University of Southern Queensland

Combining Web Data Mining Techniques for
Web Page Access Prediction

A Dissertation submitted by
Faten Khalil M.IT.
For the award of
Doctor of Philosophy

2008
Abstract

Web page access prediction gained its importance from the ever increasing number of e-commerce Web information systems and e-businesses. Web page prediction, that involves personalising the Web users’ browsing experiences, assists Web masters in the improvement of the Web site structure and helps Web users in navigating the site and accessing the information they need. The most widely used approach for this purpose is the pattern discovery process of Web usage mining that entails many techniques like Markov model, association rules and clustering. Implementing pattern discovery techniques as such helps predict the next page to be accessed by the Web user based on the user’s previous browsing patterns. However, each of the aforementioned techniques has its own limitations, especially when it comes to accuracy and space complexity. This dissertation achieves better accuracy as well as less state space complexity and rules generated by performing the following combinations. First, we combine low-order Markov model and association rules. Markov model analysis are performed on the data sets. If the Markov model prediction results in a tie or no state, association rules are used for prediction. The outcome of this integration is better accuracy, less Markov model state space complexity and less number of generated rules than using each of the methods individually. Second, we integrate low-order Markov model and clustering. The data sets are clustered and Markov model analysis are performed on each cluster instead of the whole data sets. The outcome of the integration is better
accuracy than the first combination with less state space complexity than higher order Markov model. The last integration model involves combining all three techniques together: clustering, association rules and low-order Markov model. The data sets are clustered and Markov model analysis are performed on each cluster. If the Markov model prediction results in close accuracies for the same item, association rules are used for prediction. This integration model achieves better Web page access prediction accuracy, less Markov model state space complexity and less number of rules generated than the previous two models.
Certification of Dissertation

I certify that the work reported in this thesis is entirely my own effort, except where otherwise acknowledged. This report is original and contains no material offered for the award of any other academic qualification at this or any other institution, or material previously published, except where due reference is made in the text.

_________________________  ______________________
Signature of Candidate          Date

Student Number: 50042419
Acknowledgements

I heartily thank my former principal supervisor Dr Jiuyong Li for his research direction of this project, his support, patience and guidance. Dr Li, provided me with clear direction and encouragement, thoughtful and constructive feedback on my work in a timely manner, as well as assistance and advice when needed even after he left USQ. I sincerely thank my current supervisor, Dr Hua Wang, for his continuous guidance, encouragement, valuable insights and support. Dr Wang was always there for me when I needed assistance and advice both as my associate supervisor as well as my principal supervisor later on. My gratitude also goes to the Department of Mathematics and Computing, University of Southern Queensland (USQ) for the financial support provided for attending conferences. I am extremely grateful to the Head of Department Professor Ron Addie and his secretary Ms Helen Nkansah. Lastly, but not least, I greatly appreciate the support from my parents and my family. Their care and love are what keeps me going during hard times.
Publications

The publications arising from this dissertation are as follows:


Contents

1 Introduction

1.1 Research Objectives ............................................. 2
1.2 Dissertation Structure ........................................ 4
1.3 Applications ..................................................... 5

2 Background

2.1 Introduction .................................................... 9
2.2 Web Data Mining ................................................ 11
2.2.1 Web Structure Mining ...................................... 11
2.2.2 Web Content Mining ....................................... 12
2.2.3 Web Usage Mining .......................................... 13
2.2.3.1 Preprocessing ............................................. 13
2.2.3.2 Pattern Discovery ....................................... 15
2.3 Web Usage Mining Techniques and Web Page Prediction .......... 18
3 Prediction Techniques

3.1 Introduction ................................................. 21
3.2 Markov Model ................................................. 23
3.3 Association Rules ............................................. 27
3.4 Clustering ..................................................... 31
3.5 Conclusion ..................................................... 41

4 Integrating Markov Model with Association Rules 43

4.1 Introduction .................................................. 43
4.2 Markov Model ................................................. 44
   4.2.1 Limitations of Markov Models ......................... 45
   4.2.2 Markov Model State Space Complexity .................. 47
   4.2.3 Using Markov Model Order for Prediction ............... 49
4.3 Association Rules ............................................. 50
   4.3.1 Limitations of Association Rules ........................ 51
   4.3.2 Using Association Rules for Prediction ................. 55
   4.3.3 Error Estimation of Association Rules Based Prediction 56
4.4 Integration Process ........................................... 57
   4.4.1 Motivation for Integration .............................. 57
   4.4.2 Integration Algorithm .................................... 62
5.3.2.2 Session Categorisation .......................... 101
5.3.2.3 $k$-means Distance Measures ...................... 104
5.3.2.4 Number of Clusters ($k$) ......................... 110
5.3.2.5 Markov Model Implementation ...................... 112
5.3.2.6 Item-Cluster Proximity .......................... 113
5.3.3 Integration Example ................................. 114
5.3.4 IMC Algorithm Efficiency Analysis .................. 115
5.3.4.1 Clustering Complexity ......................... 115
5.3.4.2 Prediction Complexity .......................... 116
5.4 Experimental Evaluation ............................. 116
5.4.1 Data Collection and Preprocessing .................. 116
5.4.2 Number of Clusters ($k$) .......................... 118
5.4.3 Distance Measures Evaluation ...................... 118
5.4.4 Experiments Results ............................... 122
5.4.5 Comparing IMC, Clustering and MM Accuracy ...... 128
5.4.6 Comparing IMC To a Higher Order Markov Model .. 130
5.4.6.1 Comparing State Space Complexity ............... 130
5.4.6.2 Comparing Accuracy ............................. 131
5.4.7 IMC Complexity ................................. 132
5.5 Conclusion ........................................ 134
6 Integrating Markov Model with Association Rules and Clustering

