
UNIVERSITI PUTRA MALAYSIA

# DEVELOPMENT OF A WEB ACCES CONTROL TECHNIQUE BASED ON USER ACCESS BEHAVIOR

SELMA ELSHEIKH ABDELRAHMAN.



# DEVELOPMENT OF A WEB ACCESS CONTROL TECHNIQUE BASED ON USER ACCESS BEHAVIOR

SELMA ELSHEIKH ABDELRAHMAN

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA

2004

# DEVELOPMENT OF A WEB ACCESS CONTROL TECHNIQUE BASED ON USER ACCESS BEHAVIOR

By

**SELMA ELSHEIKH ABDELRAHMAN**



Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

June 2004

To my father Elsheikh, mother Khadiga, sisters and brothers

*I dedicate this work with all my love and appreciation*

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Doctor of Philosophy

**DEVELOPMENT OF A WEB ACCESS CONTROL ECHNIQUE BASED ON USER ACCESS BEHAVIOR**

**By**

**SELMA ELSHEIKH ABDELRAHMAN**

**June 2004**

**Chairman: Professor Ir. Mohamed Daud, Ph.D.**

**Faculty: Engineering**

The development and the wide spread use of the World Wide Web allow for convenient electronic data storage and distribution all over the world. This convenience has forced organizations in both private and public sectors to make their data available on the web with restricted or limited use. These data includes sensitive data that can be released only to specific requesters. This situation calls for the need of a access control techniques capable of capturing and enforcing the different requirements that the data producer (publisher) may need to control access their data. In fact, there is a need for fine-grained access control techniques which limit access of specific individuals to resources. Previous studies have not yet designed such a system that is reliable enough for such critical applications.

This thesis discusses about designs and develops techniques and algorithms for performing web access control. The major objective of the proposed technique referred to as a Secure Web Access Control (SWAC) is to provide mechanisms for control web

access based on user access behavior. The SWAC controls access to the web pages depending on user password, date of last request, page visited (URL) and status action. In SWAC technique active user's access transaction pattern is matched with user access transaction pattern discovered from user access history based on mining techniques. A set of algorithms is used for mining user access behavior, preprocessing tasks for data preparation, association rules for defining the rules that describe the correlation between web user access transaction entries patterns, and sequential pattern discovery for finding the sequences of the web user access transaction entries pattern using PrefixSpan (Pattern growth via frequent sequence lattice) algorithms. The output is filtered using the query database system (SQL structure query language) to produce the interested web user access transaction entries pattern. Finally the rules induction is applied to the output pattern to make the access control decision (page access is permitted or denied).

The necessary steps for the proposed technique are identified, and algorithms of these steps are developed and implemented using Active Server Page (ASP) and then tested on two web pages.

The results show that proper preprocessing of the web user access transaction data is required to obtain meaningful user access transaction patterns that could be used to design web access control based on user access behavior. In SWAC the evidence combination technique is developed to provide an access control technique that allows only the authorized users to access to the web data and controls their access authorization. The technique determines which users can access web page resources and

ensures that access is restricted to authorized users who have been successfully authenticated. The results of testing the SWAC show good results.

The study concludes that limited access to web page resources based on knowledge discovery from a user access behavior gives practical and desirable web access control, and thus is an interesting research direction for future work.

Abstrak tesis dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**PEMBANGUNAN TEKNIK PENGAWALAN AKSES HALAMAN WEB BERDASARKAN KEPADA PERLAKUAN AKSES PENGGUNA**

**Oleh**

**SELMA ELSHEIKH ABELRHAMAN**

**Jun 2004**

**Pengerusi: Profesor Ir. Mohamed Daud, Ph.D.**

**Faculti: Kejuruteraan**

Penggunaan meluas WWW telah memudahkan penyimpanan dan pengagihan data elektronik di seluruh dunia. Organisasi-organisasi swasta dan awam semakin dikehendaki menyediakan data supaya mudah dicapai oleh pihak-pihak yang berkenaan. Situasi ini memerlukan sistem kawalan capaian yang berkuasa dan anjal untuk menguat-kuasakan capaian kepada pelbagai pengguna yang mempunyai keperluan yang berbeza. Penyelidikan sebelum ini belum lagi dapat mereka bentuk sebuah sistem yang boleh diharap unutk aplikasi-aplikasi yang kritikal.

