International Journal of Soft Computing 10 (1): 19-24, 2015
ISSN: 1816-9503

Cluster Optimization for Improved Web Usage Mining Using WebBEE

Anna Alphy and S. Prabakaran
Department of Computer Science and Engineering, SRM University, Chennai, India

Abstract: Rapid development of computer technology allows as accessing huge amount online information's. Next e-Business requirement will be personalizing or customizing the web pages according to the requirement of individuals. Personalization involves learning user's navigational behavior. Web personalization uses web usage mining techniques to customize the web pages. The web usage mining uses data mining techniques to discover interesting usage patterns from web data. The web pages having similar usage pattern are clustered. As users increases or growth in interest of users the size of the cluster increases and it will become inevitable need to optimize clusters. This study proposes a cluster optimizing methodology based on honey bees foraging behavior and is used for eliminating the data redundancies that may occur after the clustering done by web usage mining methods. Genetic clustering is used for the process of clustering. "WebBEE approach for cluster optimization" is presented to personalize web pages for target users.

Key words: Chromosomes, cluster optimization, swarm intelligence, user profiles, web usage mining

INTRODUCTION

The e-Business challenges the design and development of online information management because of rapid growth of the use of worldwide web. This requires understanding and identifying the navigational behaviors of each user. A user who goes online would like to get the web page links which suits his requirements. The next business requirement in the online industry will be providing the users with what they want without asking for it. Here comes the importance of web personalization/customization by displaying the web pages links according to requirements of the user based on user's previous navigational behavior without asking for it.

To improve online business customers need and interests should be understood. For this sufficient data regarding customer's needs and demands have to be collected. Analyze and learn these data for the creation of user profiles. A user profile reveals interests, avocations, needs, etc. The competition in e-Business is rapidly growing and service from the opponent is just a click away for the users' that it's important to continuously monitor users changing interests and needs.

Web usage mining uses data mining techniques to discover interesting usage patterns from web log data (Srivastava *et al*., 2000). Web customization uses web usage mining techniques for personalizing a web page for a specific user. For this, the web log files are analyzed and user sessions are identified. A user session is a sequence of web access by a user. These user sessions are used to create user profiles that portray each user. This customization improves customer relationship management by providing an easy way to access information they are interested in every time they log in.

Currently, for web personalization different clustering methods are available. But data redundancies and performance issues are high in these methods. In this study, a cluster optimization methodology is proposed for eliminating data redundancies that may occur after clustering done by the above mentioned web personalizing methods. For clustering a simple genetic clustering approach is used. In this study, information is extracted from server log files which are used for creating chromosomes and genetic clustering is performed. After applying Genetic algorithm different clusters are formed. These clusters are given as input to the bee clustering which performs optimization.

LITERATURE REVIEW

Web usage mining is the application of data mining techniques to discover usage patterns from web data. Web usage mining consists of three steps pre-processing, pattern discovery and pattern analysis (Cooley *et al*., 1997; Nasraoui *et al*., 1999, 2000; Srivastava *et al*., 2000). Web Utilization Miner WUM uses a mining language MINT that dynamically specifies the navigation pattern of users (Spiliopoulou and Faulstich, 1998). WUM utilizes aggregated materialized view of the

Corresponding Author: Anna Alphy, Department of Computer Science and Engineering, SRM University, Chennai, India