6.1 Introduction .................................................. 135

6.2 Integration Process ........................................... 136

6.2.1 Motivation For Integration .................................. 136

6.2.2 IPM Algorithm ................................................. 138

6.2.2.1 Algorithm Training process ................................. 139

6.2.2.2 Algorithm Prediction Process ................................. 143

6.2.3 Example ........................................................ 144

6.2.4 IPM Algorithm Efficiency Analysis .......................... 147

6.3 Experimental Evaluation ................................. 148

6.3.1 Clustering, Markov Model and Association Rules ............. 148

6.3.2 Experiments Results ........................................... 149

6.3.3 Comparing All Models Accuracy Results .................... 152

6.3.4 Comparing Results to a Higher Order Markov Model .......... 154

6.3.4.1 State Space Complexity Comparison ...................... 154

6.3.4.2 Accuracy Comparison ..................................... 157

6.4 Conclusion ..................................................... 157

7 Conclusions ....................................................... 159

7.1 General Discussions ........................................... 159
7.2 Conclusion of Results ........................................ 160

7.3 Strengths of Findings ........................................ 161

7.4 Limitations and Future Directions ........................ 162
# List of Figures

1.1 Dissertation structure. .................................................. 5

2.1 Web data mining architecture. .......................................... 10

2.2 Web usage mining architecture. ....................................... 18

4.1 Accuracy of all 1-, 2-, 3- and 4- frequency pruned Markov model orders. ........................................................ 50

4.2 The Integrated Markov and Association Model (IMAM) architecture. ................................................................. 61

4.3 Online computer store Web page structure. ....................... 70

4.4 Example Web log. .......................................................... 74

4.5 Frequency chart for the most frequent visited pages. .......... 76

4.6 Accuracy of 1\textsuperscript{st}, 2\textsuperscript{nd}, 3\textsuperscript{rd} and 4\textsuperscript{th} order Markov models and all 1\textsuperscript{st}, 2\textsuperscript{nd}, 3\textsuperscript{rd} and 4\textsuperscript{th} order frequency pruned Markov models for data set D1. .......................................................... 77
4.7 Accuracy of 1st, 2nd, 3rd and 4th order Markov models and all 1st,
2nd, 3rd and 4th order frequency pruned Markov models for data set D2. ..................................... 78

4.8 Accuracy of 1st, 2nd, 3rd and 4th order Markov models and all 1st,
2nd, 3rd and 4th order frequency pruned Markov models for data set D3. ..................................... 79

4.9 Accuracy of 1st, 2nd, 3rd and 4th order Markov models and all 1st,
2nd, 3rd and 4th order frequency pruned Markov models for data set D4. ..................................... 80

4.10 Number of rules generated according to different support thresh-
old values and a fixed confidence factor: 90%. ..................................... 80

4.11 No. of rules generated according to a fixed support threshold: 4%. 81

4.12 Time complexity in seconds for different support value. ......... 81

4.13 Portion of association rules results. ................................. 82

4.14 Accuracy of Association rules (AR), Frequency Pruned all 2nd
order Markov model (PMM) and IMAM model for data set D1. .. 84

4.15 Accuracy of Association rules (AR), Frequency Pruned all 2nd
order Markov model (PMM) and IMAM model for data set D2. .. 85

4.16 Accuracy of Association rules (AR), Frequency Pruned all 2nd
order Markov model (PMM) and IMAM model for data set D3. .. 86

4.17 Accuracy of Association rules (AR), Frequency Pruned all 2nd
order Markov model (PMM) and IMAM model for data set D4. .. 87
4.18 Accuracy of 3rd order Markov model (3-MM), frequency pruned all 3rd order Markov model (3-PMM) and IMAM model for all four data sets. .................................................... 88

