An Analysis and Knowledge Representation System to attain the genuine web user usage behavior

V.V.R. Maheswara Rao et. al / International Journal on Computer Science and Engineering (IJCSE)

Abstract—With the explosive growth of WWW, the web mining techniques are densely concentrated to discover the relevant behaviors of the web user from the web log data. In fact the pattern discovery techniques generate many hundreds, often thousands, of patterns, that are unwanted, unexpected, disputable and unbelievable in nature. The success of representing the real knowledge out of such patterns is highly reliant on the pattern analysis stage in investigating the web user usage behavior. To retain most genuine and interesting patterns it is necessary to filter out unqualified patterns and use more sophisticated visualization techniques to present the knowledge of web user usage effectively.

The authors in the present paper propose an Analysis and Knowledge Representation System (AKRS) that equally concentrates on both knowledge identification and representation. The key measures are combinedly used for the knowledge identification as a three phase filtering system, to determine the interestingness of patterns in the proposed AKRS. Initially, the objective measures applied on the patterns discovered by pattern discovery techniques to filter out the patterns that do not meet statistical strengths with the frame work of interest factor. Later, subjective measures are applied to identify the patterns that are of most genuine interestingness based on web knowledge. Finally, the heuristic measures evaluate the semantics of patterns based on both user specific objectives and utility of mined patterns. The measures of AKRS efficiently determine the correlation among the most interesting patterns. In addition, to meet the challenges in knowledge representation, like identifying relevant information, finding the depth of information and achieving the visualization competency, the proposed AKRS also designates the recent knowledge visualization techniques like multidimensional and specialized hierarchical. The AKRS amplifies the truthfulness and rate of success in representing final knowledge of web user behavior.

Keywords—Pattern Analysis; objective measures; subjective measures; heuristic measures; visualization techniques.

I. INTRODUCTION

The rapid advancement in web technology and declined costs of storage media, has led business to store enormous amounts of information in huge weblogs. Mining useful information and helpful knowledge from these weblogs has evolved as solid base for researchers and creates a scope for further research. Web mining is the application of data mining techniques to identify and represent useful information and patterns from the weblog data. The survey conducted by various authors and their research contributions identified three broad categories of web mining, namely Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM).

Web Content Mining is a mining technique which 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. Web Structure Mining is a mining technique which can extract the knowledge from the World Wide Web 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. Web Usage Mining, also known as web log 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.
AN INTELLIGENT OPTIMAL GENETIC MODEL TO INVESTIGATE THE USER USAGE BEHAVIOUR ON WORLD WIDE WEB

V.V.R. Maheswara Rao¹ and Dr. V. Valli Kumari²

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

²Professor, 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 evolvoes into a fertile area and got the focus by many researches and business analysts. Web mining is a methodology that blends conventional techniques with incremental algorithms. Web mining is the
AN ENHANCED PRE-PROCESSING RESEARCH FRAMEWORK FOR WEB LOG DATA USING A LEARNING ALGORITHM

V.V.R. Maheswara Rao¹ and Dr. V. Valli Kumari²

¹Professor, Department of Computer Applications, Shri Vishnu Engineering College for Women, Bhimavaram, Andhra Pradesh, India, mahesh_vvr@yahoo.com
²Professor, Department of Computer Science & Systems Engineering, College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India, 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. Before analyzing such data using web mining techniques, the web log has to be preprocessed, integrated and transformed. As the World Wide Web is continuously and rapidly growing, it is necessary for the web miners to utilize intelligent tools in order to find, extract, filter and evaluate the desired information. The data pre-processing stage is the most important phase for investigation of the web user usage behaviour. To do this one must extract the only human user accesses from weblog data which is critical and complex. The web log is incremental in nature, thus conventional data pre-processing techniques were proved to be not suitable. Hence an extensive learning algorithm is required in order to get the desired information. This paper introduces an extensive research framework capable of pre-processing web log data completely and efficiently. The learning algorithm of proposed research framework can separates human user and search engine accesses intelligently, with less time. In order to create suitable target data, the further essential tasks of pre-processing Data Cleansing, User Identification, Sessionization and Path Completion are designed collectively. The framework reduces the error rate and improves significant learning performance of the algorithm. The work ensures the goodness of split by using popular measures like Entropy and Gini index. This framework helps to investigate the web user usage behaviour efficiently. The experimental results proving this claim are given in this paper.

