Web Usage Data for Web Access Control (WUDWAC)

Dr. Selma Elsheikh*

Abstract—The development and the widespread use of the World Wide Web have made electronic data storage and data distribution possible and convenient. However, this scenario has forced organizations in both private and public sectors to make their data available on the Web with restricted or limited use. These kinds of data include sensitive information that can be released only to specific requesters. The major objective of the proposed system, referred to as WUDWAC is to provide mechanisms for Web access control based on Web usage data from server-side logs. A set of algorithms is used for mining Web usage data which is extracted from Web Server logs and integrated with user login data. The mining process includes preprocessing tasks (for Web access transactions data preparation), association rules mining (to find the correlation among Web access transaction entries) and sequential pattern mining (to find the Web access pattern sequence for the access transaction entries).

***Index Terms*—Access control, Association rules, Sequential pattern, Web Access Control, Web Usage Mining.**

I. Introduction

The World-Wide Web (WWW) is the largest distributed information space and has grown to encompass diverse information resources [3]. In order to share information effectively, adequate security mechanisms are needed for Web access, specifically, user authentication, fine-grained access control and communication encryption. Currently, adequate and affordable security tools required for such use are lacking [2], [3], [11]. The distribution and sharing of information via the Web that must be accessed in a selective way, require enforcement of security controls, ensuring that information will be accessible only to authorized entities. Web access control approaches, where the number of users is predefined are no longer adequate to deal with the dynamic nature of the Web [4], [9], and the existing access control for Web services are lacking support for global services [8],[12].

II. Related Work

In recent years, Web mining has emerged as an important branch of data mining [5]. This is mainly due to the tremendous amount of information available from the Web, which attracted many research communities, and the recent interest of e-commerce [13]. Online mining is the term used when activities of new users are monitored online and reaction to misuse of privileges is generated automatically and immediately [1], [6], [12], [14]. "Reference [10] uses both Web usage and Web content information in their study to find out whether a user is exploiting the current server's data/services to publish similar data/services". "Reference [7] proposes (ActiveWeb) XML- based active rules for deriving Web views and for defining access control by user access behaviors". The access right for a page is given to a user by the user ID and password also by the IP addresses. The limitation of ActiveWeb is in the IP addresses, since it could be continuously changing especially in mobile computing due to the user's frequent geographical data changes.

In this paper, we present an approach for Web access control based on mining Web usage data. The main goal of our work is to develop a Web access control system to provide mechanisms for Web control access based on Web usage data. The Web usage data is collected from the Web browser on Web server logs and user logins. The paper starts with an introduction and a brief description of the related work. Section III presents the WUDWAC Structure while section IV contains the WUDWAC design architecture. The WUDWAC security policy and the Mining process in WUDWAC are illustrated in Section V and VI respectively. Section VII presents experimented applications of the proposed approach followed by the conclusion in the last section.

III. The WUDWAC Structure

Figure 1 shows the WUDWAC Structure. Firstly, the Web browser (user) sends a request to a Web server. The server generates the appropriate code based on the user's request that is usually based on the data that is accessible to the Web application. In order to process the request on the Web server, the user must first authenticate and then request for the desired Web object (Web page). All the access control management resides in the Web server host.

When a user requests an object, the WUDWAC access control system determines which user is authorized to get the object and what associated access rights that user has for the requested object. The access decision is issued through pattern discovery techniques (association rule and sequential pattern). The system generates access rules and patterns on the user access transaction entries data to produce the user access transaction patterns. The analysis of this user access

* Dr. Selma Elsheikh is with the Faculty of Computing, Alghurair University (P.O Box 37374, Dubai, UAE).

patterns is done by matching the active user access transaction pattern with the previous one for the same user to make access decision (access denied or allowed to resources).

