AIN SHAMS UNIVERSITY
Faculty of Computer &
Information Sciences
Information Systems Department

A Hybrid Approach for Intelligent Recommender Systems

A Thesis submitted in partial fulfillment of the requirements for the degree of PhD in Computer and Information Sciences

To
Department of Information Systems
Faculty of Computer and Information Sciences
Ain Shams University

By

Wedad Hussein Reyad
M.Sc. in Computer and Information Sciences (2006)
Ain Shams University

Under the supervision of

Prof. Dr. Tarek Fouad Gharib
Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

Prof. Dr. Mostafa Gadal-Haqq M. Mostafa
Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University

Dr. Rasha Mohammed Ismail
Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

September - 2014

Acknowledgement

First of all thanks to God for giving me the will and strength to finish this work. Great thanks to every member of my family who has pushed me to go on and pursue my dream.

I would like to deeply thank Prof. Dr. Tarek Gharib for believing in me and always pushing me over finish lines in times when I did not have the motivation to do so.

I would also want to thank Prof. Dr. Mostafa Gadal-Haqq for guiding my thinking and always finding a way to get the work out of difficult situations.

I would also like to thanks Dr. Rasha Ismail for being my friend, my mentor and my support all through the work.

Finally, I would like to thank everyone (friends, professors or students) who trusted in my abilities, even more than I did, and pushed me to always be better.

Abstract

The anticipation of the user's next move is one the main techniques needed for web personalization. Next page prediction aims at discovering the next page the user will visit for offering recommendations as well as pre-fetching to reduce network latency. In this work we proposed a next page prediction system that is based on a hybrid framework combining memory-based and model based recommender systems.

We offered three different representations of user preferences and tested their results on different datasets. The first representation reflected usage data by building a user-page matrix. The second approach incorporated semantic information to build a user-concept matrix. Finally the third approach offered two methods to combine usage and semantic data. The approaches yielded a 12.8% and 33% improvement in prediction accuracy for the first two approaches respectively, and a 47.3% and 54.3% for the combined approach. The system also used clustering to group users and frequent patterns which caused the prediction time to be reduced by an average of 69.2%.

Summary

The world wide web (WWW) is becoming the most accessed source for searching for information and performing day-to-day activities. It is also becoming an active medium for conducting business. With this proliferation of the internet, the amount of information and products available through it is increasing exponentially. The amount of information available made the customization of content and the recommendation of products of crucial importance.

All these challenges motivated the introduction of web recommender systems as a means for representing user preferences and recommending suitable objects. There are two approaches to recommender systems, memory-based and model-based methods. Memory-based methods store all ratings of users to generalize from them, while model-based methods develop a model of user behavior.

In this thesis, we are proposing a framework for the next page prediction that uses techniques from both memory-based and model-based recommender systems. The system builds a user-item matrix representing user preferences. Clustering is then applied to this matrix to group users with similar preferences together. The clustering results are then used to group the frequent access patterns mined from server logs into groups corresponding to clusters.

When a new user accesses the website, he is matched to his cluster, or the nearest cluster if he/she is an unknown user. To make a prediction the set of

patterns assigned to the user's cluster are searched for matching patterns. We suggested three different representations of the user-item matrix.

The first representation of user preferences is a user-page matrix showing the average time spent by each user on each page. We tested this approach on three different datasets, namely, the logs from the NASA, University of Saskatchewan, and Ain Shams University web servers. The clustering showed an average reduction in the prediction time by 22.2%, 44.4% and 69.8 % for the three datasets respectively. The approach also increased the overall prediction accuracy by 0.5%, 0.4% and 12.8% respectively.

We next suggested the introduction of semantic information to the process. A user-concept matrix was suggested to represent the set of concepts the user is interested in as extracted from the text of the pages he/she visited. We tested this approach on the Ain Shams University server logs. The introduction of semantic information offered further improvement in prediction accuracy by 33% without affecting the prediction time.

Finally, we suggested two approaches to combine the previous two methods. For the first, the average time spent on a page was used to adjust concept counts. This approach improved the accuracy even further by 47.3%. The second approach used an updated distance matrix combined from the two matrices obtained in the first two methods to represent the distances between users. This approach improved the prediction accuracy by 54.3%

The chapters of the thesis are organized as follows:

Chapter 1: Introduction

This chapter offers an introduction to the field as well as the motivation behind the work. The chapter also lists the objectives of the proposed research.

Chapter 2: Background

This chapter gives an overview of the techniques used in developing recommender systems. The chapter explores the two main categories of recommender systems, namely, memory-based and model-based systems. The chapter also presents a comparison between the two techniques for recommender systems. The concepts behind the semantic web as well as its organization are also discussed in this chapter. The chapter explores the importance of semantic information as a dimension added to the information available on the web.

Chapter 3: Related Work

The chapter explores the research done in the fields of recommender systems and semantic web mining. For recommender systems, the chapter presents the different techniques used to overcome the shortcomings of these systems. For semantic web mining, the focus is how semantic information can be integrated into recommendation systems specially web mining algorithms.