Keywords: Web usage mining, intelligent pre-processing system, cleansing, sessionization and path completion.

1. 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 behaviour of web user.
An Intelligent System for Web Usage Data Preprocessing

V.V.R. Maheswara Rao¹, V. Valli Kumari², and K.V.S.V.N. Raju²

1 Professor, Department of Computer Applications,
Shri Vishnu Engineering College for Women, Bhimavaram, W.G. Dt, Andhra Pradesh, India
mahesh_vrr@yahoo.com
2 Professor, Department of Computer Science & Systems Engineering,
College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India
vallikumari@gmail.com, kvsvn.raju@gmail.com

Abstract. Web mining is an application of data mining technologies for huge data repositories. Before applying web mining techniques, the data in the web log has to be pre-processed, integrated and transformed. As the World Wide Web is continuously and rapidly growing, it is necessary for the web miners to utilize intelligent tools in order to find, extract, filter and evaluate the desired information. The data preprocessing stage is the most important phase in the process of web mining and is critical and complex in successful extraction of useful data. The web log is incremental in nature, thus conventional data preprocessing techniques were proved to be not suitable as they assume that the data is static. The web logs are non scalable, impractical and are distributed in nature. Hence we require a comprehensive learning algorithm in order to get the desired information.

This paper introduces an intelligent system, capable of preprocessing web logs efficiently. It can identify human user and web search engine accesses intelligently, in less time. The system discussed reduces the error rate and improves significant learning performance of the learning algorithm. The work ensures the goodness of split by using popular measures like Entropy and Gini index. The experimental results proving this claim are given in this paper.

Keywords: ISWUP, Human user accesses, Search engine accesses, session identification.

1 Introduction

Web mining is an application of data mining technology for huge web data repositories. 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, Web Structure Mining and Web Usage Mining.


© Springer-Verlag Berlin Heidelberg 2010
An Efficient Hybrid Successive Markov Model for Predicting Web User Usage Behavior using Web Usage Mining

V.V.R.Maheswara Rao
Professor, Department of Computer Applications,
Shri Vishnu Engineering College for Women,
Bhimavaram, Andhra Pradesh, India.

Dr. V. Valli Kumari
Professor, Department of CS&SE,
AU College of Engineering,
Visakhapatnam, Andhra Pradesh, India

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. As web log is incremental in nature, it becomes a crucial issue to predict exactly the ways how users browse websites. It is necessary for web miners to use predictive mining techniques to filter the unwanted categories for reducing the operational scope. The first-order Markov model has low accuracy in achieving right predictions, which is why extensions to higher order models are necessary. All higher order Markov model holds the promise of achieving higher prediction accuracies, improved coverage than any single-order Markov model but holds high state space complexity. Hence a Hybrid Markov Model is required to improve the operation performance and prediction accuracy significantly.

The present paper introduces An Efficient Hybrid Successive Markov Prediction Model, HSMP. The HSMP model is initially predicts the possible wanted categories using Relevance factor, which can be used to infer the users' browsing behavior between web categories. Then predict the pages in predicted categories using techniques for intelligently combining different order Markov models so that the resulting model has low state complexity, improved prediction accuracy and retains the coverage of the all higher order Markov model. These techniques eliminates low support states, evaluates the probability distribution and estimates the error associated with each state without affecting the overall accuracy as well as protection of the resulting model. To validate the proposed prediction model, several experiments were conducted and results proven this are claimed in this paper.

Keywords: Web Usage Mining, Prediction Model, Navigation Behavior, Higher order Markov Model, Web log data, Browsing Patterns, Pre-Processing.
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,
²College of Engineering, Andhra University, Visakhapatnam, India,
Email: mahesh_vvr@yahoo.com, 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.

