UNIVERSITI PUTRA MALAYSIA

A WEB-BASED RECOMMENDATION SYSTEM TO PREDICT USER MOVEMENTS THROUGH WEB USAGE MINING

MEHRDAD JALALI

FSKTM 2009 12

MEHRDAD JALALI

Doctor of Philosophy

2009

A WEB-BASED RECOMMENDATION SYSTEM TO PREDICT USER MOVEMENTS THROUGH WEB USAGE MINING

MEHRDAD JALALI

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA

2009

A WEB-BASED RECOMMENDATION SYSTEM TO PREDICT USER MOVEMENTS THROUGH WEB USAGE MINING

By

MEHRDAD JALALI

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

December 2009

Dedicated to my wife; Mahboobeh Naserzadeh
And my parents

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

A WEB-BASED RECOMMENDATION SYSTEM TO PREDICT USER MOVEMENTS THROUGH WEB USAGE MINING

By

MEHRDAD JALALI

DECEMBER 2009

Chairman: **Norwati Mustapha, PhD**

Faculty: **Computer Science and Information Technology**

Web usage mining has become the subject of exhaustive research, as its potential for Web based personalized services, prediction user near future intentions, adaptive Web sites and customer profiling is recognized. Recently, a variety of the recommendation systems to predict user future movements through web usage mining have been proposed. However, the quality of the recommendations in the current systems to predict users' future requests can not still satisfy users in the particular web sites. The accuracy of prediction in a recommendation system is a main factor which is measured as quality of the system. The latest contribution in this area achieves about 50% for the accuracy of the recommendations.

To provide online prediction effectively, this study has developed a Web based recommendation system to Predict User Movements, named as WebPUM, for online

prediction through web usage mining system and proposed a novel approach for classifying user navigation patterns to predict users' future intentions. There are two main phases in WebPUM; offline phase and online phase. The approach in the offline phase is based on the new graph partitioning algorithm to model user navigation patterns for the navigation patterns mining. In this phase, an undirected graph based on the Web pages as graph vertices and degree of connectivity between web pages as weight of the graph is created by proposing new formula for weight of the each edge in the graph. Moreover, navigation pattern mining has been done by finding connected components in the graph. In the online phase, the longest common subsequence algorithm is used as a new approach in recommendation system for classifying current user activities to predict user next movements. The longest common subsequence is a well-known string matching algorithm that we have utilized to find the most similar pattern between a set of navigation patterns and current user activities for creating the recommendations.

The quality of the navigation patterns mining in the offline phase is measured by two main parameters, namely visit-coherence which measures the percentage of the web pages inside a user session which belongs to the cluster and outliers which measures the percentage of web pages that do not belong to any cluster. In the online phase, the percentage of the accuracy, coverage and F1 are the three main parameters that are measured for qualifying the proposed recommendation system. The accuracy is number of relevant web pages retrieved divided by the total number of web pages in recommendations set, and the coverage defined as a number of relevant web pages retrieved divide by the total number of web pages that actually belong to the user sessions, and the F1 attains its maximum value when both accuracy and coverage are

maximized.

The proposed system has been tested on the two datasets, CTI dataset which is based on the Web server logs file of the host Computer Science department in DePaul University, and MSNBC dataset which is based on Web server logs for msnbc.com and news related portions of msn.com. In comparison with the previous method, named SUGGEST, the experimental results show that the percentage of the visit-coherence is higher than previous method, and the percentage of the outliers is lower than previous method in comparison with the new approach in the offline phase. Consequently, the percentages of the accuracy, coverage and F1 measurements have been enhanced by utilizing new classifying method in the online phase. Furthermore, the scalability experiments showed that the size of dataset and the number of the users in dataset are not dominant factors to the percentage of the accuracy. In this thesis, the proposed system achieves up to 56% for the accuracy, which is about 5% higher than the previous method as it means that the proposed recommendation system can predict more relevant web pages of user future movements compared to the previous method.