web server log. Presents SEWeP that uses semantics of a website and C-logs for personalizing a website (Eirinaki *et al*., 2003). C-logs are enhanced web logs that encapsulate the knowledge obtained from the link semantics. Here, for web usage mining process C-logs are used as input resulting in a wider semantically concentrated set of recommendations. User logs are examined and users that access similar pages are clustered together and links are dynamically suggested depending upon the categories an individual user falls into (Yan *et al*., 1996). An adaptive website is presented that can highlight interesting links, connect related pages and cluster similar documents thus improving themselves from user access patterns (Perkowitz and Etzioni, 1997). The study presents a complete framework and findings in mining web usage patterns from web log files (Nasraoui *et al*., 2008). In this web logs are pre-processed. Then, they are clustered using hierarchical unsupervised niche clustering. User profiles created are enriched with additional facets. And the current profiles are tracked against existing profiles. The keywords that appear in web pages are used to generate document vectors which are later clustered in the document space to further augment user profiles (Mobasher *et al*., 2000). A two level prediction model based on hierarchical characteristics of web sites (Lee *et al*., 2011). In level one, the category of web page is predicted by Markov Model and in level two, the desired web page is predicted by Bayesian model. This study propose a web content recommender system that uses semantically enhanced server log as input for analyzing the user behavior during web surfing (Fong *et al*., 2011). An Ontology based Rules Retrieval and Rummaging (O3R) for pattern analysis (Becker and Vanzin, 2010). The domain events are classified as physical level represented by URLs and conceptual level represented by ontology. The O3R prototype architecture consists of transforming the physical patterns into conceptual patterns. The patterns are filtered according to conceptual, structural and statistical properties. Then, the patterns are clustered and pattern rummaging is performed that allows interpreting the current patterns using different dimensions of interest. Proposes a web based recommendation model for predicting the user future movements (Jalali *et al*., 2010). The Longest Common Subsequence algorithm is used to predict future movement of user from the current activities. Proposes a clustering technique that group users having same visiting behavior at the same time's two time aware clustering approaches is used for tuning and binding the page and time visiting criteria (Petridou *et al*., 2008).

PATTERN RECOGNITION BASED ON WEB USAGE MINING

Web usage mining handles the discovery of interesting usage patterns using data mining techniques. This includes the following steps:

- Pre-process web log file to extract user sessions
- Pattern discovery based on data mining methods (Cooley *et al*., 1997; Nasraoui *et al*., 1999, 2000, 2008; Srivastava *et al*., 2000)

This study uses genetic clustering technique to cluster user sessions. The clusters obtained are optimized using the WebBEE:

- Generate user profiles from clusters
- Track evolving user profiles (Nasraoui *et al*., 2008)

Figure 1 describes the steps involved in web usage mining.

Preprocessing the web log file to extract user sessions: Quality of the data determines the effectiveness of data mining process. To get better results quality data is needed. Each entry in the web log file consists of the access time, IP address, URL viewed. Pre-processing of web data means a filtering crawlers requests, requests to graphics and identifying unique sessions (Nasraoui *et al*., 2008). After data cleaning chromosomes are created. A chromosome is a binary bit pattern that indicates the behavior of user. The binary bit pattern is created in such a way that if the user i access a url_j is represented as 1 otherwise by 0. Thus, chromosome gives information about the user searching behavior.

Genetic clustering: In this study, information is extracted from the server log files which are used for creating chromosomes. A typical genetic algorithm requires a genetic representation of the solution domain and fitness function to evaluate the solution domain. The Genetic Clustering algorithm starts with a population of chromosomes. Each chromosome is evaluated against a fitness function. New population is created by repeatedly selecting two parent chromosomes and performing crossover and mutation until the new population is complete. This is performed in a hope that better parents will produce better offspring.

Here, first cluster representatives by randomly selecting a user and assign the rest of the users to the closest cluster based on the similarity measure. For measuring similarity dice coefficient is used. General, parameters for binary data where determining the similarity between columns (variables in rows):

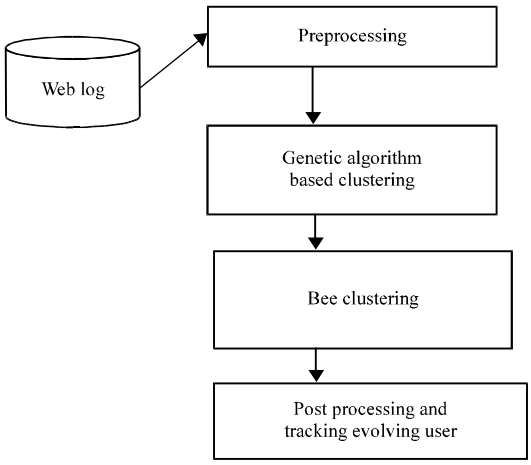

Fig. 1: Steps involved in web usage mining

$$\text{Dice (Sorenson's) Coefficient} = SD = \frac{2C}{N1+N2}$$

Where:

C = Number of positive matches between columns
A = Number of negative matches between columns
T = Total number of variables (rows)
N1 = Total number of presences in column 1
N2 = Total number of presences in column 2