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...
An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User using Web Usage Mining

V.V.R. Maheswara Rao,
Dept. of Computer Applications,
Shri Vishnu Engineering College for Women,
Bhimavaram, AP, India,
mahesh_vvr@yahoo.com
Research scholar from Computer Science Dept., Aeharya Nagarjuna University

Dr. V. Valli Kumari,
Professor, Dept. of CS & SE,
College of Engineering, Andhra University,
Visakhapatnam, AP, India,
vallikumari@gmail.com

Abstract - The Web has become the world's largest knowledge repository. Web usage mining focuses on discovering the potential knowledge from the browsing patterns of users and to find the correlation between the pages. With exponential growth of web log, the conventional data mining techniques were proved to be inefficient. As web log is incremental in nature, it becomes a crucial issue to predict exactly the ways how users browse websites. It is necessary for web miners to use predictive mining techniques to extract the usage patterns and study the visiting characteristics of user. The data on the web log is heterogeneous and non-scalable, hence to reduce the operation scope and increase the accuracy precision significantly an improved hybrid model is required.

This paper introduces an efficient hybrid predictive model, which is a combination of Markov model and Bayesian theorem. This two stage predictive model to enable the web miner to identify and analyze web user navigation patterns. In this model, the Markov model helps to reduce the operations scope by filtering possible categories and Bayesian theorem improves accuracy in predicting the web pages in identified category. To validate the proposed prediction model, several experiments were conducted and results proven this are claimed in this paper.

Key words: Prediction model, Web mining, Pre processing, web log data browsing patterns.

1. INTRODUCTION

The Web has become the world's largest knowledge repository. The popularity of WWW is rapidly developing and is a golden mount with a lot of valuable information. Extracting knowledge from the Web efficiently and effectively is becoming a tedious process. The exponential growth of the Web has greatly increased the amount of usage data in server logs. Web mining can be categorized into three categories as shown in Fig. 1, web content mining, web structure mining and web usage mining. Web content mining focuses on useful knowledge which is extracted from web pages. Web structure mining is used to analyze the links between web pages through the web structure to infer knowledge. Web usage mining is extracting the information from web log file which is accessed by users.

Web Usage Mining Procedure: The general process of web usage mining includes (i) Pre processing: Process of cleaning, Integrating and Transforming of the result of resource collection, (ii) Pattern discovery: Process of uncover general patterns in the pre process data and (iii) Pattern analysis: Process of validating the discovered patterns. The process of web usage mining procedure is as shown in Fig. 2.

Web Log File: Web usage mining is used to find out the interrelated information from web log file which is involved. All of users' browsing behavior is clearly recorded in the web log file with users' name, IP address, date, and request time etc.

Table 1: Common Web Log

The format of web log file can be divided into two types, common log files as shown in Table 1 and extended log. Common log file is constructed by access log and error log, referrer log and agent log are appended to it, and form extended log. After web log file is acquired, the procedure of web usage mining must be executed.
An Effective Intelligent Pre Processing System of Web Log data adaptable to Incremental Mining

V.V.R. Maheswara Rao, Dr. V. Valli Kumari, Dr. K.V.S.V.N Raju,
Dept. of Computer Applications, Professor, Dept. of CS & SE,
Shri Vishnu Engineering College for Women, College of Engineering,
Bhimavaram, AP, India, Visakhapatnam, AP, India,
mahesh_vvr@yahoo.com vallikumari@gmail.com kvsvn.raju@gmail.com

Abstract—Web mining is an application of data mining technologies for huge data repositories. Before applying web mining techniques, the data in the web logs has to be pre-processed, integrated and transformed. As the World Wide Web is continuously and rapidly growing, it is necessary for the web miners to utilize intelligent tools in order to find, extract, filter and evaluate the desired information in the process of web mining and is critical and complex in successful extraction of useful data. The web log is incremental in nature, thus conventional data preprocessing techniques were proved to be not suitable as they assume the data is static. The web logs are non-scalable, impractical and are distributed in nature. Hence, in order to get the desired information a comprehensive learning algorithm is required.

