal dimension ID and rapidly decreases when the number of pages contained in chromosome C is less than or greater than ID.

The chromosomes that are present in the initial population are characterized by the highest possible variability as far as the page vectors to which the number of pages belong or concerned. The evaluation of the population may alter this characteristic, creating chromosomes with high fitness where the pages belong to the same vector and are very similar to each other. Moreover, the fact that pages belonging to different vectors are different in the vectorized space may not be guaranteed, as it depends both on the nature of the data and the quality of the initial Vectorization process. For this reason, fitness function third term is introduced, which measures directly the overall dissimilarity of the pages in the chromosomes. Let $D(p_i, p_j)$ be the Euclidean distance of the vectors representing pages $p_i$ and $p_j$. Then $t_3(C) = \sum_{p_i, p_j \in C, p_i \neq p_j} D(p_i, p_j)$ is the sum of the distance between the pairs of pages in chromosome C and the measures the total variability expressed by C. The final form of the fitness function for chromosome C is then

$$ff(C) = \alpha \cdot t_1(C) + \beta \cdot t_2(C) + \gamma \cdot t_3(C)$$

Where $\alpha$, $\beta$ and $\gamma$ are parameters that depend on the magnitude of the initial score and of the vectors that represent the pages. 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. Additionally, they may be tuned to express the relevance attributed to the different aspects represented by the three terms. The goal of the GA is to find, by means of the genetic operators, a chromosome $C^*$ such that

$$ff(C^*) = \max_{C=1,\ldots,nc} ff(C)$$

3.2.3. IOG Evaluation Function:

The Fitness function $ff$ that evaluates web pages is a mathematical formulation of the user query and numerous evaluation functions. First, let us define the followings in the simplest forms for practical considerations.
1) Link quality \( F(L) \)
\[
F(L) = \sum_{i=1}^{n} #K_i
\]

2) Page quality \( F(P) \)
\[
F(P) = \sum_{j=1}^{m} F_j(L)
\]

Where ‘m’ is the total number of links per page.

3) Mean quality function \( M_q \)
\[
M_q = \frac{F_{\text{max}}(P) + F_{\text{min}}(P)}{2}
\]

Where \( F_{\text{max}}(P) \) and \( F_{\text{min}}(P) \) are the maximum and minimum values of the page qualities respectively after applying the GA. It should be noted that the upper value of \( F_{\text{max}}(P) \) is \( m*n \), and the least value of \( F_{\text{min}}(P) \) is zero. Hence, the upper limit of \( M_q \) is \( (m*n)/2 \). Application of the GA to web pages will increase some qualities of pages and decrease others.

3.2.4. IOG – Operators:

The genetic algorithm uses crossover and mutation operators to generate the offspring of the existing population. Before genetic operators are applied, parents have been selected for evolution to the next generation. One can use the crossover and mutation algorithm and produce next generation. The probability of deploying crossover and mutation operators can be changed. The genetic operators work iteratively and are:

- **Reproduction**, where individual characteristics are copied according to their fitness function values (the higher the value of characteristic, the higher of the probability of contributing to one or more offspring in the next generation)

- **Crossover**, in which the members reproduced in the new mating pool are mated randomly and, afterward, each pair of characteristics undergoes a cross change

- **Mutation**, which is an occasional random alteration of the allelic value of a chromosome that occurs with small probability

3.2.5. IOG – End condition:

GA needs an End Condition to end the generation process. If no sufficient improvement found in two or more consecutive generations, one can stop the GA process. In other cases, one can use time limitation as a criterion for ending the process.

IV. EXPERIMENTAL STUDY

The proposed IOG has been implemented in a standard environment. For the IOG algorithm number dissimilar characteristic vectors are given as input. Theses vectors are generated by the weblog data modeling of IOG, considering a weblog consists of more than 3000 pages.

A) The IOG model compared with the standard web mining techniques with respective population Mean Quality. The experimental results indicate that noticeable improvement of IOG performance over the Standard mining Techniques as shown in Fig 6.

![Fig 6. Comparison of population quality of GA w.r.t standard DM Technique](image)

B) The IOG model compared with the standard web mining techniques with respective Execution. The experimental result shows the IOG is essentially taking less time when compared with Standard mining Techniques as shown in Fig 7.

![Fig 7. Comparison of execution of GA w.r.t standard DM Technique](image)

C) The IOG ran for different cross over probabilities \( P_c = 0, 0.25, 0.5, 0.75, 1 \). The average mean quality values are tabulated for ten different enhanced contents as shown in table 1.

<table>
<thead>
<tr>
<th>S. No</th>
<th>Contents</th>
<th>( P_c=0 )</th>
<th>( P_c=0.25 )</th>
<th>( P_c=0.5 )</th>
<th>( P_c=0.75 )</th>
<th>( P_c=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Web mining</td>
<td>12.5, 17.1</td>
<td>16, 17.5</td>
<td>15.9, 16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Data mining</td>
<td>14, 18.1</td>
<td>19.3, 18.7</td>
<td>16.2, 19.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Business</td>
<td>11, 15.3</td>
<td>16.1, 15.6</td>
<td>13.1, 15.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Marketing</td>
<td>17.1, 23.1</td>
<td>25.3, 24.1</td>
<td>19.2, 25.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Education</td>
<td>11.1, 15.1</td>
<td>16.5, 15.7</td>
<td>13.5, 15.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Intelligence</td>
<td>18.3, 24.1</td>
<td>25.3, 21</td>
<td>21.1, 23.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>DBMS</td>
<td>21, 27.6</td>
<td>28.1, 24.1</td>
<td>30.1, 28.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Security</td>
<td>28, 36.3</td>
<td>39.1, 34.2</td>
<td>30.1, 34.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Testing</td>
<td>29, 38.1</td>
<td>42.3, 35.1</td>
<td>29.1, 35.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Oracle</td>
<td>14, 19.1</td>
<td>21.1, 20.1</td>
<td>16.1, 20.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

D) The IOG model experimented for different cross over probabilities \( P_c = 0, 0.25, 0.5, 0.75, 1 \) for different no. of iterations. It is showing high performance at the average mean \( P_c \) as in Fig 8.
E) The IOG ran for different cross over probabilities (Pc=0, 0.25, 0.5, 0.75, 1) for different number of iterations. The IOG is executing with significantly less time at the average mean Pc as shown in Fig 9.

F) In addition, the standard analysis algorithms are applied on the collective output (desired patterns) generated by IOG, the web user usage interests are identified as shown in Fig 10.

V. CONCLUSION

The present model has proven experimentally the relevance of web intelligent methodologies in the identification of the desired patterns. Associating the balanced weights generated by IOG model worked significantly better than that technique of characteristic selection. According to the results the proposed technique is more effective in improving the page quality, Mean quality and reduces the execution time and hence the proposed IOG model helps to investigate the usage behavior of the web user in any application domain.

VI. ACKNOWLEDGEMENTS

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

REFERENCES