Tesis ini membangunkan dan melaksanakan rekabentuk dan algoritma untuk menguat-kuasakan kawalan kepada capaian web. Objektif utama sistem yang dicadangkan ialah untuk mendefinasikan kawalan kepada capaian web yang berdasarkan kepada kelakuan capaian pengguna. Corak kelakuan pengguna yang aktif dipadankan dengan data yang ditemui daripada sejarah kawalan pengguna pada web menggunakan teknik-teknik

penggalian. Suatu kumpulan algoritma-algoritma digunakan untuk menggali sejarah capaian pengguna, pra-pemprosesan untuk penyediaan data, peraturan dan penemuan corak berturutan untuk mendefinasikan corak peraturan capaian berdasarkan kepada algoritma "PrefixSpan". Kemudian corak capaian pengguna dianalisa menggunakan pengaturcaraan induktif logik (ILP) dan sistem pengkalan data SQL untuk menentukan keputusan kawalan capaian (halaman dibenarkan atau dilarang).

Langkah-langkah yang diperlukan untuk membangunkan sistem yang dicadangkan dikenalpasti. Algoritma-algoritma untuk langkah-langkah tersebut dibangunkan dan dilaksanakan menggunakan Halaman Pelayan Aktif "ASP" yang kemudiannya diujikan kepada dua halaman web.

Keputusan yang didapati menunjukkan pra-pemprosesan yang sesuai mengenai data transaksi web oleh pengguna adalah diperlukan untuk mendapatkan corak capaian pengguna yang bermakna. Ini boleh digunakan untuk merekabentuk kawaln capaian web. Sistem kombinasi yang digelar Kawalan Capaian Web Selamat (SWAC) telah dibangunkan untuk menyediakan kawalan capaian web dan capaian yang selamat kepada sumber-sumber web secara berkesan. Sistem ini dapat menentukan pengguna-pengguna yang boleh mencapai sumber halaman web dan memastikan capaian adalah dihadkan kepada pengguna-pengguna yang dibenarkan iaitu pengguna-pengguna yang telah disahkan. Keputusan daripada pengujian menunjukkan SWAC mempunyai prestasi yang baik.

Penyelidikan ini merumuskan bahawa kawalan capaian berdasarkan penemuan pengetahuan daripada sejarah capaian pengguna memberikan kawalan capaian web seperti yang dikehendaki dan praktikal, justeru itu ia adalah bidang penyelidikan yang menarik untuk diteruskan pada masa hadapan.

## ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Professor Dr. Ir. Mohamed Daud, for whose expertise, understanding and patience added considerably to my graduate experience. I appreciate his assistance in writing reports (i.e., research reports, conference papers, and this thesis), and his generosity in sending his students to the top conferences, which opened our eyes for research.

My deep appreciation and gratitude also goes to Professor Dr. Ir. Dato' Mohd Zohadie Bardaie, member of my supervisory committee for his patience, kind co-operation and thoughtful suggestions, for my study. I would also like to acknowledge with great gratitude for the invaluable guidance and suggestions of my supervisory committee member Assoc. Prof. Dr. Hj. Md. Nasir Sulaiman.

I must also acknowledge many of my lab mates for their suggestions and discussion. We have worked together and developed true friendships that will benefit my future career.

I must acknowledge with gratitude the government of Sudan and University of Gezira for sending me to pursue my postgraduate degree. I would also like to acknowledge the government of Malaysia for funding this research project through its Intensified Research Priority Area (IRPA) grand program.

