ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

An Efficient Web Mining Algorithm To Mine Web Log Information

R.Shanthi[1], Dr.S.P.Rajagopalan[2]

Research Scholar, Dept. Of Computer Applications, Sathyabama University, Chennai, India[1]

Professor, Dept. Of Computer Applications, Dr.M.G.R.University, Chennai, India[2]

***ABSTRACT*:** This paper focuses on the efficient application of the Web Mining Algorithm for web log analysis which is applied to identify the context associated with the web design of an e commerce web portal that demands security. As priority is given to efficiency, the comparative study made with other similar algorithm like E-web Miner Algorithm and Apriori All, it has been proved that this proposed Web Page Collection web mining algorithm as the best [or say the most suited] performer to manage time and space complexity .thus this algorithm, better known as Efficient web Miner possesses valid by computational comparative performance analysis. The number of data base scanning drastically gets reduced in Web Page Collection algorithm. Here it may be noted that E -Web Miner can be applied successfully in any weblog analysis which includes information centric network design.

Keywords: E-Web Miner IMPROVED Apriori-ALL Algorithm, E Web, weblog.

I. INTRODUCTION

Web mining is the latest variation of data mining concerning web data which are mostly structured are the result of various web activities. Web mining is broadly categorized as Web Content mining, Web Structure Mining and Web Usage Mining and all these categories function using the data of different web contents such as web page content of HTML/XML code for the pages (web structure). This can be linear or hierarchical or any actual linkage structure access information of web pages (web usage) e.g. Number of hits/subscription/visit (Dunham 2003) User profile, cookies etc. , In short this research paper attempts to prove that E Web Miner has lower complexities of time and space and is better than improved Apriori- All Algorithm.

A. WEB LOG ANALYSIS

Web Log mining is the outcome of web usage mining which contains information of web access of different users. Here any kind of access (Hans and Kamber 2001) information's recorded by the web server into log file for corresponding data. Analysis of log files provides the full details of the access patterns of the users, example profiles of the user's behavior, operating system used, particular time period of usage in the way of successful/unsuccessful transactions etc., thus summarizing all these information in a pre decided format. E.g. A log file of Microsoft Internet Information Server 5.0 having format Of W3 C extension norm (Bair, 2003) completing the task of analyzing these log files exposes the two way results depending upon the perspective chosen, viz from the client part of view or the server point of view. If it is, for instance, from the server point of view, the web log analysis reveals the in details about the availability of these servers, vulnerability of servers, security loopholes of servers, user etc. By this analysis, the web designer can fill up the gaps by improving the required services and web site design. Similarly, the clients are also benefited through the information provided by the web log analysis regarding the frequency of the usage of particular web page etc.

B. CONVENTIONAL SEARCH ENGINE

Information on any kind of searches largely depends on the selection of apt and appropriate key words used through such engines linked with the web sites. But here the individual's query fetches only small set of information known as "Over abundance" which refers to the limited number of web pages, are accessed through Search engines within short interval. (E.g. 12 Sec, 17 sec) etc., Problem of limited coverage refers to access of indices, created and by search engines and are updated periodically. These indices are directly accesses for retrieval on the request of the query (Bin. Deond Xiang 2003).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

If the user's background and knowledge are found inadequate, then the problem of limited customization of semantics, or ontology crops up. Thus the other problem of limited query is based only on the Single keyboard without any context or associated semantics: popularity/usage of number of access of a web page is strongly retrieves on the basis of its popularity.

C. SEQUENTIAL PATTERN

It is defined as an ordered set of web pages that satisfy a given support of defined confidentiality. It is the client who creates this support and sequential pattern. Thus a user can invoke many sessions; a sequential pattern over total number of clients may emerge in many sessions which excludes / rules out the contiguously accessed pages. So sequential pattern is said to be having a n number of web pages between its home page (fast click) and the destination home page.