Chapter 4: Hybrid Framework for Next Page Prediction

This chapter introduces the hybrid recommender system we are proposing that incorporates memory-based and model-based techniques. Here

clustering of user-item matrix was suggested for focusing the search for the user's next page on relevant patterns. The chapter also explains the experiments conducted to test the proposed approach along with the obtained results.

Chapter 5: Semantic Data for Improved Prediction

In this chapter semantic information extracted from the text of web pages was suggested to be added to the proposed framework. Also a decision fusion approach from both usage and semantic data was introduced. The experiments conducted to test the proposed approaches were presented and compared to traditional approaches.

Chapter 6: Conclusions and Future Work

The final conclusions of our work are presented in this chapter. Also, this chapter presents the possible directions for future research.

Table of Contents

Chapter **Page**

Abstract..... I
Summary..... II
Table of Contents..... VI
List of Figures..... X
List of Tables..... XII
List of Abbreviations..... XIII

1- Introduction **1**
1.1 Motivation..... 1
1.2 Objective..... 3
1.3 Thesis Organization..... 4

2- Background **6**
2.1 Recommender Systems..... 7
2.1.1 Memory-Based Recommender Systems..... 8
A. Collaborative Filtering..... 9
B. Content-based Recommender Systems..... 16
C. Hybrid Recommender Systems..... 17
2.1.2 Model-Based Recommender Systems..... 20
A. Data Preparation..... 22
B. Behavior Modeling..... 23
C. Recommendation..... 26
2.1.3 Memory-Based vs. Model-Based Techniques..... 27
A. Incorporating New Data..... 27
B. Scalability..... 28
C. Handling New Users..... 28
D. Individuality..... 29
E. Prediction Time..... 29
F. Robustness..... 29
2.2 Semantic Web Architecture..... 29
2.2.1 Universal Resource Identifier..... 30
2.2.2 Extensible Markup Language..... 31

2.2.3 Resource Description Framework..... 32
2.2.4 RDF Schema..... 32
2.2.5 Ontology..... 33
2.2.6 Logic, Proof and Trust..... 34
2.3 Semantic Web Mining..... 35
2.3.1 Integration Areas..... 35
2.3.2 Mining for Ontology Construction and Modification..... 36
A. Ontology Construction..... 36
B. Ontology Modification..... 36
2.3.3 Semantic Annotation..... 37
2.3.4 Semantic Data for Improved Recommendations..... 38
2.4 Summary..... 39

3- Related Work 40
3.1 Memory-Based Recommender Systems..... 41
3.1.1 Collaborative Filtering..... 41
3.1.2 Content-Based Systems..... 42
3.1.3 Hybrid Recommender Systems..... 43
3.2 Next Page Prediction..... 45
3.2.1 Association Rules..... 45
3.2.2 Clustering. 46
3.2.3 Sequential Patterns..... 47
3.2.4 Markov Models. 49
3.3 Semantic Data Integration..... 50
3.3.1 Ontology Construction and Modification..... 50
3.3.2 Semantic Web Usage Mining..... 51
3.3.3 Semantic Data in Recommender Systems..... 55
3.3.4 Improving Search Results..... 57
3.4 Summary..... 57

4- Hybrid Framework for Next Page Prediction 59
4.1 The Proposed Framework..... 60
4.2 Frequent Patterns Mining..... 61
4.2.1 Data Preparation..... 61
A. Data Cleaning..... 61

B. Session and Transaction Identification..... 62
4.2.2 Pattern Mining..... 64
4.3 The User Page Matrix..... 64
4.4 User Clustering and Centroid Pruning..... 65
4.5 Frequent Patterns Clustering..... 66
4.6 Next Page Prediction..... 67
4.6.1 Conventional Next Page Prediction..... 67
4.6.2 Prediction Based on Usage Data..... 68
4.7 Results and Discussion..... 69
4.7.1 Datasets..... 69
4.7.2 Experiments' Settings..... 69
4.7.3 Prediction Speed..... 72
A. Prediction Time..... 72
B. Patterns / Query..... 74
4.7.4 Prediction Accuracy..... 75
4.7.5 Other Prediction Quality Measures..... 77
4.8 Summary..... 82

5- Semantic Data for Improved Prediction 83

5.1 Integrating Semantic Information..... 84
5.1.1 Semantic Annotation..... 84
5.1.2 User Concept Matrix..... 86
5.1.3 Decision Fusion..... 86
A. Combined User-Concept Matrix..... 87
B. Combined Distance Matrix..... 88
5.2 User and Pattern Clustering..... 89
5.3 Next Page Prediction..... 91
5.4 Results and Discussion..... 93
5.4.1 Prediction Speed..... 93
A. User Concept Matrix..... 93
B. Decision Fusion..... 94
C. Results Summary and Analysis..... 94
5.4.2 Prediction Accuracy..... 96
5.4.3 Other Prediction Quality Measures..... 98
5.5 Summary..... 102