My deep appreciation and sincere gratitude to my parents, sisters, brothers and friends for their encouragement, patience and strong support, which have been a source of inspiration that kept me going. And finally I feel special gratitude to my brother Abobekr and my sister Samia

## TABLE OF CONTENTS


| | | Page |
|---|---|---|
| **DEDICATION** | | ii |
| **ABSTRACT** | | iii |
| **ABSTRAK** | | iv |
| **ACKNOWLEDGEMENTS** | | v |
| **APPROVAL** | | vi |
| **DECLARATION** | | vii |
| **LIST OF TABLES** | | viii |
| **LIST OF FIGURES** | | ix |
| **CHAPTER** | | |
| 1 | INTRODUCTION | 1.1 |
| | 1.1 Background | 1.1 |
| | 1.2 Problem Statement | 1.3 |
| | 1.3 Research Goal | 1.6 |
| | 1.4 Objectives of the Research | 1.6 |
| | 1.5 Contribution of the Research | 1.7 |
| | 1.6 Scope of the Study | 1.8 |
| | 1.7 Organization of the Study | 1.9 |
| 2 | Access Control Techniques | 2.1 |
| | 2.1 Introduction | 2.1 |
| | 2.2 Security Concepts | 2.4 |
| | 2.3 Access Control | 2.5 |
| | 2.3.1 Access Control Definition | 2.6 |
| | 2.3.2 Identification and Authentication | 2.7 |
| | 2.3.3 Access Control Fundamentals | 2.8 |
| | 2.3.4 Access Control Structures | 2.9 |
| | 2.3.5 Role-based Access Control | 2.13 |
| | 2.4 Security Policies | 2.15 |
| | 2.4.1 Mandatory Access Controls (MAC) | 2.15 |
| | 2.4.2 Discretionary Access Control (DAC) | 2.16 |
| | 2.5 Review of Security and Access Models | 2.17 |
| | 2.5.1 Access Control Matrix Model | 2.18 |
| | 2.5.2 The Bell–LaPadula Clearance Classification Model | 2.19 |
| | 2.5.3 The Shen–Dewan Collaborative Access Model | 2.20 |
| | 2.6 Distributed Access Control | 2.21 |
| | 2.7 Operating Systems | 2.25 |
| | 2.7.1 UNIX | 2.27 |
| | 2.7.2 Windows NT | 2.27 |
| | 2.8 Approaches to Access Control Management | 2.28 |

| | | | |
|---|---|---|---|
| | 2.8.1 | IP Source Address Filtering | 2.28 |
| | 2.8.2 | Firewalls | 2.29 |
| | 2.8.3 | Credential Based Approaches | 2.30 |
| | 2.8.4 | Cookies | 2.32 |
| 3 | LITERATURE REVIEW | | 3.1 |
| | 3.1 | Introduction | 3.1 |
| | 3.2 | The Scope of Data Mining | 3.3 |
| | 3.2.1 | Automated Prediction of Trends And Behaviors | 3.3 |
| | 3.2.1 | Automated Discovery of Previously Unknown Patterns | 3.3 |
| | 3.3 | What kind of Data can be mined? | 3.4 |
| | 3.3.1 | Flat files | 3.4 |
| | 3.3.2 | Relational Databases | 3.5 |
| | 3.3.3 | Data Warehouses | 3.6 |
| | 3.3.4 | Transaction Databases | 3.6 |
| | 3.3.5 | Multimedia Databases | 3.6 |
| | 3.3.6 | Time-Series Databases | 3.7 |
| | 3.3.7 | World Wide Web | 3.7 |
| | 3.4 | WEB Mining | 3.9 |
| | 3.5 | Web Content Mining | 3.10 |
| | 3.6 | Web Structure Mining | 3.11 |
| | 3.7 | Web Usage Mining | 3.13 |
| | 3.8 | Pre-Processing Tasks | 3.13 |
| | 3.8.1 | Data Cleaning | 3.14 |
| | 3.8.2 | Transaction Identification | 3.15 |
| | 3.9 | Pattern Discovery | 3.15 |
| | 3.9.1 | Path Analysis | 3.16 |
| | 3.9.2 | Statistical Analysis | 3.16 |
| | 3.9.3 | Classification and Clustering | 3.17 |
| | 3.9.4 | Dependency Modeling | 3.17 |
| | 3.9.5 | Association Rules | 3.23 |
| | 3.9.6 | Sequential Discovery Patterns | 3.33 |
| | 3.10 | Pattern Analysis | 3.33 |
| | 3.10.1 | Visualization Techniques | 3.34 |
| | 3.10.2 | On-Line Analytical Processing (OLAP) Techniques | 3.34 |
| | 3.10.3 | Data and Knowledge Querying | 3.35 |
| | 3.11 | Rule Induction | 3.36 |
| | 3.12 | Data Sources | 3.36 |
| | 3.12.1 | Server-Level Collection | 3.37 |
| | 3.12.2 | Client Level Collection | 3.38 |
| | 3.12.3 | Proxy Level Collection | 3.38 |
| | 3.13 | WEBMINER Architecture | 3.38 |
| | 3.14 | Web Personalization | 3.41 |
| | 3.15 | Web Mining in Web Security | 3.42 |
| | 3.16 | Summary | 3.45 |

