IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 65-71
www.iosrjournals.org

Implementation of a Model for Web Mining Based on Web Usage

Jayaprabha P[1], Dr.Paulose Jacob[2] , Dr.Preetha Mathew[3], Bindu P K[4]
[1](Department of IT, CUSAT, India)
[2](Department of CSE, CUSAT, India)
[3](Department of CSE, CUSAT, India)
[4](Department of CSE, CUSAT, India)

Abstract: *Web search engines are used to retrieve relevant information from the web. Current search engines are very fast in terms of their response time to a user query. From the huge volume of information, extracting relevant information is an application of data mining. There are different techniques used for extracting relevant information, such as sequential rules, self organizing map, association rules, genetic algorithm, cluster computing etc. In this paper we have explode the possibility of genetic algorithm for extracting relevent information. Out of the lot of information's in the web, to get a few relevant ones with the help of fitness function and ranking function using genetic algorithm. The parameters for this method are extracted from the web usage and web content mining. Implementation results show that the proposed approach performed better than the existing Self-Organizing Map (SOM) network .*

Keywords: *Genetic Algorithm,Self-Organizing Map,Web search engines.*

I. Introduction

Web usage mining is the application of data mining techniques to discover usage patterns from web data. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Various approaches have been discussed in literature for web usage mining. These are association rules, apriori algorithm, sequential rules, self-organizing map, genetic algorithm and cluster computing[1].

The Self-Organizing Map (SOM) is one of the most popular neural network models. It belongs to the category of competitive learning networks based on unsupervised learning, In this case no human intervention is needed during the learning process and little need to be known about the characteristics of the input data. SOM is used for clustering the data without knowing the class. The SOM can be used to detect features inherent to the problem and it also called SOFM, the Self-Organizing Feature Map, It provides a topology preserving mapping from the high dimensional space to map units. Map units, or neurons, usually form a two-dimensional lattice and the mapping is from high dimensional space onto a plane. The property of topology preserving means that the mapping preserves the relative distance between the points. Points that are near to each other in the input space are mapped to nearby map units in the SOM. The SOM can thus serve as a cluster analyzing tool of high-dimensional data[3].

SOM is applied to preprocessed data in web usage mining in order to find visitor's navigation behavior. Web sites are organized based on the co-occurrence frequencies between web pages which are obtained by user access pattern. To reduce the search depth and information overload for users two constraints are used.

1. The number of outward links from each page.
2. The length of shortest path from home page to each page. Web personalization is the way of providing service to web visitor for retrieving the information of their interest. This is achieved by predicting the next page access by the user. Pair wise nearest neighbor clustering is used for identifying similar access pattern. This method provides good prediction accuracy and minimizes complexity[2].

GA is a natural selection theory based algorithm used for solving optimization problems. It is an adaptive heuristic algorithm based on concept of survival of the fittest. Selection, crossover, mutation and acceptance are the main steps used for finding the solution to a problem. Fitness function is used for finding the goodness of any solution and mutation escapes the population from problem of local optima. The input of the GA will be the extracted from the log file using data mining. GA consists of the following steps

(i) Chromosome Representation - The chromosomes are used for representing initial population. Each chromosome shows a candidate solution. For representing the web page we will assign a unique number id to each unique URL taken from web server log.

(ii) Fitness Function - Fitness function is an objective function used for selection of best individual among all individuals. It is used for quantifying the optimality of a solution. It measures the goodness of a solution by providing ranks to solution. (iii) Selection- Selection is the process of choosing the fitter chromosomes from the

population. The main objective of selection is to give importance to good solution and ignoring bad solution. We are using binary tournament selection which picks two individuals randomly from large set of population.

(iv) Crossover- Crossover is the method which exchanges the genetic material of both the parents to get new offspring. Main function of crossover is to recombine two strings to get a new better string.

(v) Mutation- Mutation is the third operator of GA that performs the function of maintaining diversity in the population by altering some bits present in the chromosome. It randomly distributes genetic information and avoids the probability of algorithm to suffer from the problem of local optima[4].