6- Conclusions and Future Work **104**

6.1 Proposed Hybrid System..... 105
6.2 Adding Semantic Information..... 107
6.3 Recommendations and Future Work..... 109
6.3.1 System Recommendations..... 109
6.3.2 Future Work..... 109

Publications…....…….. 110
References…….. 111

Arabic Summary

List of Figures

Fig. 2.1	Taxonomy of Recommender Systems.....	8
Fig. 2.2	Switching Hybrid Recommender Systems.....	18
Fig. 2.3	Cascade Hybrid Recommender Systems.....	19
Fig. 2.4	Weighted Hybrid Recommender Systems.....	20
Fig. 2.5	Architecture of Web Usage Mining.....	21
Fig. 2.6	An Example of Web Server Logs.....	22
Fig. 2.7	An Example of 1st and 2nd Order Markov Models.....	26
Fig. 2.8	Evolution of the Web.....	30
Fig. 2.9	Architecture of Semantic Web.....	31
Fig. 2.10	Example for using RDF.....	32
Fig. 2.11	Example for Using an RDF Schema.....	33
Fig. 2.12	Semantic Annotation Process.....	38
Fig. 3.1	Using Data Mining to Build Ontologies.....	51
Fig. 3.2	The Knowledge and Data Co-Evolution Cycle.....	52
Fig. 3.3	Integrating Semantic Data into Usage Mining.....	53
Fig. 3.4	Architecture of the ORGAN System.....	55
Fig. 4.1	Proposed Framework.....	61
Fig. 4.2	Example of a User Path.....	63
Fig. 4.3	Algorithm for Generating MFRs.....	63
Fig. 4.4	Steps for Building User-Page Matrix.....	65
Fig. 4.5	Prediction Based on Usage Data.....	68
Fig. 4.6	Silhouette Coefficient Values.....	70
Fig. 4.7	Prediction Time (ms.) Comparison for User-Page Matrix...	73
Fig. 4.8	Patterns / Query Comparison for User-Page Matrix.....	75
Fig. 4.9	Prediction Accuracy Comparisons for User-Page Matrix....	77
Fig. 4.10	Comparisons of Precision Values.....	79
Fig. 4.11	Comparisons of Coverage Values.....	80
Fig. 4.12	Comparisons of F1 Measure Values.....	81
Fig. 5.1	System Architecture using User-Concept Matrix.....	85
Fig. 5.2	Steps for Semantic Annotation.....	85
Fig. 5.3	Steps for Building User-Concept Matrix.....	86
Fig. 5.4	Steps for Building Combined User-Concept Matrix.....	87
Fig. 5.5	Steps for Building Combined Distance Matrix.....	89

Fig. 5.6 Centroid Pruning with Concept Data..... 90
Fig. 5.7 Centroid Pruning with Combined Distance Matrix..... 91
Fig. 5.8 Prediction Based on Semantic Data..... 92
Fig. 5.9 Prediction Time (ms.) Summary Results..... 95
Fig. 5.10 Patterns / Query Summary Results..... 96
Fig. 5.11 Prediction Accuracy Comparisons..... 97
Fig. 5.12 Comparisons of Precision Values..... 99
Fig. 5.13 Comparisons of Coverage Values..... 100
Fig. 5.14 Comparisons of F1 Measure Values..... 101

List of Tables

Table 2.1	Comparison of Memory-based and Model-based Systems.....	28
Table 2.2	Web Mining Applications for the Semantic Web.....	36
Table 4.1	Datasets Description.....	69
Table 4.2	Prediction Time (ms.) Comparison for User-Page Matrix.....	72
Table 4.3	Patterns / Query Comparison for User-Page Matrix.....	74
Table 4.4	Prediction Accuracy Comparison for User-Page Matrix.....	76
Table 4.5	Comparison of Prediction Quality Measures.....	78
Table 5.1	Prediction Speed Comparison for User-Concept Matrix.....	93
Table 5.2	Prediction Speed Comparison for Decision Fusion.....	94
Table 5.3	Prediction Speed Summary results.....	95
Table 5.4	Prediction Accuracy in User-Concept Matrix & Decision Fusion.....	96
Table 5.5	Comparison of Precision Values.....	97
Table 5.6	Comparison of Coverage Values.....	98
Table 5.7	Comparison of F1 Measure Values.....	99
Table 5.8	Summary Values for Prediction Quality Measures.....	100
Table 6.1	Results for experiment (1).....	106
Table 6.2	Results for experiment (2).....	108

List of Abbreviations

DFS	Depth First Search
GSP	Generalized Sequential Patterns
LCS	Longest Common Subsequence
MAE	Mean Absolute Error
MFR	Maximal Forward Reference
OWL	Web Ontology Language
PNN	Pair-wise Nearest Neighbor
RDF	Resource Description Framework
RMSE	Root Mean Square Error
SVD	Singular Value Decomposition
URI	Universal Resource Identifier
URL	Uniform Resource Locator
VLMC	Variable Length Markov Chains
XML	Extensible Markup Language