D. SEQUENTIAL PATTERN ALGORITHM FOR WEB LOG ANALYSIS

Traversal system, a web mining technique applied on click stream data (Tough Pilion 2005) reveals a set of web pages visited by a user in a session wherein the design of a web page can be improved from this information. The data provided by contiguous web pages help in creating/forming new links which in turn helps/supports for easy forward traversal to benefit the real time system. A performance analysis, incorporating such links changes the sequence visited by a user with a limited set of meaningful patterns, reduces the route length of pre-fetching and caching. As data generated in web log is asynchronous in nature, the access pattern of a user can never be monitored in time synchronous manner. Web log mining is used in many applications like pre-fetching, server side transformation and customization. It is type of operation where the web sites information is converted into more suitable information for the web users, who visit and discover their idea. In different aspects the research has been made for the issues of creating web interfaces from the web log based on the user behaviours.

II BACKGROUND

Several approaches have been proposed for efficient application of the Web Mining Algorithm for web log analysis. Dynamic techniques avoid many problems faced by static techniques and are subject of recent studies. The application of meta heuristic techniques in the field of mining is proposed.

2.1 APRIORI ALGORITHMS

Apriori is a classic algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. This has applications in domains such as market basket analysis.

2.2 APRIORI-ALL ALGORITHMS

The algorithm happens to be a modification of Apriori Algorithm [Dunham, 2003]. The modification allows to put the data in correct order by using User-ID and time-stamp sort. The major difference between Apriori and Apriori-All is that Apriori-All makes use of full join for candidate sets. In case of Apriori, it is only forth joined. Thus, Apriori-All is more appropriate for web usage mining rather than Apriori. Apriori is found suitable for web log mining. The sorting of candidate sets identifies the sequential patterns that are complete reference sequence for a user across various transactions. It is iterative in the senses that first scan finds large 1-itemset. Initially, a frequent 1-itemset is the same as a frequent 1-sequence. The subsequent scan divulges more candidate sets from this larger item sets of the previous scan and it may be counted for reference. The counting indicates support.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

III PROPOSED STRATEGY

On the decades various web mining algorithms have been developed in order to cater various clients and server side needs. Some of the following subsections present a bird's view as:

A. THE APRIORIALL ALGORITHM

The Algorithm is a modification of Apriori Algorithm (Dunham 2003) which puts the data in order by using I D and time stamp sort. The major difference between A Priori and Apriori All is that Apriori-All uses the optimum use of candidate's sets. Where as in the Apriori, it is been fourthly joined. This difference obviously holds that Apriori-All as more appropriated for the web usage mining while Apriori is identified /treated as more suitable for web log mining. (The sorting of candidate sets identified the complete sequential pattern with the reference /sequence for a user across various transactions. The large 1-intemset resembles frequent-I-sequence. This further reveals /exposes more candidate sets). This is performed in Apriori in the following manner and the Algorithm is

Input: U= { U1,U2,…,Ui } // The set of users

D={t1,t2,…,tk} //Database of sessions with UserID

S //Support

Output: sequential patterns Ck

D`=sort D` on UserID and time of first page reference in

Each session;

L1 with UserID={large 1-itemsets};

For (k=2; Lk-1! =null; k++) do

Begin

Ck=Apriori-gen(Lk-1,U);//new candidate set

For all transaction ti _ D` do

Begin

Ci=subset(Ck, ti);

For all candidate c _ Ci do

c.count++;

End

Lk={c _ Ck, c.count>S};//S:support

End

Find maximal reference sequences from L;

Procedure Apriori-gen(Lk-1,S,U)

Ck=null;

For each itemset Li _ Lk-1

For each itemset Lj _ Lk-1

Begin

If Li and Lj has same U

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

begin

C=Li join Lj;

If has infrequent-subset(c,Lk-1) Delete c; Else Add c to Ck;

End

End

Return Ck;

Procedure has infrequent-subset(c,Lk-1)

For each (k-1) subset s of c

If s Lk-1 then returns False; Else True.