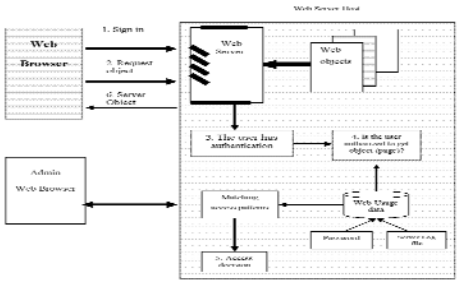


Figure 1: The WUDWAC Structure

IV. THE WUDWAC DESIGN ARCHITECTURE

In the proposed model, the data sources are the server log data and the login data. The system utilizes these data through different stages as in figure 2. The first stage is preprocessing; the main goal of this stage is to create minable objects for knowledge discovery through data transformation and integration. The mining algorithms are identified for pattern discovery in stage two. The SQL engine and the ' if…Then' rules are used as access control enforcement to produce Web access patterns. Stage three includes identification of the user access patterns. The goal of this stage is to eliminate the irrelative rules and to extract the interesting patterns from the output of the previous stage. In the last stage the system checks the user access rights depending on access query patterns and then the access decision is made- either allow access or deny it.

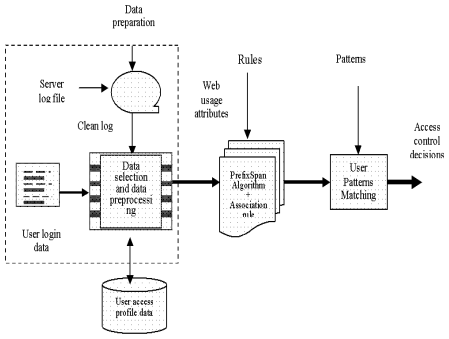


Figure 2: The WUDWAC Design Architecture

V. WUDWAC SECURITY POLICY

The system administrator is responsible to set the system security policies and configures the access permission. The access rights are associated with the objects and the users must ensure that they have sufficient access rights before gaining access to the objects (access the Web). The new user provides his/her login data or the registration data. The system administrator makes sure that the user has used his/her access right and browsed the Web page or pages at least once. This is done after the administrator has given the user authority (user ID and password) to access the specific Web page/ pages.

VI. MINING PROCESS IN WUDWAC

In the proposed model, mining usage data for Web access control is divided into two separate stages. The first stage is the preprocessing and data preparation. This includes data cleaning, filtering, and transaction identification. The second stage is the mining stage where usage patterns are discovered via association-rule discovery and Sequential Pattern Discovery. Each of these components is discussed below.

A. Preprocessing task

The data go through a preprocessing stage to clean the data from irrelevant or redundant data in Web server logs. Next, the data is formatted appropriately according to the association rules and the sequential patterns. The output is server log database file. This database integrated with the user login data to produce the Web access transaction data, which contains password, date of last request (the time duration from when the user Web accessed last, e.g. today, yesterday, the day before yesterday, etc), URL of the page(s) visited, and the Status action. The Web server log file contains a complete history of file access by the users.

Due to local caches, not all page requests made to a server are recorded in the server log files. Since the browser finds in its cache a copy of a document being requested by the user, the request is not made to the server and the stored copy of the document is displayed. Therefore, although the page is viewed by the user, the request is not recorded in the server log file, because it is recorded in the browser's cache. In this proposed model, the browser's cache is disabled; therefore, the user's requests can be recorded. The status action is set by the Web server and indicates the action taken in response to the request. For example Codes from 200 through 299 indicate success, 300 through 399 indicate some form of redirection, 400 through 499 indicate an error serving the particular request and 500 through 599 indicate a problem in the Web server [4].

1) Definition1

Let L= $\{l_1, l_2 \ldots\ldots, l_m\}$ be the set identified Web access transactions entry, where l Є L is defined as a tuple, l_i = $\{password_i , date_i .url_i, statu_i\}$, for $l_i \subseteq L$, where $1< i<m$. Table 1: shows the set of Web access transactions entry (sample).