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

SISTEM CADANGAN BERASASKAN WEB UNTUK MERAMAL PERGERAKAN MELALUI PERLOMBONGAN PENGGUNAAN WEB

Oleh

MEHRDAD JALALI

Disember 2009

Pengerusi: **Norwati Mustapha, PhD**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Perlombongan penggunaan web telah menjadi subjek penyelidikan yang meluas, sebagaimana potensinya dalam menawarkan perkhidmatan peribadi berasaskan Web, meramal keinginan pengguna, laman web adaptif dan mengenalpasti profil pelanggan. Kini, pelbagai sistem cadangan untuk meramal pergerakan pengguna melalui perlombongan penggunaan web telah dicadangkan. Walau bagaimanapun, kualiti cadangan pada sistem semasa untuk meramal permintaan pengguna masih tidak memuaskan pengguna dalam laman web tertentu. Ketepatan ramalan pada sistem cadangan adalah faktor utama yang diukur sebagai kualiti sistem. Sumbangan terkini dalam bidang ini mencapai kira-kira 50% bagi ketepatan cadangan.

Untuk menyediakan ramalan yang berkesan dalam talian, kajian ini telah membina sistem cadangan berasaskan web untuk meramalkan pergerakan pengguna, dinamakan sebagai WebPUM, untuk ramalan dalam talian melalui sistem perlombongan penggunaan web dan mencadangkan satu pendekatan baru untuk mengkelaskan corak

pelayaran pengguna untuk meramal tujuan pengguna. Terdapat dua fasa utama dalam WebPUM; fasa luar talian dan fasa dalam talian. Pendekatan dalam fasa luar talian adalah berdasarkan algoritma pemetakan graf yang baru bagi memodelkan corak pelayaran pengguna untuk perlombongan corak pelayaran. Dalam fasa ini, satu graf tidak terarah berdasarkan pada laman web sebagai bucu-bucu graf dan darjah hubungan di antara laman web sebagai pemberat bagi graf yang dibina dengan mencadangkan formula baru untuk berat bagi setiap sisi graf. Tambahan lagi, perlombongan corak pelayaran telah dibuat dengan mencari komponen-komponen yang berkaitan dalam graf. Dalam fasa dalam talian, algoritma *longest common subsequence* digunakan sebagai satu pendekatan baru dalam sistem cadangan untuk mengkelaskan aktiviti pengguna semasa untuk meramal pergerakan pengguna seterusnya. *Longest common subsequence* adalah satu algoritma padanan rentetan yang popular yang telah kami gunakan untuk mencari corak paling serupa di antara satu set corak pelayaran dan aktiviti pengguna semasa untuk membina cadangan.

Kualiti bagi perlombongan corak pelayaran dalam fasa luar talian diukur oleh dua parameter utama, iaitu lawatan-kekoherenan yang mengukur peratusan laman web itu dalam sesi pengguna yang menjadi kepunyaan kelompok dan nilai tersisih yang mengukur peratusan laman-laman web yang tidak dipunyai mana-mana kelompok. Dalam fasa dalam talian, peratusan ketepatan, liputan dan F1 adalah tiga parameter utama yang diukur untuk kelayakan pada sistem cadangan yang dicadangkan. Ketepatan adalah bilangan laman web berkaitan yang dicapai dibahagikan dengan jumlah keseluruhan laman web dalam set cadangan, dan liputan ditakrifkan sebagai jumlah laman web berkaitan yang dicapai dibahagikan dengan jumlah keseluruhan laman web

yang sebenarnya dipunyai oleh sesi pengguna, dan F1 mencapai nilai maksimumnya semasa kedua-dua ketepatan dan liputan adalah dimaksimakan.

Sistem cadangan telah diuji pada dua dataset, dataset CTI yang berdasarkan fail log pelayan bagi hos *Computer Science department in DePaul University* dan dataset MSNBC yang berasaskan log pelayan Web untuk msnbc.com dan bahagian-bahagian berkaitan berita bagi msn.com. Dalam perbandingan dengan kaedah sebelumnya, dinamakan SUGGEST, keputusan eksperiman menunjukkan peratusan lawatan-kekoherenan adalah lebih tinggi daripada kaedah sebelumnya, dan peratusan nilai tersisih adalah lebih rendah daripada kaedah sebelumnya dalam perbandingan dengan pendekatan baru dalam fasa luar talian. Oleh yang demikian, peratusan bagi pengukuran ketepatan, liputan dan F1 telah meningkat dengan menggunakan kaedah pengkelasan baru dalam fasa dalam talian. Selain itu, eksperimen penskalaan telah menunjukkan saiz dataset dan bilangan pengguna bukan faktor dominan untuk peratusan ketepatan. Dalam tesis ini, sistem cadangan telah mencapai sehingga 56% ketepatan, dimana 5% lebih tinggi daripada kaedah sebelumnya yang bermakna sistem cadangan yang telah dicadangkan boleh meramal lebih banyak laman web yang berkaitan bagi pergerakan pengguna berbanding dengan kaedah sebelumnya.