The authors Tong and Pi-lian introduced/took the lead to suggest the AprioriAll Algorithm by reducing the size of the candidate sets, where in the number of scanning data base is reduced in generating the large set proving efficiency. (Check technically) Here user property and candidate set pressing is not considered while Apriori All Algorithm is (Tony and PiLion 2005) rather presented as:

Input: U= {U1, U2… Ui} // the set of users

D= {t1, t2…, tk} //Database of sessions with UserID

S //Support

Output: sequential patterns Ck

D`=sort D` on UserID and time of first page reference in each session;

L1 with UserID= {large 1-itemsets};

For (k=2; Lk-1! =null; k++) do

Begin

Ck=Apriori-gen (Lk-1,U);//new candidate set

For all transaction ti _ D` do

Begin

Ci=subset(Ck, ti);

For all candidate c _ Ci do

c.count++;

End

Lk={c _ Ck, c.count>S};//S:support

End

Find maximal reference sequences from L;

Procedure Apriori-gen(Lk-1,S,U)

Ck=null;

For each itemset Li _ Lk-1

For each itemset Lj _ Lk-1

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

Begin

If Li and Lj has same U

begin

C=Li join Lj;

If has infrequent-subset(c, Lk-1) Delete c; Else Add c to Ck;

End

End

Return Ck;

Procedure has infrequent-subset(c, Lk-1)

For each (k-1) subset s of c

If s Lk-1 then return False;

Else True

B. WEB MINING EFFORT

The Collector Engine System- a collaborative web mining system was introduced / suggested in E Commerce (GUO 2006) by GUO. This is a multi agent system provides post retrieval analysis and user collaborative across web search and web mining. Here the users are benefited, who can activate search session and share the same with other users also. Collector Engine System consists of four types of software agent's viz. User Agent, collaborator agent, Scheduler Agent and Web Agent. Each of these bears various responsibilities and accordingly, as the name suggests.

***User Agent*:** Retrieves pages from the Web, the collaborator agent-facilitates the sharing of information among different user agents, The Scheduler Agent maintains a list of monitoring and carrying out tasks based on users schedules, The Web Agent maintains the exact filtering the trashy information (Pitkow 1997).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

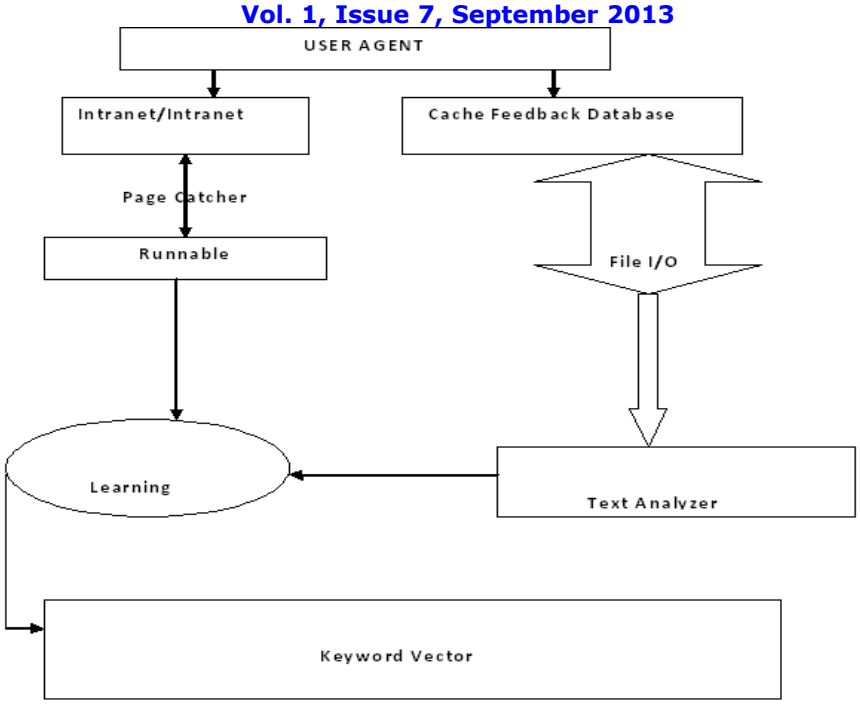

Figure 1: Agent running model of CES

The figure-1 also explains the functions of CES with these agents interacting with the information/fulfilling their tasks accordingly as explained already. I.e. consisting the actual network of collected information with local data structure. The main disadvantage with CES is the system will be of a non automated collection of user profiles. But this can be avoided by including more content based or collaborative information recommendation functionalities.

C. COOKIE PICKER

Cookies are used mainly to track/are exploited to the maximum by websites to track a User profile which violates privacy of the user and security of communicator. This has paved way for the introduction of Cookie Picker, Cookie management Scheme (check technically) through which Usefulness of cookies from a web site is validated. This sets the cookies usage permission on behalf of the users.