Pseudo code

Genetic clustering:

1. Input: set of users say U1, U2, ..., Up where p is No. of users.
   Output: set of cluster say C1, C2, ..., Ck
   Initialize done [1, 2, 3, ..., p] as 0.
2. Each user has a chromosome based on their searching pattern.
3. Initialize i as 1.
4. Repeat the steps 5-13 p-1 times
5. If done[i] = 1 then goto step 14 else goto step 6
6. Set done[i] as 1 and move Ui to cluster Ck.
7. Initialize j as 1.
8. Repeat the steps 9-11 until j is equal to p.
9. If (i ! = j) and (done[j]! = 1) then
9.1. Calculate similarity between Ui and Uj by using
Dice (Sorenson's) Coefficient = SD = 2C/N1+N2 and stored in S1
9.2. If S1>threshold value then add the user Uj to cluster Ck and set done[j] as 1.
9.3: Perform crossover and mutation and new chromosome to the cluster Ck
10. Increment j by 1.
11. Goto step 8
12. Increment i by 1 and k by 1.
13. Goto step 4

After applying Genetic Clustering algorithm clusters that represent common user profiles are formed.

WebBEE: In recent years, swarm intelligence has gained attention of research scholars. Without the foresight about the distribution of the environment bees, ant, birds, fish can communicate, self organize and cooperate. These swarm intelligence concepts can be used for cluster optimization.

After applying Genetic algorithm different clusters are formed. In WebBee fore-aging behavior of honey bees is used cluster optimization. Cluster representative will acts as employed bee. Employed bee in the first cluster will go to second cluster and find the similarity between itself and members of the second cluster. If the similarity is greater than the similarity between member and its representative of the second cluster then add the new bee to first cluster.

WebBEE:

1. Input: cluster of user formed by Genetic algorithm output:set of cluster say C1, C2, ..., Cn.
2. Let there are n clusters of users say C1, C2, ..., Cn where each cluster have an representative say E1, E2, E3, ..., En, respectively. Let members in a cluster Ci is represented as Ui1, Ui2, Ui3, ..., Uim where m represents No. of members in a cluster Ci.
3. Initialize i as 1.
4. Repeat the steps 5-18 n-1 times
5. Initialize j as i+1.
6. Repeat the steps 7-15 until j is equal to n.
7. Initialize k as no. of members in cluster Cj.
8. Repeat the steps 9-13 until k is equal to 0.
9. Ei of Ci cluster is compared with user Ujk of cluster Cj. 10. Find similarity between Ei and Uik by using
Dice (Sorenson's) Coefficient = SD =2C/N1+N2 and stored in S1
11. Also, calculate similarity between Ej and Ujk and stored in S2
12. If S1>S2 then move the user Ujk to cluster Ci.
13. Decrement k by 1.
14. Goto step 8
15. Increment j by 1.
16. Goto step 6
17. Increment i by 1.
18. Goto step 4

General parameter:

k = Number of cluster
p = Number of users
C = Number of positive matches between columns
A = Number of negative matches between columns
T = Total number of variables (rows)
N1 = Total number of presences in column 1
N2 = Total number of presences in column 2

Once all the iterations are completed researchers get N optimized clusters.

IMPLEMENTATION

The first step is to implement the preprocessing phase. In this all the irrelevant entries are deleted. Each web long entry contains the information's about the user's access pattern. All log entries with filename suffixes such as gif, jpeg, GIF, JPEG, jpg, JPG (Cooley *et al*., 1997) are removed. Now the data used for further analysis is cleaned. Next is to create chromosomes from the preprocessed data and genetic clustering is performed. Figure 2 depicts the creation of chromosomes. Chromosomes which are binary bit patterns represent a users browsing behavior. Users with similar browsing pattern will come under same cluster. These clusters are

userid	programming_la...	college	companies	food	entertainment
albert	1	0	1	0	1
anu	0	1	0	1	0
balu	1	1	1	1	0
christy	0	0	0	0	1
emilin	0	0	0	1	0
felix	1	1	0	1	0
geo	0	1	1	1	1
hari	1	0	1	0	1
irine	1	1	1	0	0
jacob	1	0	0	1	1
jinu	1	0	1	1	1
kiran	0	0	0	1	0
lakshmi	1	1	1	0	0
manju	1	0	1	0	1