This paper introduces an Intelligent Pre Processing system, capable of preprocessing web logs efficiently. It can identify human and search engine accesses intelligently, in less time. The work ensures the goodness of split by using popular measures like Entropy and Gini index. The system discussed reduces the error rate significantly. To validate the learning performance of proposed system, several experiments were conducted and results proven this are claimed in this paper.

I. INTRODUCTION

Web mining is an application of data mining technology for huge web data repositories. The web mining can be used to discover hidden patterns and relationships with in the web data. The web mining task can be divided into three general categories, known as Web Content Mining, Web Structure Mining and Web Usage Mining.


(i) Resource collection: The conventional data mining techniques assumes that the data is static, and is retrieved from the conventional databases. In web mining techniques the nature of the data is incremental and is rapidly growing. One has to collect the data from web which normally includes web content, web structure and web usage. Web content resource is collected from published data on internet in several forms like unstructured plain text, semi structured HTML pages and structured XML documents. The web structured data can be captured via inter-page-linkage among the pages of a website and site map of a website. The web usage data can be captured from web logs, click streams and database transactions. The above three resources are captured from a website or a group of related websites.

(ii) Information pre processing: In conventional data mining techniques information pre processing includes data cleaning, integration, transformation and reduction. In web mining techniques the information pre processing includes a) Content pre processing, b) Structure pre processing and c) Usage pre processing. Content Preprocessing: Content preprocessing is the process of converting text, image, scripts and other files into the forms that can be used by the usage mining. Structure Preprocessing: The structure of a website is formed by the hyperlinks between page views. The structure preprocessing can be treated similar to the content pre processing. Usage Preprocessing: The inputs of the preprocessing phase may include the web server logs, referral logs, registration files, index server logs, and optional usage statistics from a previous analysis. The outputs are the user session files, transaction files, site topologies and page classifications.

(iii) Pattern discovery: All the data mining techniques can be applied on preprocessed data. Statistical methods are used to mine the relevant knowledge.

(iv) Pattern analysis: The goal of pattern analysis is to eliminate the irrelevant rules and to extract the interesting rules from the output of patterns discovery process. As the World Wide Web is continuously and rapidly growing, it is necessary for users to utilize intelligent
A Novel Lattice Based Research Frame Work for Identifying Web User’s Behavior with Web Usage Mining

V.V.R. Maheswara Rao¹ and V. Valli Kumari²

¹ Department of Computer Applications, Shri Vishnu Engineering College for Women, Bhimavaram, AP, India and
Research scholar from Computer Science Dept., Acharya Nagarjuna University mahesh_vvr@yahoo.com

² Dr. V. Valli Kumari, Professor, Dept. of Computer Science & Systems Engineering, College of Engineering, Andhra University, Visakhapatnam, AP, India vallikumari@gmail.com

Abstract. Web mining is one of the mining technologies, which applies data mining techniques in large amount of web log data. Web navigational mining discovers users’ access patterns from web logs. This information can be used to identify the behavior of the web user. However, the web data will grow rapidly in the short time, and some of the web data may be antiquated. The user behavior may be changed when the new web data is inserted into and the old web data is deleted from web logs. Therefore, the user behavior must be re-discovered from the updated web logs. However, it is very time-consuming to re-find the users’ access patterns. Hence, many researchers pay attention to the incremental mining, which utilizes the previous mining results and finds new patterns just from the inserted or deleted part of the web logs such that the mining time can be reduced.

The present paper proposes an efficient incremental web navigational mining algorithm for discovering web navigational patterns when the user sequences are inserted into and deleted from original database. It avoids re-finding the original web navigational patterns and re-counting the original candidate sequences. It uses lattice structure to keep the previous mining results such that just new candidate sequences need to be computed. Hence, the web navigational patterns can be obtained rapidly when the navigational sequence database is updated. Besides, maximal web navigational patterns can also be obtained easily by traversing the lattice structure. The experimental results show that the present algorithm is more efficient than the other approaches.

Keywords: Incremental Mining, web navigational patterns, lattice storage structure.