D .WEB LOG MINING

GAO (GAO2010) used patterns reorganizations system to study the client behavior from the web data towards providing quality service. Client Behavior Pattern Recognition System is described in figure 2 that reflects the five hierarchical structures in the recognition system:

Application layer: includes a machine interactive interface diagram, used for information exchange between client and system.

***System Control layer*:** Includes a question formation and result using function with the coordination and control of the functional models.

***Data analysis layer*:** it consists of data mining and online analysis process, used for system log data analysis by log mining online analysis technique. The paper (Gao 2010) generates interest at the framework level and various application parameters of the framework prompted to present the following architecture on which algorithm would be applied. The following system (figure 3) based on web log mining architecture for proposing the basic interactive elements for E-Web miner.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

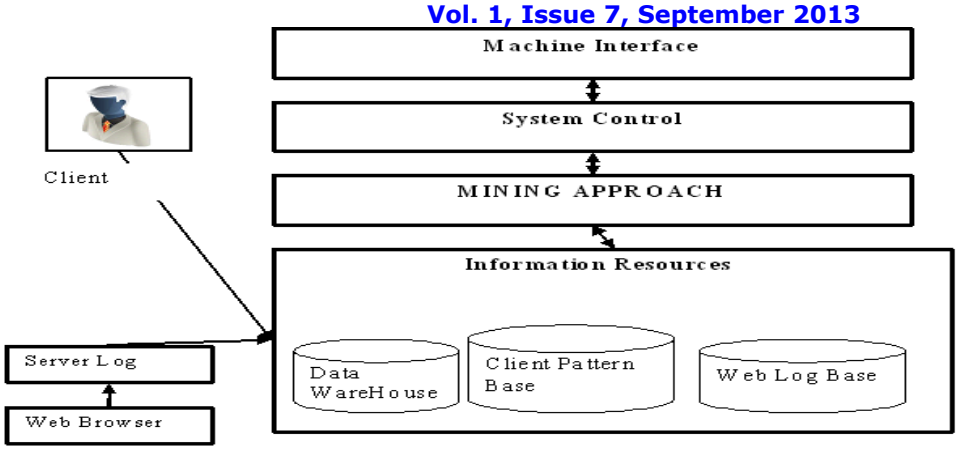

Figure 2: Pattern Recognition System

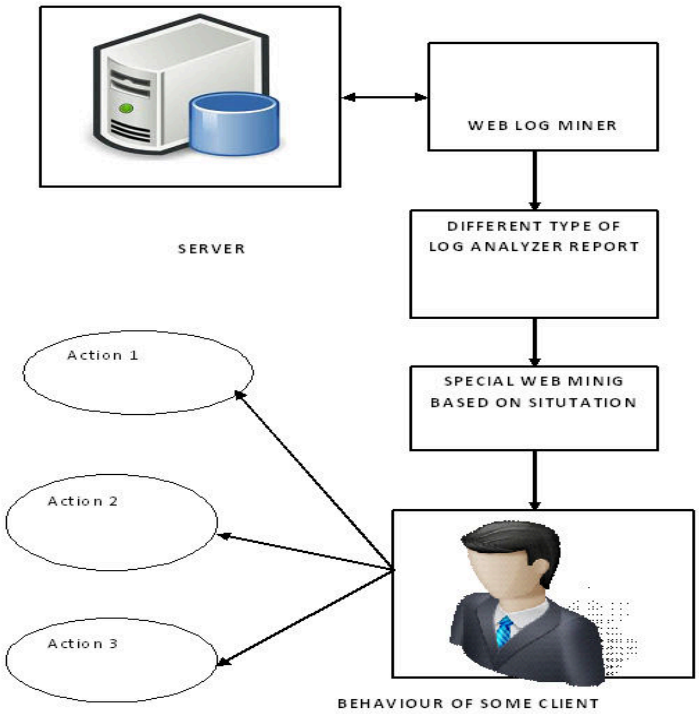

Figure 3: Proposed Web Miner Frameworks

The Rule base of web miner consists of the following factors

-Learning mechanism in data /knowledge base

-Rule base configuration of baseline.