Fig. 2: Chromosome generation

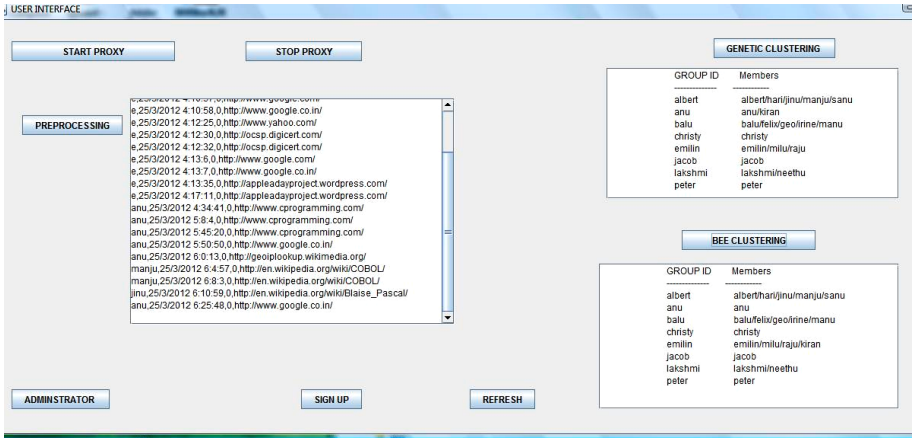

Fig. 3: Results obtained after applying genetic clustering and WebBEE

given as input to Bee Clustering algorithm. Here, positive matches between the clusters are identified. Hence, researchers have clusters with higher densities. Figure 3 shows the user profiles after applying genetic clustering and WebBEE.

PERFORMANCE ANALYSIS

A systematic approach to profile evolution validation (Karaboga and Ozturk, 2011) is performed in terms of precision and coverage.

Precision: A summary profile's items are all correct or included in the original input data that is they include only the true data items. And Precij = sj∩n pi/pi (Nasraoui *et al*., 2008).

Coverage/recall: A summary profile's items are complete compared to the data that is summarized that is they include all the data items. And Covij = sj∩n pi/sj where sj is a summary of input sessions and pi represents discovered mass profile (Nasraoui *et al*., 2008).

Precision will gratify the profiles with true and accurate representations of user input. But, the coverage will gratify the complete and largest possible profiles. Efficiency recommender system depends on well the precision and coverage can be balanced. In Fig. 4, a graph is plotted with user profile id on x-axis and profile evolution validation on y-axis. Quality measure is analyzed in terms of coverage and precision after applying genetic clustering. As in implementation shown in Fig. 3 some groups have more densities and for some groups some elements will be vaporized depending on similarity

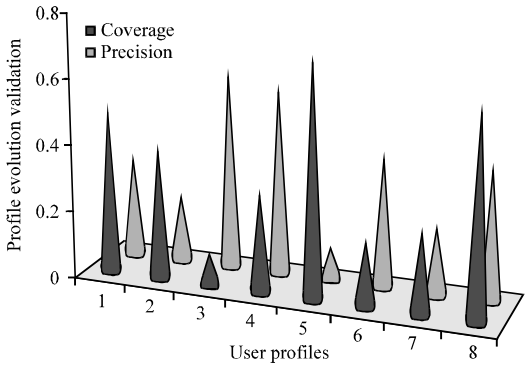

Fig. 4: Quality measure after applying genetic clustering

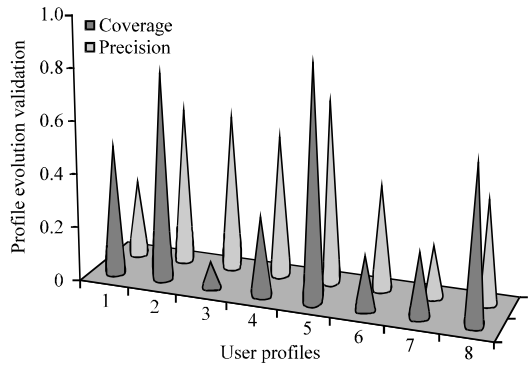

Fig. 5: Quality measure after applying WebBEE

measure. In Fig. 5 also a graph is plotted with user profile id on x-axis and profile evolution validation on y-axis. Quality measure is analyzed in terms of coverage and precision after applying WebBEE. The results shows that after applying WebBEE there is an increase in coverage and precision and also the gap between the coverage and precision is reduced for some user profiles.