Table1: Set of Web Access Transactions Entry

page_log

Password	URL	Date	Action-code
mm_125	/ESRG Research /indx/ESRG_project_database.asp	11/27/2007	304
samia338	/ESRG_project_database.asp ESRG Research /indx/ESRG_project_database	11/28/2007	200
Samia_338	/ ESRG Research /index/ ESRG_project_database.asp	11/29/2007	301
samia338	/ ESRG Research /index/ ESRG_project_database.asp	11/30/2007	200
samia338	../ESRG Research/ index/Ecotorism Ranking	11/31/2007	200
samia338	../ESRG Rearchs/index/ technical_project .asp	12/03/2007	200
samia338	/ ESRG Research /index/ ESRG_project_database.asp	12/04/2007	200
mm_125	ESRG Research /indx/ESRG_project_database.asp	12/05/2007	304
samia338	ESRG_project_database.asp ESRG Research /indx/ESRG_project_database	12/05/2007	200
Aaa_2007	/ ESRG Research /index/ ESRG_project_database.asp	12/05/2007	301
samia338	/ ESRG Research /index/ ESRG_project_database.asp	12/05/2007	200
samia338	../ESRG Research/ index/Ecotorism Ranking	12/06/2007	304
Samia_338	../ESRG Rearchs/index/ technical_project .asp	12/06/2007	304
Samia_338	/ ESRG Research /index/ ESRG_project_database.asp	12/13/2007	304
mm_125	ESRG Research /indx/ESRG_project_database.asp	12/13/2007	304
samia338	ESRG_project_database.asp ESRG Research /indx/ESRG_project_database	12/28/2007	200

2) Definition 2

Let U be a set of attributes to identify the Web access transaction entries, U⊆ L, such that U= {password, url, date of last request, status action}. Specifying a value for each attribute in a class set U order set, where the class set U order set forms a web access transactions entry data. In the proposed model the transaction identification is based on the password (class set identifier), and the order time set is based on the transaction date (the date of last request). The Web access transaction entries set < password, date of last request, url, status>.

B. Pattern Discovery

Pattern discovery is the stage of the knowledge extraction. In this stage, association rule discovery techniques are applied to the formatted data of Web access transactions entry. Sequential patterns are used to find the maximal sequences among all sequences that have a certain user specified minimum support and confidence.

1) Association Rules Discovery

In WUDWAC, the association rules capture the relationships among items based on their patterns of co-occurrence across transactions. Association rule is used for inter-transaction mining; each rule has two measures, support (prevalence) and confidence (predictability).

a) Definition 3

Given U= {password, date, url, status} define user access transaction entries pattern such that U ⊆ T. The association rule is an implication of the form X ⇒Y, where X, Y ⊂ U. The rule X ⇒ Y holds in transaction U with confidence c if c% of transaction in that U contain X and also contain Y, and the rule X ⇒ Y has support s in the transaction U if s% of transaction in U contain X∪Y.

In (WUDWAC) the application of association rule mining is to discover the associations and correlations among Web access transaction entries {password, date of the request (Date), page visited (URL), and status action}, where the presence of one set of Web access entries pattern in the transaction implies the presence of others with 100% confidence, and minimum support ≥2.

2) Sequential Discovery Pattern

In (WUDWAC) the Sequential pattern mining is used to find the intra-transaction patterns in Web access Transaction entry. The present work uses the Prefix Sequential pattern mining [6] to discover the single Web access transaction entries patterns in order to predict future ones. The algorithm starts by scanning the input data to find frequent events that can be assembled to the last element or added as such to the end of the prefix to form a sequential pattern. Recursively new prefix projected database is constructed and explored similarly. This is done by selecting the Web access transaction entry patterns as the first projected database <Password, date, url, status action >. Then, select the password as a prefix, get the password form the active user and match the pattern (minimum support ≥ 2). For the second projected database < date, url, status action >, select the date (date of last request) as prefix, and get the date of last request from the user, pattern match with (minimum support=2). For the fourth projected database= < url, status action >, select the url of the page visited as prefix, get the URL of the page requested, the URL pattern match (minimum support ≥ 2). For the fifth projected database <status action> check the status action taken in response to successful request, if the status action = success (the request is fulfilled), then the process ends. Table 2 shows the output patterns from pattern discovery algorithms (ESRG Homepage usage data Sample).