The rule based generations secure web log miner mines data in a secured way by defining a property set and rule based configuration mechanism.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

E. WEB LOG MINING

Of course web log mining is comparatively is paid less attention in the application of E Commerce (Yen men 2010) Web mining technique has been able to generate interest in establishing personalization service system. The relationship of web data mining with E Commerce is well established wherein data security is the main concern in web data (Jan 2007) ensuring quality issues of communication such as security. Thus these Web data are analyzed towards designing a security Protocol. Towards fulfilling this technique ANSIX 959 Standard has been formed (Levi and koe 2001) developed towards preparing convenient and secures electronic payment Protocol.

F. E-WEB MINING ALGORITHM

E-Web Mining is the improvisation of the Web mining algorithm which removes the loopholes in the Apriori-All Algorithm.

E- Web Mining Algorithm

1. Make the set of web pages in the ascending order for the various users

2. Now assign the set of pages in the string array a for the user u

3: Initialize the f=0, max= 0, where f is frequency and max is maximum

4. Consider I vary from 1 to n; also J varies from 0 to (n-1)

5. If substring (a[I], a[J])

f=f+1;

END IF

b [I] =f;

IF max<= f

Max =f;

END IF

6. Discover the positions in array b, where the value is nearly equal to maximum value and chooses the substring from all values.

7. Repeat the step 6 for all the substring with their positions.

IV PROPOSED WEB PAGE COLLECTION ALGORITHM

Consider the given web access log, a major task is to identify the collection of pages which tend co-occurred in the specific visits. Usually the clustering techniques are used for the web log access to analysis the web log effectively. In the clustering the documents are represented in N-dimensional space formed by term factor. A cluster is defined as collections of documents close to each other and relatively distinct from other clusters. In the proposed approach the Web Page Collection Algorithm is introduced that uses cluster mining in order to find a group of connected pages at a web site. The proposed algorithm takes web server access log as input and maps it into form clustering. Afterwards the cluster mining is applied to the output data. Finally the output is obtained by mining the web user's logs.

WEBPAGECOLLECTION ALGORITHM

Input: web access log
Output: Mined web users log

1. Take the web log access of users for the particular domain
2. Select a random user into the unique profile set

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

3. Find the co-occurrence frequencies among web pages and
3. Next cluster the similar results
4. Then rank the clusters that is found
5. Create the web page that contain the related links of the specific domain

V COMPARISION WITH EXISTING WORK

The comparative study of Proposed Web Page Collection Mining algorithm and the E-Web Mining Algorithm are carried out by running over a number of transactions from the item set. The result is tabulated in Table 1

Number of Transactions	Number of Items	E-Web Mining (Time)	Web Page Collection (Time)
8	10	0.235	0.011
9	12	0.543	0.005
10	12	0.176	0.002
11	8	0.235	0.009
12	16	0.781	0.010
13	20	0.111	0.014
14	22	0.432	0.058

Table 1: Simulation Results

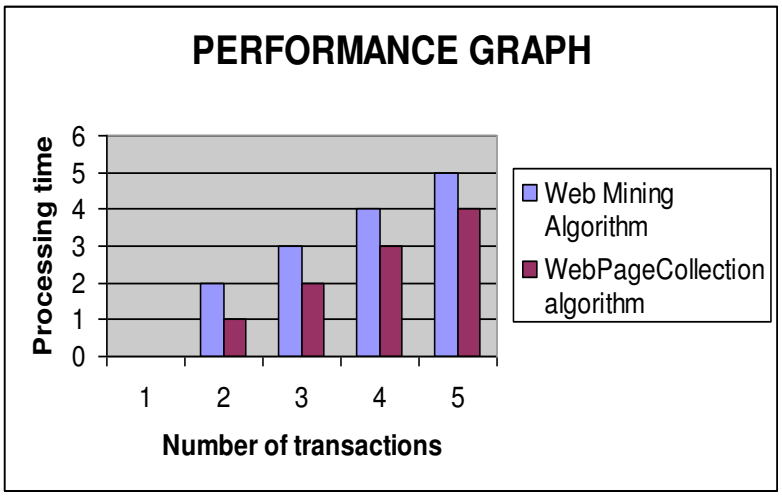

Figure 4: Performance Graph of Proposed Web Mining Algorithm

The figure 4 explains the execution time among the proposed web mining time and Apriori All Time. The main objective of this experiment is to reduce the number of elements in every candidate set without any repetition but with changes in the larger sets. The observation can be studied as follows:

- Firstly candidate set pruning gets reduced in steps.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

- Secondly by applying pruning, the number of elements of candidate set is decreased. Repetitive scanning of data base is eliminated totally. Proving the E web miner as effective.

