

سامية محمد مصطفى



شبكة المعلومات الجامعية

بسم الله الرحمن الرحيم



سامية محمد مصطفى



شبكة المعلومات الجامعية



شبكة المعلومات الجامعية التوثيق الالكتروني والميكروفيلم



سامية محمد مصطفى



شبكة المعلومات الجامعية

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات



يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار



سامية محمد مصطفى



شبكة المعلومات الجامعية



بعض الوثائق الأصلية تالفة



سامية محمد مصطفى



شبكة المعلومات الجامعية



بالرسالة صفحات لم ترد بالأصل



Cairo University
Institute of Statistical Studies and Research
Department of Computer and Information Sciences

Web Personalization by Pattern Discovery
from web Usage

Master Thesis

Submitted By

Osama Hashim Khamis Mubarak

Supervised By

Prof.Dr. Mahmoud Riad

Dr.Eng Laila Nassef

Dr. Mohammed EL Said EL Telbany

A thesis submitted to the Institute of Statistical Studies and Research, Cairo
University in partial fulfillment of the requirements for the degree of
master of computer science, in the department of computer and
information sciences.

2005

B

18724

APPROVAL SHEET

Web Personalization by Pattern Discovery from Web Usage

M.S.c Thesis

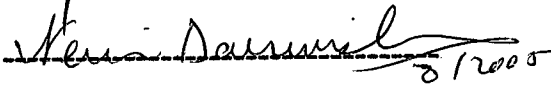
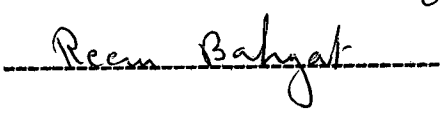
By

Osama Hashim Khamis

This thesis is for M.SC. Degree in Computer and Information Science,
Department of Computer and Information Science, Institute of Statistical
Studies and Research, Cairo University, has been approved by:

Name
Prof.Dr. Mahmoud R.Mahmoud
Prof.Dr. Nevin M.Darwish
Prof.Dr. Reem M.Bahgat

Signature

 8/2005


2005

STATEMENT

I certify that this work has not been accepted or submitted in candidature for any other degree.

Any portion of this thesis for which I am indebted to other sources are mentioned and explicit references are given.

Osama Hashim Khamis Mubarak

Date:

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

قال تعالى:

{ وَقُلْ رَبِّ زِدْنِي عِلْمًا }

سورة طه (114)

صَلَّى الْعِظِيمِ

*This Thesis is dedicated to my Parents and my
Brothers*

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor professor Mahmoud Riad for his supervision, continuous follow-up and guidance throughout this work. I also would like to express my deepest appreciation to my supervisor Dr. Laila Nassef for her faithful assistance and potential help in terminating this work. I am also appreciating my deep thanks to my previous supervisor Dr. Mohammed EL Said EL Telbany for providing guidance and moral supports.

Finally my deepest thanks belong to my family for their continuous encouragement and potential support.

ABSTRACT

Discovering and extracting useful information about the world wide web visitors is an important area that affects e-commerce, e-services, and e-learning, etc. Generally, most of the efforts focus on extracting useful patterns and rules using data mining techniques in order to understand the users' navigational behaviour, so that decisions concerning site restructuring or modification can then be made by site managers. Some of the more advanced systems provide much more functionality, introducing the notion of adaptive web site and providing means of dynamic changing a site structure. One possible approach to web usage personalization is to mine the users' profiles from vast amount of historical data stored in access logs. An Active Node Technique is proposed to overcome drawbacks of prior systems in web usage personalization. This technique deals with creating users' profiles in an incremental fashion. The created profiles are only based on the prior traversal patterns of the user on the web site and do not involve providing any declarative information or the user to log in. User profiles are dynamic in nature. The traversal patterns of a user change with time. In order to reflect these changes on the recommendations generated for the user, the profiles have to be regenerated taking into account the existing profile. Instead of creating a new profile, information from a user profile is added or removed to save time as well as physical storage requirements. This technique considers the significance function and the scanning function (or profiling function).

The significance function is used to eliminate all pages that are insignificance to the user on each session. Page significance on each user session is based on the page weight as well as the pre specified threshold.

The scanning function is used to create each user profile by scanning all site layers to determine the nodes (pages) that the user is interested with. User nodes are represented as a vector of nodes and their associated weight for all previously visited nodes by the user.

Two types of recommendations are proposed and implemented: ***Node Recommendation and Batch Recommendation.***

In the Node Recommendation, the recommendation process is based on the current active node. Therefore, the recommendation for the same user is different from one node to another during the same session. The recommendation agent will take actions base on the requested page, which is the active node.

In the Batch Recommendation, the recommendation process is based on all previously visited pages by the current user and have weights greater than or equal the pre specified threshold. The recommendation agent scans user profile and then rank nodes based on their weights to the user, fetch top N pages and then send the requested page associated with hyperlinks of these top pages.

Two systems for personalization are proposed in the thesis. ***The first*** system is based on historical data collection form users' log files. This system consists of three phases. *The data preparation phase, the pattern discovery phase*, and *the recommendation phase* which is the only online phase on such system. ***The second*** system is based on real time data collection about the users' clickstreams. Such system consists of three online phases. *The data collection phase, the pattern discovery phase*, and *the recommendation phase*.

CONTENTS

Acknowledgement	vi
Abstract	vii
Table of contents	ix
List of figures	xiv
List of tables	xvi
 Chapter 1 Introduction	 1
1.1 Introduction	2
1.2 Web Mining and Information Retrieval	3
1.3 Web Mining and Information Extraction	4
1.4 Web Mining and Machine Learning Applied on the Web	4
1.5 Web Mining and the Agent Paradigm	5
1.5.1 User interface agents	5
1.5.2 Distributed agents	5
1.6 Web Usage Mining	6
1.7 The Proposed Technique	11
1.8 The Proposed Systems	12
1.9 Thesis Organization	13
 Chapter 2 Web Mining and Web Personalization	 14
2.1 Introduction	15
2.2 Web mining categories	15
2.2.1 Web Content Mining	16
2.2.2 Web Structure Mining	17
2.2.2.1 Web Structure Mining Tasks	17
2.2.3 Web Usage Mining	19

2.2.3.1 Data Sources for Web Usage.....	20
2.2.3.2 Web Usage Mining Architecture.....	21
2.3 Web Personalization.....	24
2.3.1 Overview	24
2.3.2 Goals of web usage personalization	26
2.3.3 Web Personalization Phases	26
2.3.3.1 Data Preparation for Personalization Phase	26
2.3.3.1.1 Pre-Processing of Usage Data	30
2.3.3.1.2 Post – Processing of User Transaction Data	34
2.3.3.2 Pattern Discovery Techniques Phase	37
2.3.3.2.1 Association Rules	38
2.3.3.2.2 Statistical analysis	40
2.3.3.2.3 Classification	40
2.3.3.2.4 Sequential and Navigational Patterns	41
2.3.3.2.5 Clustering	42
2.3.3.3 Pattern Analysis Phase	43
2.3.3.4 Recommendation Phase.....	46
2.4 Tools and Applications for Web Personalization.....	46
2.5 Web Personalization Previous Related Works.....	46
2.6 web personalization and the users' Privacy	48
2.6.1 Privacy Risks	49
2.6.2 Principals of Applying fair Information Practice	50
2.6.3 Approaches to Reducing Privacy Risks in Personalization	51
Chapter 3 The Proposed Active Node Technique	56
3.1 Introduction	57
3.2 The Active Node Technique (ANT).....	58
3.3 Implementation of the Active Node Technique	61