Table 2: The Output Patterns from Pattern Discovery Algorithms (ESRG Homepage usage data Sample)

Projected partition the sequential web access transaction entry patterns	Prefix	Association rules (Pattern growth) minimum support ≥ 2	Confidence (%)
<Login password = samia_338 , date -of - last request = 7/27/2007 , $page_{requested}$ = /userpages/index SRG_project_databas.asp , status action =200 >	< login password= samia-338 >	< login password= samia-338 >	25%
< date -of - last request = 7/29/2007, $page_{requested}$ = /userpages/index/ SRG_project_database.asp , status action =200 >	<date -of - last request = 7/27/2007>	< login password= samia-338 ⇒ date -of - last request = 7/27/2007 >	50%
< $page_{requested}$ = .. /userpages/index/ SRG_project_database.asp, status action =200 >	< $Page_{requested}$ = /userpages/index/ ESRG_project_database.asp > >	Login password = samia_338 ⇒ date -of - last request = 7/27/2007 ⇒ $page_{requested}$ = /userpages/index/ SRG_project_database.asp >	75%
< status action = 200>	<status action = 200 >	< Login password = samia_338 ⇒ date -of - last request = 7/27/2007 ⇒ $page_{requested}$ = /userpages/index/ SRG_project_database.asp ⇒ status action =200 >	100%

C. Pattern Analysis and Access decision

To identify the interesting patterns, Structure Query Language (SQL) was used. The pattern analysis procedure filters the patterns using the SQL algorithm. SQL is used to check which of these patterns match the requirement specified by the user (minimum support ≥ 2) and confidence 100%. The interested Pattern is the pattern with sequential and association of password, date-last-request, page visited (URL), and status action, with minimum support ≥ 2 and confidence 100%.

The WUDWAC makes access decision by analyzing the user access pattern (IF condition THEN action). The Web access transaction entries data are used to match the active user's access pattern. The system makes the access decision based on analyzing the user access pattern. The data of the Web access transaction entries pattern is used to match the active user's access pattern, then makes decision either access is allowed or denied (activate the page or not).

The access rules are described in IF Condition THEN action. The Query below presents a complete representation of the rule.

SELECT Password, date-last-request, $page_{visitited}$, status
If login password = user password,
Then If date-of- last- request = date_of _last- request
Then If $page_{requested}$ = $page_{visitited}$
Then If status action = successful access Then
Activate the page link (page access permitted).
(With minimum support=2 and confidence of the rule is 100%).

VII. Experiment

We have applied the proposed approach to the Expert System Research Group (ESRG) homepage between July 2007 and December 2007. ESRG homepage includes comprehensive expert systems for information capturing, sharing and managing for ecotourism and related industries which consist of different projects information. Each project has a specific research group. Each group has authorization to update or to modify the research information in their page only, while forbidding other groups to do so and vice- versa. The WUDWAC system is developed using Active Server Page script (Server-side Scripting). Figure 3 illustrates the procedure to generate the Web access transactions data. The system administrators have the power to give access authorization to the users, and to make sure that the users use their access rights and activate the links of the pages which they have been given authorization to navigate and view. WUDWAC prototype was tested in this Web page. Testing has been carried out to verify that the system behaves as intended following the designed algorithms and procedures. The outcome of WUDWAC was promising and as expected. . Figure 4 displays the screen of the registration menu, main Web page and subpage in ESRG homepage. Figure 5 shows the password and the date of last request menu, main Web page and subpage in ESRG homepage.

```
bjRecordset1.Open "tracking", Conn, adOpenStatic,
adLockPessimistic, adCmdTable
                objRecordset1.AddNew
        objRecordset1.Fields("page_visited") =
request.servervariables("SCRIPT_NAME")
        objRecordset1.Fields("time") = Time
        objRecordset1.Fields("date") = Date
        objRecordset1.Fields("ACTION_CODE") =
objRecordset1.Update
```

Figure 3: The Procedure Generates the Web Access Transaction

Figure 4: Screen Display the registration Menu, Main Web Page and subpage in ESRG homepage

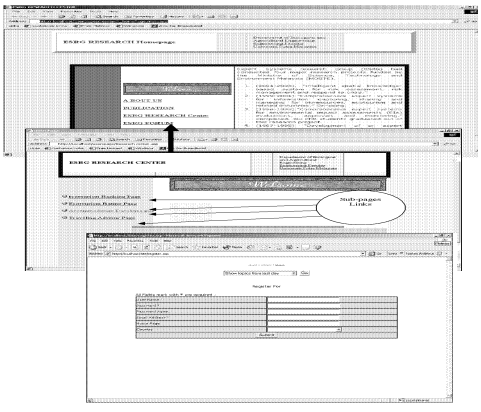
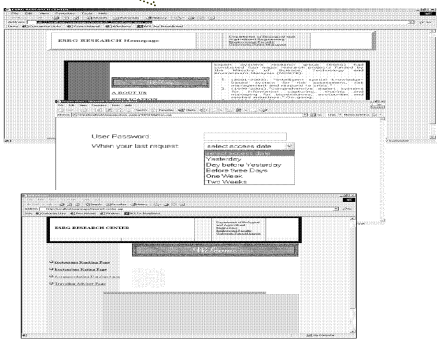