1 Introduction

With the trend of the information technology, huge amount of data would be easily produced and collected from WWW. It causes the web log data to grow up at amazing speed. Hence, how to obtain the useful information and knowledge efficiently from
Study of Visitor Behavior by Web Usage Mining

V.V.R. Maheswara Rao1, V. Valli Kumari2, and K.V.S.V.N Raju2

1 Professor, Department of Computer Applications,
Shri Vishnu Engineering College for Women, Bhimavaram, W.G.Dt, 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, kvsvn.raju@gmail.com

Abstract. Web usage mining focuses on discovering the potential knowledge from the browsing patterns of users and to find the correlation between the pages on analysis. With exponential growth of web log, the conventional data mining techniques were proved to be inefficient, as they need to be re-executed every time. As web log is incremental in nature, it is necessary for web miners to use incremental mining techniques to extract the usage patterns and study the visiting characteristics of user. The data on the web log is heterogeneous and non scalable, hence we require an improved algorithm which reduces the computing cost significantly.

This paper discusses an algorithm to suit for continuously growing web log, based on association rule mining with incremental technique. The algorithm is proved to be more efficient as it avoids the generation of candidates, reduces the number of scans and allows interactive mining with different supports. To validate the efficiency of proposed algorithm, several experiments were conducted and results proven this are claimed.

Keywords: Incremental Frequent Pattern Tree, Total Site Reach, Stickiness.

1 Introduction

The World Wide Web is fertile ground for web mining techniques. When users visit the web pages, the web servers accumulate and record the abundant data, which is structurally complex, non scalable and exponentially growing in nature. The web mining includes Web Content Mining, Web Structure Mining and Web Usage Mining. Web Usage Mining is a special kind of web mining, which can discover the knowledge in the hidden browsing patterns and analyses the visiting characteristics of the user.

Web usage mining is a complete process that includes various stages of data mining cycle, including Data Preprocessing, Pattern Discovery & Pattern Analysis. Initially, the web log is preprocessed to clean, integrate and transform into a common log. Later, Data mining techniques are applied to discover the interesting characteristics in the hidden patterns. Pattern Analysis is the final stage of web usage mining which can validate interested patterns from the output of pattern discovery.
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 Raju³,
¹Shri Vishnu Engineering College for Women, Bhimavaram, India,
Email: mahesh_vvr@yahoo.com
²College of Engineering, Andhra University, Visakhapatnam, India,
Email: vallikumari@gmail.com, kvsvnraju@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 consider practical measures like page, link and mean qualities. The genetic operators are intended in drawing the stickiness among the characteristics of a 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.

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 modeling is to generate the semantic enhanced content data. The integration of semantic content with IOG can potentially
A Complete Research Frame Work to Investigate Web User Usage Behavior

Prof. V. V. R. Maheswara Rao and Dr. V. Valli Kumari
1 Shri Vishnu Engineering College for Women, Dept of Computer Applications, Bhimavaram, India
2 Andhra University, Department of CS&SE, Visakhapatnam, India
mahesh_vvr@yahoo.com
vallikumari@gmail.com

Abstract—With the massive influx of information on the World Wide Web, weblog has become potentially high dimensional and diversified, which got the attention of present researchers. Web usage mining focuses on investigating and analyzing the potential knowledge from navigational patterns to serve the needs of the web user. The conventional techniques applied for data mining are not suitable for dynamically changing weblog scenario. Moreover without careful analysis on the discovered sequences, performing repeatedly the mining techniques yields an inevitable result. However, the past research work in web mining focused attention on individual stages of web mining process alone instead of establishing a whole system.

The authors in the present paper propose a complete research framework, Investigation of Web User Usage Behavior System IWUUBS, which equally concentrates on all stages of web mining process and more adaptable to incremental web log scenario. The IWUUBS is an organic system which consists of an Improved pre processing system, Lattice based pattern discovery system and an advanced pattern analysis system. All the parts of IWUUBS have close connection with each other and work as an organic system. The IWUUBS is proved to be more efficient in investigating the web user usage behavior efficiently. IWUUBS is an organic system which consists of Improved Pre processing System IPS, Incremental Lattice Pattern Discovery System ILPDS and Advanced Pattern Analysis System APAS, where all these parts have close connection with each other.