CONCLUSION

This study deals with a cluster optimization technique. The web logs are cleaned to remove irrelevant items such as request to graphics and crawler's request. Chromosomes are created from cleaned web logs. Chromosomes resemble the users browsing behavior. Here, clustering technique is used for discovering interesting usage patterns. Clustering is done based on user sessions. Here, genetic clustering is used. The user with the same browsing pattern comes under the same cluster. The clusters obtained are again feed into the proposed WebBEE, a Cluster Optimization algorithm that uses foraging behavior of honeybees. Based on the user profiles the web page is personalized. As a future enhancement scalability can be taken into consideration.

REFERENCES

Becker, K. and M. Vanzin, 2010. O3R: Ontology-based mechanism for a human-centered environment targeted at the analysis of navigation patterns. Knowl. Based Syst., 23: 455-470.

Cooley, R., B. Mobasher and J. Srivatsava, 1997. Web mining: Information and pattern discovery on the world wide web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, November 3-8, 1997, Newport Beach, CA., pp: 558-567.

Eirinaki, M., M. Vazirgiannis and I. Varlamis, 2003. SEWeP: Using site semantics and a taxonomy to enhance the Web personalization process. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2003, Washington, DC., USA., pp: 99-108.

Fong, A.C.M., B. Zhou, S.C. Hui, G.Y. Hong and T.A. Do, 2011. Web content recommender system based on consumer behavior modeling. IEEE Trans. Consum. Electron., 57: 962-969.

Jalali, M., N. Mustapha, M.N. Sulaiman and A. Mamat, 2010. WebPUM: A Web-based recommendation system to predict user future movements. Expert Syst. Appl., 37: 6201-6212.

Karaboga, D. and C. Ozturk, 2011. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Applied Soft Comput., 11: 652-657.

Lee, C.H., Y.L. Lo and Y.H. Fu, 2011. A novel prediction model based on hierarchical characteristic of web site. Expert Syst. Appl., 38: 3422-3430.

Mobasher, B., H. Dai, T. Luo, Y. Sun and J. Zhu, 2000. Integrating web usage and content mining for more effective personalization. LNCS, Vol. 1875, Proceedings of the 1st International Conference in Electronic Commerce and Web Technologies, September 4-6, 2000, London, UK., pp: 165-176.

Nasraoui, O., H. Frigui, R. Krishnapuram and A. Joshi, 2000. Extracting web user profiles using relational competitive fuzzy clustering. Int. J. Artificial Intell. Tools, 9: 509-526.

Nasraoui, O., M. Soliman, E. Saka, A. Badia and R. Germain, 2008. A web usage mining framework for mining evolving user profile in dynamic web sites. IEEE Trans. Knowl. Data Eng., 20: 202-215.

Nasraoui, O., R. Krishnapuram and A. Joshi, 1999. Mining web access logs using a relational clustering algorithm based on a robust estimator. Proceedings of the 8th International World Wide Web Conference, May 11-14, 1999, Canada, pp: 40-41.

Perkowitz, M. and O. Etzioni, 1997. Adaptive sites: Automatically learning from user access patterns. Proceedings of the 6th International World Wide Web Conference, April 7-11, 1997, Santa Clara, California.

Petridou, S.G., V.A. Koutsonikola, A.I. Vakali and G.I. Papadimitriou, 2008. Time aware web users clustering. IEEE Trans. Knowl. Data Eng., 20: 653-667.

Spiliopoulou, M. and L.C. Faulstich, 1998. WUM: A web utilization miner. Proceedings of the 1st International Workshop on Web and Databases (WebDB'98), Spain, pp: 241-253.

Srivastava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorat., 1: 1-12.

Yan, T.W., M. Jacobsen, H. Garcia-Molina and U. Dayal, 1996. From user access patterns to dynamic hypertext linking. Comput. Networks ISDN Syst., 28: 1007-1014.