| | | | |
|---|---|---|---|
| 4 | METHODOLOGY | | 4.1 |
| | 4.1 | Introduction | 4.3 |
| | 4.2 | Basic Web Architecture | 4.3 |
| | 4.3 | The General Framework of the SWAC | |
| | 4.4 | Access Control Infrastructure of SWAC | 4.7 |
| | 4.5 | The SWAC Access Control Structure | 4.9 |
| | 4.6 | SWAC Architecture | 4.11 |
| | 4.7 | SWAC Technical Design | 4.13 |
| | | 4.7.1 Data Preprocessing | 4.14 |
| | | 4.7.2 Pattern Discovery | 4.16 |
| | | 4.7.3 Pattern Analysis | 4.22 |
| | | 4.7.4 Make Decision | 4.22 |
| | 4.8 | Data Sources | 4.23 |
| | 4.9 | Development Environment and Tools | 4.25 |
| | 4.10 | SWAC Design Flow Chart | 4.26 |
| 5 | SWAC TECHNIQUE BASED ON USER BEHAVIOUR | | 5.1 |
| | 5.1 | Introduction | 5.1 |
| | 5.2 | Web Data Preprocessing | 5.3 |
| | | 5.2.1 Data Interesting | 5.4 |
| | 5.4 | Pattern Discovery | 5.8 |
| | | 5.4.1 Association Rules Discovery | 5.8 |
| | | 5.4.2 Sequential Discovery Pattern | 5.11 |
| | 5.5 | Pattern Analysis | 5.22 |
| | | 5.5.1 Structure Query Language (SQL) Algorithms | 5.22 |
| | 5.6 | Make decision | 5.23 |
| 6 | SWAC Prototype Development | | 6.1 |
| | 6.1 | Overview | |
| | 6.2 | SWAC Requirements Definition | 6.1 |
| | | 6.2.1 SWAC Functional Requirements | 6.1 |
| | | 6.2.2 Non-Functional Requirements | 6.2 |
| | 6.3 | System Development | 6.3 |
| | | 6.3.1 Data Flow in SWAC Description | 6.4 |
| | 6.4 | SWAC Development Components | 6.6 |
| | | 6.4.1 Preprocessing and Cleaning | 6.8 |
| | | 6.4.2 Association Rule Mining | 6.19 |
| | | 6.4.3 Sequential Pattern Mining | 6.21 |
| | | 6.4.4 Pattern Analysis | 6.24 |
| | | 6.4.5 Make Access Decision | 6.25 |
| 7 | TESTING AND VALIDATION | | 7.1 |
| | 7.1 | Introduction | 7.1 |
| | 7.2 | SWAC Testing (Verification) | 7.1 |
| | | 7.2.1 Date Preprocessing | 7.3 |
| | | 7.2.2 Pattern Discovery | 7.4 |

| | | | |
|---|---|---|---|
| | 7.2.3 | Pattern Analysis | 7.7 |
| | 7.2.4 | Make Decision | 7.7 |
| 7.3 | | Validation | 7.7 |
| 7.4 | | SWAC Capabilities | 7.18 |
| 8 | | CONCLUSION AND FUTURE WORK | 8.1 |
| 7.1 | | Conclusion | 8.1 |
| 7.3 | | Future Work | 8.3 |
| REFERENCES | | | R.1 |
| APPENDICES | | | A.1 |
| BIODATA OF THE AUTHOR | | | D.1 |