Preprocessing is a first and complex stage of the complete knowledge discovery process that can take majority of the time in the whole process, thus developing improvent methods for efficient pre processing is an essential task in view of improving the performance of mining techniques. To perform this work effectively, the organic part IPS of IWUUBS introduced by the authors in the present paper takes raw weblog data and preprocesses in different stages and delivers suitable data to the mining techniques.

Pattern discovery is second stage of web usage mining process which can take the output generated by pre processing stage. To identify the web user behavior, it is essential to study the relationship among the pages browsed by the web user. In order to study the relationship among the pages one has to mine the web navigational patterns incrementally. To perform incremental mining on web navigational patterns, one can use the previous mining results to discover new patterns such that the mining time can be reduced. To store previous mining results, the organic part ILPDS of IWUUBS introduced by the authors in the present paper, effectively mines the web navigational patterns from the first level to the last level of the lattice structure.

Pattern analysis is the last stage in the overall Web Usage mining. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The discovery of Web usage patterns, carried out by pattern discovery techniques described earlier would not be very useful unless there were mechanisms to develop descriptions about browsing sequences to help an analyst.

Index Terms—web usage mining, usage behavior, analysis system and incremental mining

I. INTRODUCTION

With the trend of the information technology, huge amount of data would be easily produced and collected from WWW. It causes the web log data to grow up at amazing speed. Hence, how to obtain the useful information and knowledge efficiently from web log data is becoming a typical task for web miners. Web Usage Mining focuses on investigating potential knowledge from browsing patterns of the web users and finds the correlation between the pages.

The overall web usage mining process can be divided into mainly three inter-dependent stages: Pre Processing, Pattern Discovery and Pattern Analysis as shown in Fig. 1. Each stage is very crucial and critical, but combine helps in investigation of web user usage behavior. However, the past research frame work aimed at any one of the stages, but not all.
A PLAUSIBLE COMPREHENSIVE WEB INTELLIGENT SYSTEM FOR INVESTIGATION OF WEB USER BEHAVIOUR ADAPTABLE TO INCREMENTAL MINING

V.V.R. Maheswara Rao¹, Dr. V. Valli Kumari² and Dr. K.V.S.V.N. Raju²

¹Professor, Department of Computer Applications, Shri Vishnu Engineering College for Women, Bhimavaram, Andhra Pradesh, India, mahesh_vvr@yahoo.com
²Professor, Department of Computer Science & Systems Engineering, College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India, vallikumari@gmail.com, kvsvnraju@gmail.com

ABSTRACT

With the continued increase in the usage of the World Wide Web (WWW) Web mining has been established as an important area of research. The WWW is a vast repository of unstructured information, in the form of interrelated files, distributed on numerous web servers over wide geographical regions. Web mining deals with the discovering and analyzing of useful information from the WWW. Web usage mining focuses on investigating the potential knowledge from the browsing patterns of users and to find the correlation between the pages on analysis. To proceed towards web intelligence, obviating the need for human interaction, need to incorporate and embed artificial intelligence into web tools. Before applying mining techniques, the data in the web log has to be pre-processed, integrated and transformed. The data pre-processing stage is the most important phase in the process of web mining and is critical and complex in successful extraction of useful data. The web log is non scalable, impractical and distributed in nature thus conventional data pre-processing techniques are proved to be not suitable as they assume that the data is static. Hence intelligent system is required for capable of pre processing weblog efficiently. Due to the incremental nature of the web log, it is necessary for web miners to use incremental mining techniques to extract the usage patterns and study the visiting characteristics of user, hence one can require a comprehensive algorithm which reduces the computing cost significantly.

This paper introduces an Intelligent System IPS for pre-processing of web log, in addition a learning algorithm IFP-tree model is proposed for pattern recognition. The Intelligent Pre-processing System (IPS) can differentiate human user and web search engine accesses intelligently in less time, and discards search engine accesses. The present system reduces the error rate and improves significant learning performance of the algorithm. The Incremental Frequent Pattern Tree (IFP-Tree) is to suit for continuously growing web log, based on association rule mining with incremental technique. IFP-Tree is to store user-specific browsing path information in a condensed way. The algorithm is more efficient as it avoids the generation of candidates, reduces the number of scans and allows interactive mining with different supports. The experimental results that prove this claim are given in the present paper.