VI CONCLUSION

This paper presents strategy for automatic web log information mining by Web Page Collection algorithm and is been proved to be more effective. It stands above other web mining algorithms. With the mined results, the web applications is developed and provides adaptive user interface

Further the other type of explore in web applications will focus in the future work. It also includes the information integration of content knowledge and knowledge extraction from the various web sites.

REFERENCES

1] Tong, Wang and Pi-lian, He, Web Log Mining by an Improved AprioriAll Algorithm World Academy of Science, Engineering and Technology, Vol 4 2005 pp 97-100.
[2] Dunham., Margaret H., Data Mining Introductory and Advanced Topics. Beijing: Tsinghua University Press, 2003, p195-220.
[3] Han Jiawei and Kamber Micheline Data Mining Concepts and Techniques [M].Beijing: China Machine Press, 2001, p290-297.
[4] Bain Tony SQL Server 2000 Data Warehouse and Analysis Services. Beijing: China Electric Power Press, 2003, p443-470.
[5] Bin, Lin jie, de, Liu ming and xiang, Chen Data mining and OLAP Theory & Practice [M]. Beijing: Tsinghua University Press, 2003, p194- 244.
[6] Tong, Wang and Pi-lian, He, Web Log Mining by an Improved AprioriAll Algorithm World Academy of Science, Engineering and Technology, Vol 4 2005 pp 97-100.
[7] Wen-Hai Gao Research On Client Behaviour Pattern Recognition System Based On Web Log Mining, Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010, DO1, 978-1-4244-6527-9/10/$26.00 ©2010 IEEE, pp 466-470
[8] Levi Albert, and Koç Çetin Kaya CONSEPP: Convenient and Secure Electronic Payment Protocol Based on X9.59 Proceedings, the 17th Annual Computer Security Applications Conference, pages 286-
295, New Orleans, Louisiana, IEEE Computer Society Press, Los Alamitos, California, December 10-14, 2001
[9] Guo, Di, Collector Engine System: A Web Mining Tool for E-Commerce, Proceedings of the First International Conference on Innovative Computing, Information and Control (ICICIC'06) DOI computer society 0-7695-2616-0/06 2006 IEEE Yuewen, LI, Research on E-Commerce Secure Technology DOI 978-1-4244-3709- 2/10 2010 IEEE
[10] Mei Li and Cheng Feng, Overview of WEB Mining Technology and Its Application in E-commerce, 2010 2nd International Conference on Computer Engineering and Technology ,Volume 7.DOI 978-1-4244- 6349-7/10 .2010 IEEE pp V7-277-V7-280
[11] Yue, Chuan, Xie, Mengjun, and Wang, Haining, Automatic Cookie Usage Setting with Cookie Picker, 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07) DOI 0-7695-2855-4/07 2007 IEEE

BIOGRAPHY

R.Shanthi, receieved B.Sc and M.C.A Degree from University of Madras in 1999 and 2002. She is working as a Assistant Professor in Computer Science Department of A.M.Jain College, Chennai, India. She has 10 years of teaching experience. She is pursuing Ph.D in Sathyabama University, Chennai, India.

Dr.S.P.Rajagopalan received M.Sc from IIT Madras, M.Phil and Ph.D from Madras University, Chennai, India. He is working as a Professor (Emeritus) in School of Computer Science & Engineering of M.G.R.University, Chennai. He has 40 years of teaching experience. He has published 4 Books. He has about 100 publications in International Journals and National Journals. His special fields of interest include Quantitative Techniques, Data Processing and Project Management, Management Information System, Programming Languages, Simulation, Text generation, Cryptography and Data Mining.