## LIST OF TABLES


| Table | | Page |
|---|---|---|
| 2.1 | Access Control Matrix | 2.11 |
| 2.2 | Capabilities | 2.12 |
| 2.3 | Access Control Lists (ACL) | 2.13 |
| 3.1 | Comparison between GSP Algorithm, SPADE Algorithm, FreeSpan Algorithm and Prefix Span Algorithm | 3.32 |
| 4.1 | Web Server Log Field Description (Example) | 4.24 |
| 5.1 | Set of Users Access Transactions Entry (Sample) | 5.5 |
| 5.2 | Examples of Users Transaction Entries | 5.7 |
| 5.3 | The Minimum Support (Frequency) of the Web User Transaction Entries Pattern for selected Single User Access | 5.9 |
| 5.4 | All the Possible Output Patterns from Pattern Discovery Stage | 5.12 |
| 5.5 | Web Access Transaction Entry Patterns Sequence and Web Access Transaction Entries Pattern | 5.19 |
| 6.1 | The Web User Access Transactions Entry Data Description | 6.13 |
| 6.2 | Web User Access Transaction Procedures Sample Output | 6.15 |
| 6.3 | The Output of Running Association Rules (Sample) | 6.20 |
| 7.1 | Algorithms Result Output (Sample) | 7.5 |
| 7.2 | The Possible Output Patterns from Pattern Discovery Algorithms (Sample) | 7.6 |

## LIST OF FIGURES


| Figure | | Page |
|---|---|---|
| 2.1 | The Access Control Fundamentals | 2.8 |
| 2.2 | Model of Role-Based Access Control (RBAC) | 2.14 |
| 2.3 | Distributed Access Control | 2.22 |
| 3.1 | Taxonomy of Web Mining | 3.8 |
| 3.2 | Web Usage Mining Phases | 3.12 |
| 3.3 | WEBMINER Architecture | 3.40 |
| 4.1 | SWAC Phases of Activities | 4.2 |
| 4.2 | Simplified Web Access Diagram | 4.5 |
| 4.3 | The General Framework of SWAC | 4.6 |
| 4.4 | The Access Control Infrastructure of SWAC | 4.8 |
| 4.5 | SWAC Access Control Structure | 4.10 |
| 4.6 | SWAC Design Architecture | 4.12 |
| 4.7 | SWAC Detailed Design | 4.14 |
| 4.8 | Prefix- based Frequent Sequence | 4.21 |
| 5.1 | The Control Flow of Mining Process in SWAC | 5.2 |
| 5.2 | The Sequence Selection Process in the Selected Example | 5.16 |
| 5.3 | Frequent Sequence of the Web Access Transaction Entries Pattern (Example) | 5.20 |
| 6.1 | An Overview of Data Flow Diagram in SWAC Prototype | 6.5 |
| 6.2 | SWAC Development Components | 6.7 |
| 6.3 | The Procedure Capturing the Server Log File (sample) | 6.8 |