Mutation is applied after crossover in order to change genetic material between parents and forms offspring. Then on the basis of Darwinism (is a theory of biological evolution developed by Charles Darwin and others, stating that all species of organisms arise and develop through the natural selection of small, inherited variations that increase the individual's ability to compete, survive, and reproduce. Also called Darwinian theory, it originally included the broad concepts of transmutation of species or of evolution which gained general scientific acceptance when Charles Robert Darwin published On the Origin of Species, including concepts which predated Darwin's theories, but subsequently referred to specific concepts of natural selection or in genetics.) the offspring which survives most is chosen to be fittest.

In this paper we have used a collection of web page to represent chromosome in web usage mining problem. Inorder to find the web pages is of most importance to user. Unique number has been assigned to web pages. The pages taken from web server log. This number acts as an ID. Each chromosome shows a candidate solution. For further processing this unique no. ID is used instead of URL of the page visited by the user.

II. Background And Related Work

Web content mining and structure mining uses the real or primary data, but web usage mining (WUM)mines secondary data generated by the users' interaction with the web. Web usage data gets data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse-clicks and scrolls, and any other data generated by the interaction of users and the web. WUM works on user profiles, user access patterns, and mining navigation paths which are being heavily used by e-commerce companies for tracking customer behavior on their sites.

In addition to this learning access patterns, one needs to use "collaborative filtering" for listing other users with similar interests, which is an application of clustering. Collaborative recommender systems allow personalization for e-commerce by exploiting similarities and dissimilarities in users' preferences. A new algorithm is suggested in [5],[6] for specifically catering to association rule mining in collaborative recommendation systems.

It uses feature reduction techniques to reduce the dimension of the rating data and then NNs are applied on the simplified data to make a model for collaborative recommendation. However, the discovery of patterns from usage data by itself is not sufficient for performing personalization tasks. A way of deriving good quality and useful "aggregate user profiles" from patterns is suggested in [7]. It evaluates two techniques based on clustering of user transactions and clustering Recommender systems can be effectively utilise aggregate profiles for real-time personalization. A framework for web mining has been proposed in WEBMINER [8] for pattern discovery from WWW transactions.

The web creates new challenges to different task of web mining. The amonnt of information on the web is increasing and changing rapidly without any control. Therefore the existing systems find very difficult to handle the problems during information retrieval,extraction, generalization and analysis.

III. Proposed Work

In this paper, to improve the user navigation in a search engine by prioritizing web links based on web usage and content data using data mining techniques by implementing a new approach to Genetic Algorithm.
Fitness function is order-based, where the fitness value depends on the order in which the relevant documents are retrieved. For fitness function comparison purposes, we also use P_Avg which is defined as

$$P\_Avg = \frac{\sum_{i=1}^{|D|}\left(r(d_i)*\left(\frac{\sum_{j=1}^{i} r(d_j)}{i}\right)\right)}{T\ Re\ l} \quad (1)$$

Where r(d) (0,1) is the relevance score assigned to a document, being 1 if document is relevant and 0 otherwise. TRetr.d is the total number of retrieved documents. TRel is the total number of relevant documents for the query. The choice of fitness function is important for retrieval performance in the cases such as adhoc task for feedback queries routing task for short queries and routing task for feedback queries. A sample of the ranking fn. discovered by GA + S where S is the structural information within documents such as anchor, title,

abstract, body and so on. Ranking function is otherwise called retrieval function and is restricted to a polynomial regression function or a logistic/ log-linear function.
A sample of the ranking function discovery by GA is :

$$\left(\frac{tf\_Doc}{tf\_\max Doc} \times \frac{df\_\max Doc}{df\_Doc} \times \frac{length\_avg\_Abstract\_Col}{tf\_avg\_Abstract}\right) \quad (2)$$

where $(tf_{Doc})/(tf\_\max Doc)$ is the normalized token frequency,$(df\_\max Doc)/(df\_Doc)$ is the normalized inverse document frequency and $(length\_avg\_Abstract\_Col)/(tf\_avg\_Abstract)$ is the structural part of the ranking function.