ACKNOWLEDGEMENTS

My thanks to God for all things throughout my voyage of knowledge exploration.

First and foremost, I would like to express my sincere gratitude to my supervisor Dr. Norwati Mustapha for giving me an opportunity to start off this project. Through the course of my study, I have had the great fortune to get to know and interact with her. Her comments and suggestions for further development as well as her assistance during writing this thesis are invaluable to me. Her talent, diverse background, interest, teaching and research style has provided for me an exceptional opportunity to learn and made me become a better student.

I would like to express my sincere thanks and appreciation to the supervisory committee members Associate Professor Dr. Md Nasir Sulaiman and Associate Professor Dr. Ali Mamat for their guidance, valuable suggestions and advice throughout this work in making this a success.

My deepest appreciation to my wife Ms. Mahboobeh Naserzadeh, who has been very supportive and patiently waiting for me to complete my study. Finally, I owe my sincere thanks to my parents for their encouragement and affirmation, which made it possible for me to achieve this work

Mehrdad Jalali
December 2009

APPROVAL

I certify that an Examination Committee met on 28th Dec 2009 to conduct the final examination of **Mehrdad Jalali** on his Doctor of Philosophy thesis entitled " WebPUM: A Web based recommendation system to predict user movements through Web usage mining" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee were as follows:

RAHMITA WIRZA O.K. RAHMAT,PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

SHYAMALA C. DORAISAMY, PhD
Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

MASRAH AZRIFAH AZMI MURAD, PhD
Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

MUDUR SUDHIR P., PhD
Professor
Faculty of Engineering and Computer Science
Concordia University, Canada
(External Examiner)

_____
BUJANG KIM HUAT, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of philosophy. The members of the Supervisory Committee were as follows:

Norwati Mustapha, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

MD. Nasir B Sulaiman, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Ali Mamat, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

_____

HASANAH MOHD. GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 17 March 2010

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Motivation and Background

Given the growth rate of the Web, proliferation of e-commerce, Web services, and Web-based information systems, the volumes of clickstream and user data collected by Web-based organizations in their daily operations has reached huge proportions. Meanwhile, the substantial increase in the number of websites presents a challenging task for webmasters to organize the contents of the websites to cater to the needs of users. Modeling and analyzing web navigation behavior is helpful in understanding what information of online users demand. Following that, the analyzed results can be seen as knowledge to be used in intelligent online applications, refining web site maps, web based personalization system and improving searching accuracy when seeking information. Nevertheless, an online navigation behavior grows each passing day, and thus extracting information intelligently from it is a difficult issue. Web usage mining refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites (Cooley, Mobasher, & Srivastava, 1997; Srivastava, Cooley, Deshpande, & Tan, 2000; Wang & NetLibrary, 2006). Web usage mining has been used effectively as an approach to automatic personalization and as a way to overcome deficiencies of traditional approaches such as recommendation systems based on user's profile.

The goal of personalization based on Web usage mining is to recommend a set of objects to the current (active) user, possibly consisting of links, advertisement, text, products, and so forth, tailored to the user's perceived preferences as determined by the matching usage patterns. This task is accomplished by matching the active user session with the usage patterns discovered through Web usage mining.

Web based recommendation systems help users overcome information overload, support them in finding items of interest for efficiency, and assist them navigating large information spaces in the web. The input to such systems can be explicit ratings provided by users representing their likes and dislikes, or implicit indicators gathered through analysis of users' interactions. In general, implicit information is easier to gather in quantity than ratings, which require user initiative. For example, Amazon.com monitors each visitor's activity and uses this information to build a navigational profile. Such recommenders model and analyze users' navigational behavior as stored in access logs, a process that is commonly referred to as web usage mining (Cooley et al., 1997; Srivastava et al., 2000).