| | | |
|---|---|---|
| 6.4 | Server Log Raw Data (Sample) | 6.9 |
| 6.5 | Registration Data Procedure (Sample) | 6.11 |
| 6.6 | Registration Menu | 6.12 |
| 6.7 | The Procedure Generates the Web User Access Transactions | 6.14 |
| 6.8 | Password Procedure (Sample) | 6.16 |
| 6.9 | Screen Display the Password and the Date of Last Request Menu | 6.17 |
| 6.10 | Procedure to Generate the Date of the Last Request (Sample) | 6.18 |
| 6.11 | Data Flow Diagram for PrefixSpan Algorithm in SWAC | 6.23 |
| 6.12 | Access Decision Procedures (Sample) | 6.27 |
| 7.1 | ASP Scripts to Prevent the Page from Being Cached (Example) | 7.3 |
| 7.2 | Medical Homepage Hierarchy | 7.10 |
| 7.3 | Link Structure in Medical Center Homepage | 7.10 |
| 7.4 | Screen Display for New User (Example 1) | 7.11 |
| 7.5 | Screen Display for Registered User (Example1) | 7.12 |
| 7.6 | Screens Display the Main and its Sub-pages (Exampe1) | 7.13 |
| 7.7 | ESRG Page Hierarchy | 7.15 |
| 7.8 | ESRG Page Link Structure | 7.15 |
| 7.9 | Screen Display for New User (Example 2) | 7.16 |
| 7.10 | Screens Display the Main Page and the Sub-pages Link (Example 2) | 7.17 |

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

The evolution of the Internet is one of the most important phenomena in information technology. Its users are rapidly increasing in number and variety, from companies with high-speed networks to individuals with slow modem connections. A major technology on both Internet and Intranets is the World Wide Web (WWW), often called simply "the web". The rapid development of the web, with increasing popularity and ease of use of its tools, becomes the most important media for collecting, sharing and distributing information.

The World Wide Web consists of both a web server and a web browser (or client). The former delivers HTML and other media to browsers through the Hyper Text Transfer Protocol (HTTP) and the latter provides a user interface to navigate through information by pointing and clicking. The main function of the web browser is to use the Uniform Resource Locator (URL) to retrieve documents from the web server (Oppliger, 2000). Moreover, web browsers allow users to access aspects of the Internet via the World

Wide Web. Nearly all types of transactions can occur online today: banking, shopping, education, communication, etc. Web browsers facilitate the completion of these transactions, but they also provide a means of keeping one's personal information private.

As the web has become a social infrastructure for data sharing and information management, the need to process and classify a large amount of diverse information resources within an enterprise, and make them available to a larger set of diverse users has increased, making security issues become more critical. However, networks are all about the sharing of programs and data, both internally and externally. Accordingly, implementing and maintaining security is a prerequisite to protect data and systems from unexpected loss or unauthorized access while allowing all necessary processes to take place with a minimum impact on the users.

Information and system security is a multi-faceted discipline and deserving of interest from researchers and funding agencies. The principal security technologies today are cryptography, authentication, intrusion detection, assurance, and access control (Labs, 2002). The access control is used to determine which user is allowed to access to a site and what information. Access control for Internet information processing, in contrast to access control in a traditional operating system, has a higher demand for dealing with much larger scale problems in real time, due to the large amounts of information and number of users in the internet/intranet environment.

## 1.2 Problem Statement

Currently, many people are trying to figure out how to use the web effectively. In most cases, the primary focus is on using it to create, manage, find, and deliver stored information. An excellent web design will inevitably be tarnished by the lack of thought given to security and access control management. Since the web becomes a viable method for corporations to connect to their users, partners, branch offices and remote employees, sensitive information should only be accessible to a group of users depending on their right information access.

The data distribution process is selective; the data cannot just be released to anybody. Rather, specific data can usually be released only to specific requesters or under specific conditions. There are data which are subject to embargoes and can be released to the general public only after a specific time; there are data that can be released only for non-commercial purposes; and data which do not bear sensitivity, but whose release is subject to payment. Many and many examples can be mentioned, but these few can already give an idea of the variety of protection requirements that may need to be enforced. In this distributed environment, where a large a mount of diverse information resources within an enterprise will be made available for group of diverse to query. Existing widely deployed access control tools are inflexible, and do not provide all functionality need. This situation calls for the need of web access control techniques able to capture and enforce the different requirements that the data producers may need to enforce on the web data access. These techniques should have abilities to support the