The proposed Genetic Algorithm based approach combines the information from both content as well as usage of a web page in order to provide the required and relevant pages to user.
Various parameters are required for calculating the fitness of a solution as
Access frequency - which measures number of times a particular page is visited by user irrespective of user id.
Number of unique visitors - This factor shows the importance of any web page on the bases of unique visitors visited this page.
Time duration - The amount of time spent on a page shows the relevance of page for the user.
Number of bytes received - The quantity of data downloaded by user from the web page shows that page has content which is relevant for user.
Common entry and exit points - The entry point of the user begins his/her search by clicking on a link which forwarded him/her toward a page of web site. The exit point signifies the designation of the visitor. It tells what visitors are looking for in the website.
Number of advertisements- The importance of any web page can also be recognized by analyzing the number of advertisement present on any particular page. Advertisements are placed on the pages which have higher frequency of visits by user so they signify the importance of page.
A number of parameters present in the content and usage pattern of web links are included in the fitness function. Genetic Algorithm initiates by randomly selecting a set of initial population and then applying crossover and mutation on the population for many generations until the population gets converged and result is produced.

IV. Output

Code for Thresold Calculation :

```
public void calculateReleventThreshold() {
floatreleventThreshold = 0;
for (int i = 0; i <no_of_urls; i++) {
relevent_fitness[i] = (float) ((float) ((acces_count[i] *
Constants.acess_constant) + (duration[i] *
Constants.duration_constant) + (no_of_adds[i] *
Constants.no_of_adds_constant)) + (no_of_users[i] *
Constants.no_of_users_constant));
releventThreshold += relevent_fitness[i];
System.err.println("Relevent of " + i + "   :    " +
relevent_fitness[i]);
        }
releventThreshold = releventThreshold / no_of_urls;
releventThreshold = releventThreshold /4;
System.err.println("Relevent threshold " + "   :    " +
releventThreshold);
this.releventThreshold = releventThreshold;
    }
```

Code for Selection :

```
intselectedUidArr[];
intrequiredLength = 0;
requiredLength = uid_arr.length - (uid_arr.length % 4);
          //calculated required length
          Random rand = new Random();
selectedUidArr = new int[requiredLength];
intrandomUid ;
int count =0;
while(true)
          {
randomUid = rand.nextInt(uid_arr.length);
if(relevent_fitness[randomUid]>releventThreshold )
                {
selectedUidArr[count] = uid_arr[randomUid];
                    ++count;
                }
if(count==requiredLength)
                {
break;
                }
          }
          //picked required
loging.append("Selected array\n");
for(int i=0 ; i<requiredLength;i++)
          {
System.out.println("Seleteduids : "+selectedUidArr[i]);
loging.append(selectedUidArr[i]+ "\t");
          }
```

Code for Cross Over :

```
intcrossArrLength = requiredLength/2;
int crossArr1 [];
int crossArr2 [];

          crossArr1 = new int[requiredLength];
          crossArr2 = new int[requiredLength];
int cross2index=0;
int cross1index=0;
for(int i=0;i<requiredLength;i++)
          {
if(i<requiredLength/4)
                    {
crossArr1[cross1index] = selectedUidArr[i];
cross1index++;
                    }
else if((i>requiredLength/4) && (i<requiredLength/2))
                    {
crossArr2[cross2index] = selectedUidArr[i];
cross2index++;
                    }
else if((i>=requiredLength/2) && (i<(requiredLength-
(requiredLength/4))))
                    {
crossArr1[cross1index]=selectedUidArr[i];
cross1index++;
                    }
else
                    {
crossArr2[cross2index] = selectedUidArr[i];
cross2index++;
                    }
          }
```

Code for Mutation:

```
public String flipBitMutation(String binary)
     {
          String flipBitMutationValue = "";

for(int i=0; i<binary.length();i++)
          {
chareachBit = binary.charAt(i);
if(eachBit=='0')
flipBitMutationValue = flipBitMutationValue+"1";
else
flipBitMutationValue = flipBitMutationValue+"0";
          }

returnflipBitMutationValue;
     }
```