KEYWORDS

Web usage mining, intelligent pre-processing system, incremental frequent pattern tree.

1. INTRODUCTION

Web mining deals with the application of data mining techniques to the Web for extracting interesting patterns and discovering knowledge. Web mining, though essentially an integral part of data mining, has emerged as an important and independent research direction due to the
Understanding User Behavior using Web Usage Mining

V.V.R. Maheswara Rao
Shri Vishnu Engg. College for Women
Bhimavaram, W.G.Dt,
Andhra Pradesh, India

Dr. V. Valli Kumari
AU College of Engineering,
Visakhapatnam,
Andhra Pradesh, India

Dr. KVSVN Raju
AU College of Engineering,
Visakhapatnam,
Andhra Pradesh, India

ABSTRACT
Web usage mining is about analyzing the user interactions with a web server based on resources like web logs, click streams and database transactions. It helps in discovering the browsing patterns of the user and in relating the pages visited by him. This knowledge can be of help in making business decisions, refining the web site design to derive personalized pages. Conventional data mining techniques were proposed to be inefficient, as they need to be re-executed every time the log changes which in turn requires multiple database scans.

This paper discusses an algorithm to suit the dynamically changing log scenario, based on association rule mining with incremental technique. The algorithm proved to be more efficient as there are redundant no of scans of database; it avoids the generation of candidate keys and allows interactive mining with different supports. The experimental results proving this claim are given in this paper.

Categories and Subject Descriptors
Data Mining – Web Mining: Incremental approach

General Terms
Web usage mining, Web content mining, Web logs, Click streams, Data preprocessing, Pattern discovery, Pattern analysis, Incremental mining, Web data collection, Algorithm, Session identification, Stickiness, Association rules, Candidate keys, Interactive mining, supports.

Keywords
IFPT, TSR.

1. INTRODUCTION
The World Wide Web is fertile ground for web usage mining techniques. When users visit the web pages they leave abundant information in web server log, which is structurally complex, large in size and ever growing in nature.

According to the web usage mining objectives, there are three domains, known as Web Content Mining, Web Structure Mining and Web Usage Mining.

Web Content Mining is a mining technique which can extract the knowledge from content published on the internet, usually as semi-structured (HTML), Unstructured (Plain text) and structured (XML) Documents. Web Structure Mining is a mining technique which can extract the knowledge from the World Wide Web and links between references in the Web. Web Usage Mining is a special kind of data mining to address the behavior of user in web access logs.

Web usage mining analyzes user interactions with web server, includes web logs, click streams and database transactions. It identifies user browsing patterns and ideal Web site design structure.

Web usage mining can be divided into Data Preprocessing, Pattern Discovery & Pattern Analysis. The inputs for Data Preprocessing are the Web Server Logs, Referral Logs, Registration Files and Index Server Logs. The outputs are the User session file, Transaction file, site Topology and Page classification. In Web Usage Mining, the Server logs are first preprocessed to clean and transform the data. Later Data mining techniques are applied to extract the interesting patterns. Pattern Discovery, in Web domain, the pages, which are most often accessed together, put in single server session by association rule mining. Association rule mining techniques can be used to discover correlation between pages found in a web server log. Pattern Analysis is the final stage of usage mining which can extract interested patterns from the output of pattern discovery. There are two most common approaches for the pattern analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations. All the methods assume the output of the previous phase that has already been structured.

Typical Web server log will grow rapidly in the short time, and some of the web data may be antiquated. The user behavior may be changed when the new web data is inserted into and the old web data is deleted from the web logs. Therefore, one must re-discover the user behaviors from the updated web logs. However, it is very time-consuming to re-find the users' access patterns. Hence, many researchers focused their attention on the incremental mining in recent years. The essence of incremental mining is that it utilizes the previous mining results and finds new