Figure 5: Screen Display the Password and the Date of Last Request Menu, Main Web Page and subpage in ESRG homepage



VIII. Conclusion

The (WUDWAC) model bases on mining Web usage data to explore whether a user is authorized to access specific Web page. The model manipulates the Web usage data to generate Web access transaction data. The data extracted from the server log files are integrated with the user logins and exploited to retrieve Web access transaction patterns of the similar user. This information is then provided to determine whether the user (requestor) is authorized to perform the action or not (has rights to access the Web page or not). The contribution of WUDWAC is its capability to make a decision about which user can access the Web resources and ensuring that access is only allowed to authorize users based on Web usage data or the user navigation history. This can add substantial levels of protection for the Web pages. Future work should be done on the reliability of the Web usage data in Web access control.

References

[1] A. Herzberg, Y. Mass, J. Mihaeli, " Access Control Meets Public Key Infrastructure," Proc. of IEEE Symposium on Security and Privacy, pp 2-14, 2000.

[2] B. Atkinson, G. Della-Libera, S. Hada, M. Hondo, P. Hallam-Baker, C. Kaler, J. Klein. LaMacchia, P. Manferdelli, and J. H Maruyama. (2002, April 5). Web Services Security (WS-Security) Version 1.0, Available: http://msdn.microsoft.com/Webservices/default.aspx?pull=/library/en-us/dnglobspec/html/ws-security.asp.

[3] C.Shahabi and F. B.Kashani "Efficient and Anonymous Web-Usage Mining for Web Personalization, " *INFORMS Journal on Computing*, Vol. 15 No.2, pp.123-147, 2003.

[4] E. A. Selma." Development of a Web Access Control Technique Based on User Access Behavior", PhD Thesis, Faculty of Engineering, University Putra Malaysia, Malaysia, 2004.

[5] J. Han, and M. Kamber. "Data Mining: Concepts and Techniques," San Francisco, Morgan Kaufmann Publishers, 2 001.

[6] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, and Q. Chen."Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach," *Journal of Transactions On Knowledge And Data Engineering 10(16),* 2003.

[7] H. Kiyomitsu, A. Takeuchi, and K. Tanaka, "ActiveWeb: XML-Based Active Rules for Web View Derivations and Access Control". Pages 31-39. Proceedings of IEEE Conference on Internet and Computer Security (ICSC '01), Las Vegas, USA 2001.

[8] L. Bauer, S. Edward, and , W. Felton. "A General and Flexible Access-Control System for the Web, Secure Internet Programming Laboratory," *in Proc. 11th USENIX Security Symposium, Department of Computer Science. San Francisco, USA,* 2002.

[9] M. Mohania, V. Kumar, Y. Kambayashi, and B. Bhargava, "Secured Web Access," Proceedings of the Kyoto International Conference on Digital Libraries, Research and Practice,Kyoto:, Japan, 2000.

[10] M. Mahoui, B. Bhargava and M. Mohania, "Data Mining For Web Security: UserWatcher," Proceedings of the IC'2001 Conference, Las Vegas, USA 2001.

[11]P. Bonatti, P. Samarati, "A unified framework for regulating access and information release on the Web", Journal of Computer Security, Vol. 10 No.3, pp.241-72, (2002).

[12] R. Bhatti, J. B. Joshi, E. Bertino, and A. Ghafoor. " Access Control in Dynamic XML-based Web-Services with XRBAC ," in Proc. 1th International Conf. Web Services, Las Vegas, USA, June 23-26, 2003.

[13] R. Kosala, H. Blockee. Web Mining Research: A Survey. SIGKDD Explorations, July 2000.

[14] Y. Zhong, and B. Bhargava, "Authorization on Web Access," Report No. 50-155, ,CERIAS, Purdue, USA, 2001.