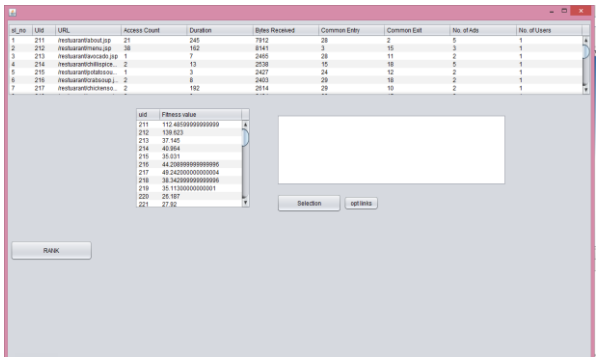

fig:1 Fitness

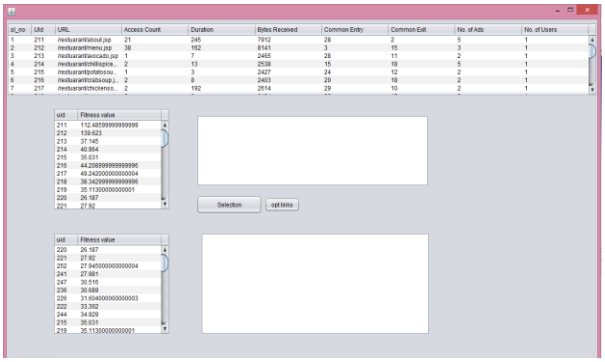

fig:2 Rank

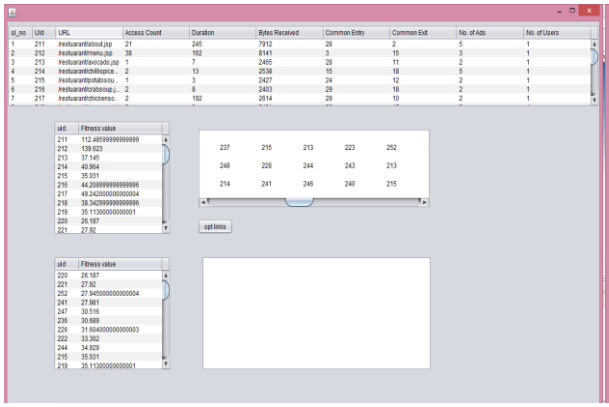

fig:3 Selection

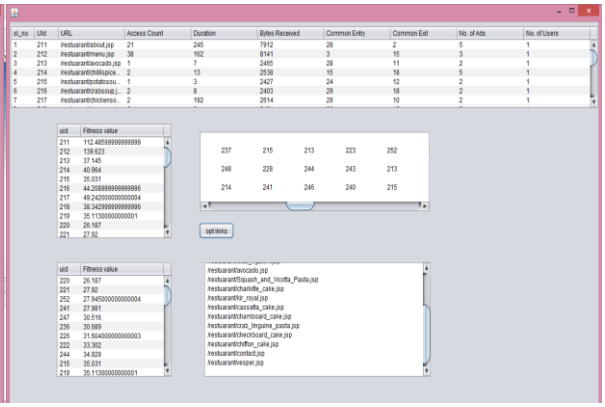

fig:4 Optlinks

V. Advantages Of GA Over SOM

The Genetic Algorithm is a natural optimization and adaptive heuristic search technique whose basic idea depends upon process of natural evolution. The mechanism of evolution is parallel in nature and has been used for solving several computational problems. GA is used for solving general purpose optimizations. Fitness function is an objective function is used for quantifying the optimality of a solution. It measures the goodness of a solution by providing ranks to solution. Genetic Algorithm can be used to optimize any type of fitness function. It does not require the fitness fn. to be continuous or differentiable. Many fitness functions in information retrieval are discrete in nature and GA is well suited for such a task. Secondly, GA has been shown to be very useful for nonlinear function discovery. Because of these reasons Genetic Algorithm is more advantageous than SOM.