Web based recommendation systems through Web usage mining usually process the information related to users sessions, that is, a sequence of pages requested by the same user. For instance the sequence: **Home ➔ Faculty ➔ Computer Science ➔ Courses ➔ Advanced Algorithms** could be a typical session of a person browsing a University Web site with an interest in **Advanced Algorithms**. Web servers capture all users sessions and keep them into the logs file. Furthermore, a set of clusters of Web pages is

created by navigation pattern mining. To create prediction for the users, the recommendation system captures current user activities in the particular Web site. Subsequently, based on this information and the set of clusters, the system classifies current user activities and creates a set of Web pages as predictions for the user. For instance, if a user browse **Advanced Algorithms** Web page as above example, the recommendation system predicts **Java**, **Advanced Programming Languages** and **Visual Basic.Net** Web pages, because regarding to logs file and extracted patterns from logs file by navigation pattern mining, for many users, if they are interested in Advanced Algorithm, They are also interested in **Java**, **Advanced Programming Languages** and **Visual basic.Net**.

All recommendation systems attempt to find architecture and algorithm to improve quality of the personalized recommendation, but the recommendations still does not meet satisfaction in term of the percentage of the accuracy as main parameter to measure the quality of the recommendations (Baraglia & Silvestri, 2007; H. Liu & Kešelj, 2007).

To provide online prediction effectively in term of accuracy and coverage, we have developed a Web based recommendation system to Predict User Movements, named as WebPUM, for online prediction through web usage mining system and propose a novel approach for classifying user navigation patterns to predict users' future intentions and movements. The approach is based on the new graph partitioning algorithm to model user navigation patterns for the navigation patterns mining phase. Furthermore, longest common subsequence algorithm is used for classifying current user activities to predict user next movements.

In this thesis, the proposed system achieves the prediction accuracy of nearly 56%, which is about 5% higher than the prediction accuracy by the previous work. The accuracy of prediction in a recommendation system is a main factor which is used to measure to quality of the system.

1.2 Problem Statement

The Web is an integral part of today's business dealings as e-business. Companies and institutions exploit the Web to conduct their business; customers make daily use of the Net to perform all kinds of transactions. In addition, most users browse through pages of personal interest. The Web, as we know, is massive and its data collected from countless sources. Consequently, search tools should be able to accurately extract, filter, and select what is "hidden" from such tools.

Web usage mining (WUM) typically extracts knowledge on the clickstream data (usage logs file) by analyzing historical data such as Web server access logs, browser caches, or proxy logs. WUM techniques are important for several reasons. It is possible to model user behavior for predicting their future movements. The information mined can subsequently be used in order to personalize the contents of Web pages, to improve Web server performance, to structure a Web site according to the preferences expressed by the users, or to help the business to carry out a specific users' target.

Some works have been done in the area of the Web usage mining recommendation systems (Cooley et al., 1997; Mobasher, Cooley, & Srivastava, 1999). This continuous effort has one common goal to achieve, that is to produce an accurate recommendation system for prediction of user future movement in the particular Web sites (Baraglia & Silvestri, 2007; H. Liu & Kešelj, 2007).

The main problem in Web usage mining recommendation systems is the accuracy of recommendations. In the Web usage mining recommendation systems, accuracy of recommendation is the percentage of the best matching between current user sessions by a cluster in navigational pattern profiles. In spite of many efforts to improve the accuracy of the recommendations, these studies cannot lead the users to the accurate results (Baraglia & Silvestri, 2007; Fu, Sandhu, & Shih, 1999; H. Liu & Kešelj, 2007; Mobasher, Cooley, & Srivastava, 2000; Perkowitz & Etzioni, 2000a, 2000b; Shahabi, Banaei-Kashani, Chen, & McLeod, 2001).

In the current Web usage mining recommendation systems, the quality of recommendations still does not satisfy users in the particular web sites. The quality of the recommendations in Web usage mining recommendation systems can be measured by some parameters in terms of the percentage of the accuracy, coverage and F1 measure. Moreover, some works have done to improve the architecture of Web based recommendation systems through Web usage mining. The architecture of this type of the recommendation systems should be designed to achieve more accurate recommendations. According to earlier discussion, this problem can be broken down into the following two sub problems.