5.1 The stages of clustering before Markov model implementation. .......... 95
5.2 The integration model (IMC) architecture. ................................. 98
5.3 ISODATA improves the k-means clusters. ................................. 111
5.4 Silhouette value of D1 with 7 clusters. ........................................ 119
5.5 Silhouette value of D2 with 9 clusters. ........................................ 119
5.6 Silhouette value D3 with 14 clusters. ......................................... 120
5.7 Silhouette value of D4 with 10 clusters. ...................................... 120
5.8 Silhouette value of Euclidean distance measure with 7 clusters. .... 122
5.9 Silhouette value of Hamming distance measure with 7 clusters. .... 123
5.10 Silhouette value of City Block distance measure with 7 clusters. ... 123
5.11 Silhouette value of Correlation distance measure with 7 clusters. ... 124
5.12 Silhouette value of Cosine distance measure with 7 clusters. ....... 124
5.13 The mean value for 2...10 clusters using different distance mea-
sures. .................................................................................. 124
5.14 Flowchart illustrating prediction accuracy calculation process. .... 127
5.15 Accuracy of clustering, Markov model of whole data set and Markov model accuracy using clusters based on Euclidean, Correlation and Cosine distance measures with $k = 7$ for data set D1. ................................. 128

5.16 Accuracy of clustering, PMM and IMC for data set D1. ................. 129

5.17 Accuracy of clustering, PMM and IMC for data set D2. .................. 130

5.18 Accuracy of clustering, PMM and IMC for data set D3. .................. 130

5.19 Accuracy of clustering, PMM and IMC for data set D4. .................. 131

5.20 Accuracy of 3rd order Markov model (3-MM), frequency pruned all 3rd order Markov model (3-PMM) and IMC model for all four data sets. ......................................................... 132

5.21 Running time of clusters for all four data sets. ............................... 133

5.22 Prediction time of IMC model for all four data sets. ........................ 133

6.1 IPM model architecture. ................................................................. 139

6.2 Accuracy of Clustering, AR, PMM, and IPM for data set D1. .......... 150

6.3 Accuracy of Clustering, AR, PMM, and IPM for data set D2. .......... 151

6.4 Accuracy of Clustering, AR, PMM, and IPM for data set D3. .......... 151

6.5 Accuracy of Clustering, AR, PMM, and IPM for data set D4. .......... 152

6.6 Accuracy of Clustering, AR, PMM, IMAM, IMC and IPM for data set D1. ................................................................. 153
6.7 Accuracy of Clustering, AR, PMM, IMAM, IMC and IPM for data set D2. .................................................. 154

6.8 Accuracy of Clustering, AR, PMM, IMAM, IMC and IPM for data set D3. .................................................. 155

6.9 Accuracy of Clustering, AR, PMM, IMAM, IMC and IPM for data set D4. .................................................. 156

6.10 Accuracy of Clustering, AR, PMM, IMAM, IMC and IPM for all four data sets. ........................................ 157

6.11 Accuracy of 3-MM and 3-PMM compared to that of IMAM, IMC and IPM for all four data sets. .................. 158
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Number of states of all 1- to 4- Markov model orders</td>
<td>48</td>
</tr>
<tr>
<td>4.2</td>
<td>Number of states of frequency pruned Markov model orders</td>
<td>49</td>
</tr>
<tr>
<td>4.3</td>
<td>Example: Four Web transactions</td>
<td>51</td>
</tr>
<tr>
<td>4.4</td>
<td>User sessions</td>
<td>69</td>
</tr>
<tr>
<td>4.5</td>
<td>Pageviews frequencies</td>
<td>70</td>
</tr>
<tr>
<td>4.6</td>
<td>User sessions after frequency and support pruning</td>
<td>70</td>
</tr>
<tr>
<td>4.7</td>
<td>User sessions history</td>
<td>71</td>
</tr>
<tr>
<td>4.8</td>
<td>Confidence of accessing page M using subsequence association rules</td>
<td>72</td>
</tr>
<tr>
<td>4.9</td>
<td>Confidence of accessing page N using subsequence association rules</td>
<td>72</td>
</tr>
<tr>
<td>4.10</td>
<td>Sessions</td>
<td>76</td>
</tr>
<tr>
<td>4.11</td>
<td>IMAM number of states</td>
<td>86</td>
</tr>
<tr>
<td>5.1</td>
<td>Example: initial Web sessions</td>
<td>103</td>
</tr>
<tr>
<td>5.2</td>
<td>Example: Preprocessed Web sessions</td>
<td>103</td>
</tr>
</tbody>
</table>
5.3 Web sessions after categorisation

5.4 Sessions

5.5 Sessions distances

5.6 Example of user sessions

5.7 The first cluster

5.8 The second cluster

5.9 Number of categories

5.10 Session categorisation

5.11 Entropy measures for different clusters

5.12 Web sessions grouped into 7 clusters

5.13 IMC number of states

6.1 Accuracy according to $z_{\alpha/2}$ value

6.2 User sessions

6.3 First cluster

6.4 Second cluster

6.5 User sessions history

6.6 Confidence of accessing page E using subsequence association rules

6.7 Confidence of accessing page G using subsequence association rules

6.8 Prediction accuracy using all models for all four data sets
6.9 Number of states for 3-PMM, IMAM, IMC and IPM and 3-MM using D1, D2, D3 and D4 . . . . . . . . . . . . . . . . . . . . . . . . . . . 155

7.1 Accuracy values standard deviation . . . . . . . . . . . . . . . . . . . 162