VI. conclusion

This paper proposes an efficent model for web mining based on web usage. The proposed model uses GA for finding the web usage. Therefore the inherant advantages of GA will also be the advantages of this proposed work.s

References

[1] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," in Proc. 6th Int. WWW Conf., 1997, pp. 391–404

[2] S. K. Pal, A. Ghosh, and M. K. Kundu, Eds., Soft Computing for Image Processing. Heidelberg, Germany: Physica-Verlag, 2000

[3] Amit Kumar Mishra, Mahendra Kumar Mishra, Vivek Chathurvedi, Santhosh Kumar Guptha, Jaiveer Singh. "Web Usage Mining Using Self Organized Map."International Journal of Advanced Rearch in Computer Science and Software Engineering. Vol.3, June 2013

[4] W. Fan, M. Gordon and P. Pathak, "Genetic programming-based discovery of ranking functions for effective web search," Journal of Management Information Systems, vol 21(4), pp 37-56, 2005.

[5] W. Y. Lin, S. A. Alvarez, and C. Ruiz. Collaborative recommendation via adaptive association rule mining. presented at Int. Workshop Web Mining for E-Commerce (WEBKDD'00). [Online]http://robotiocs.stanford.edu/~ronnyk/WEBDD2000/papers/alvarez.pdf

[6] J. Pazzani and D. Billsus, "Learning collaborative information filters," presented at the Proc. 15th Int. Conf. Machine Learning, Madison, WI, 1998, pp. 46–54.

[7] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Y. Sun, and J. Wiltshire, "Discovery of aggregate usage profiles for web personalization," presented at the Proc. KDD-2000 Workshop Web Mining E-Commerce, Boston, MA, Aug. 2000.

[8] B. Mobasher, N. Jain, E.-H. Han, and J. Srivastava, "Web Mining: Patterns from WWW Transactions," Dept. Comput. Sci., Univ. Minnesota, Tech. Rep. TR96-050, Mar. 1997 [1) F. Picarougne, N. Monmarche, A. Oliver and G. Venturini, "GeniMiner: Web Mining with a Genetic-Based Algorithm," ICWI, pp. 263-270, 2002.

[9] D. Zhang, and Y. Dong, " A novel web usage mining approach for search engines," Computer Networks, vol 39(3) ,pp 303-310, 2002.

[10] Charu C Aggarwal and Philip S Yu. "On disk caching of web objects in proxy servers". In *CIKM 97*, pages 238-245, Las Vegas, Nevada, 1997.

[11] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules". In *Proc. of the 20th VLDB Conference*,pages 487-499, Santiago, Chile, 1994.

[12] Virgilio Almeida, Azer Bestavros, Mark Crovella, and Adriana de Oliveira. Characterizing reference locality in the www. Technical Report TR-96-11, Boston University,1996.

[13] Martin F Arlitt and Carey L Williamson. "Internet web servers: Workload characterization and performance implications". *1EEE/A CM Transactions on Networking*,5(5):631-645, 1997.

[14] M. Balabanovie and Y. Shoham. "Learning information retrieval agents: Experiments with automated web browsing". In *On-line Working Notes of the AAAI Spring Symposium Series on Information Gathering from Distributed, Heterogeneous Environments*, 1995.

[15] Alex Buchner and Maurice D Mulvenna. "Discovering internet marketing intelligence through online analytical web usage mining". *SIGMOD Record*, 27(4):54-61,1998.

[16] L. Catledge and J. Pitkow. "Characterizing browsing behaviors on the world wide web". *Computer Networks and ISDN Systems*, 27(6), 1995.

[17] M.S. Chen, J. Hart, and P.S. Yu. "Data mining: An overview from a database perspective". *IEEE Transactions on Knowledge and Data Engineering*, 8(6):866-883, 1996.

[18] M.S. Chen, J.S. Park, and P.S. Yu. "Data mining for path traversal patterns in a web environment". In *16th International Conference on Distributed Computing Systems*, pages 385-392, 1996.

[19] Roger Clarke. "Internet privacy concerns conf the case for intervention". 42(2):60-67, 1999.

[20] E. Cohen, B. Krishnamurthy, and J. Rexford. Improving end-to-end performance of the web using server volumes and proxy filters. In *Proe. ACM SIGCOMM*, pages 241-253, 1998.

[21] Dr. G. K. Gupta, "Introduction to Data Mining with Case Studies", PHI Publication, 2005.

[22] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Vol. 1, No. 2, Pp. 12-23, 2000.

[23] Adel T. Rahmani and B. Hoda Helmi, "EIN-WUM an AIS based Algorithm for Web Usage Mining", Proceedings of GECCO'08, Atlanta, Georgia, USA, ACM978-1-60558-130-9/08/07, Pp. 291-292, 2008.

[24] Shailey Minocha, Nicola Millard, Lisa Dawson, "Integrating Customer Relationship Management Strategies in (B2C) Ecommerce Environments", IFIP Conference on Human-Computer Interaction- INTERACT, 2003.

[25] C. Ramya, G. Kavitha, K. S. Shreedhara, "Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process", Computing Research Repository - CORR, vol.abs/1105.0, 2011.

[26] V. Chitraa, Antony Selvdoss Davamani, "A Survey on Preprocessing Methods for Web Usage Data", Computing Research Repository-CORR, Vol. abs/1004.1, 2010.

[27] Nizar R. Mabroukeh, Christie I. Ezeife, "A taxonomy of sequential pattern mining algorithms", ACM Computing Surveys - CSUR, Vol. 43, No. 1, Pp. 1-41, 2010.

[28] Francesco Moscato, Nicola Mazzocca, Valeria Vittorini,Giusy Di Lorenzo, Paola Mosca, Massimo Magaldi,"Workflow Pattern Analysis in Web Services", High Performance Computing and Communications - HPCC, Pp.395-400, 2005.
[29] Hussain, T.; Asghar, S.; Masood, N.; "Web usage mining: A survey on preprocessing of web log file", International Conference on Information and Emerging Technologies (ICIET), Pp. 1 – 6, 2010.
[30] Tanasa, D.; Trousse, B.; "Advanced data preprocessing for intersites Web usage mining", IEEE Intelligent Systems, Vol.19, No. 2, Pp. 59 – 65, 2004.
[31] Othman, Z.A.; Abu Bakar, A.; Hamdan, A.R.; Omar, K.;Shuib, N.L.M.; "Agent based preprocessing", International Conference on Intelligent and Advanced Systems (ICIAS), Pp. 219 – 223, 2007.
[32] Khasawneh, N.; Chien-Chung Chan; "Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining", IEEE/WIC/ACM International Conference on Web Intelligence, Pp. 325 – 328, 2006.
[33] Tanasa, D.; Trousse, B.; "Data preprocessing for WUM", IEEE Potentials, Vol. 23, No. 3, Pp. 22 – 25, 2004.
[34] Xidong Wang; Yiming Ouyang; Xuegang Hu; Yan Zhang; "Discovery of user frequent access patterns on Web usage mining", The Proceedings 8th International Conference on Computer Supported Cooperative Work in Design, Vol. 1, Pp. 765 – 769, 2004.
[35] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused crawling: A new approach to topic-specific web resource discovery," presented at the 8th World Wide Web Conf., Toronto, ON, Canada, May 1999.
[36] C. V. Negotia, "On the notion of relevance in information retrieval," Kybernetes, vol. 2, no. 3, pp. 161–165, 1973.
[37] O. Etzioni and O. Zamir, "Web document clustering: A feasibility demonstration," in Proc. 21st Annu. Int. ACM SIGIR Conf., 1998, pp. 46–54.
[38] L. A. Zadeh, "Fuzzy logic, neural networks, and soft computing," Commun. AGM, vol. 37, pp. 77–84, 1994.
[39] P. Britos, D. Martinelli, H.. Merlino, R. García-Martínez, Web Usage Mining Using Self Organized Map, PhD Computer Science Program, National University of La Plata. Software & Knowledge Engineering Center, Buenos Aires Institute of Technology, Intelligent Systems Laboratory, University of Buenos Aires , Argentina, 2007
[40] Qianhui Althea LIANG; Jen-Yao CHUNG; Steven MILLER; Yang OUYANG; "Service Pattern Discovery of Web Service Mining in Web Service Registry-Repository", IEEE International Conference on e-Business Engineering (ICEBE '06), Pp. 286 – 293, 2006.