MST(EC)
04
HO

# GA-BASED COLLABORATIVE FILTERING FOR ONLINE RECOMMENDATION

By

**HO YI FONG, Yvonne**

**Master of Science in E-Commerce Technology**

**2007**




**Faculty of Science and Technology**

**University of Macau**

# Table of Contents

**List of Figures** ..... x

**List of Tables** ..... xii

**List of Charts** ..... xiii

**List of Abbreviations** ..... xiv

**Chapter 1 Introduction** ..... 1

1.1 Motivation and Objectives ..... 1

1.2 Background ..... 2

1.3 Advantages ..... 3

*1.3.1 Make the visitor becomes buyer* ..... 3

*1.3.2 Increase Cross-sell* ..... 4

*1.3.3 Improve Loyalty* ..... 4

*1.3.4 Improve efficiency of finding interested items* ..... 5

*1.3.5 Sharing with friends* ..... 5

*1.3.6 Knowing other customers' choice* ..... 5

1.4 Current Recommender Applications ..... 6

*1.4.1 MovieLens* ..... 6

1.4.1.1 Overview ..... 6

1.4.1.2 Functions ..... 7

*1.4.2 Amazon.com* ..... 10

1.4.2.1 Overview ..... 10

1.4.2.2 Functions ..... 11

*1.4.3 ChoiceStream* ..... 13

1.4.3.1 Overview ..... 14

1.4.3.2 Functions ..... 14

*1.4.4 Google News* ..... 17

1.4.4.1 Overview ..... 18

1.4.4.2 Functions ..... 18

1.5 Thesis Organization ..... 20

**Chapter 2 Overview of Recommendation Technology** ..... 21

2.1 Background ..... 21

2.2 Methods ..... 21

*2.2.1 Content-based Filtering* ..... *22*
*2.2.2 Collaborative Filtering (CF)* ..... *23*
*2.2.3 Hybrid Model* ..... *25*

**Chapter 3 Problems Definition** ..... **26**

3.1 Limited Content Analysis ..... 26
3.2 Over-specialization ..... 26
3.3 Limited Items ..... 27
3.4 Too many information ..... 27
3.5 Redundant features ..... 27
3.6 Inaccurate Prediction ..... 27
3.7 Model Build ..... 28

**Chapter 4 Background of GA and Related Techniques** ..... **29**

4.1 Genetic Algorithm ..... 29
*4.1.1 Chromosomes* ..... *30*
*4.1.2 Reproduction* ..... *31*
4.1.2.1 Selection ..... 31
4.1.2.2 Crossover ..... 31
4.1.2.3 Mutation ..... 32
*4.1.3 Fitness Function* ..... *33*
*4.1.4 Simple GA* ..... *33*
4.2 Feature Selection ..... 34
4.3 Feature Weight ..... 35
*4.3.1 Repair Algorithm* ..... *37*
4.4 Pearson Coefficient ..... 37
*4.4.1 Fitness measure* ..... *38*

**Chapter 5 Literature Survey** ..... **40**

5.1 Content-based Filtering ..... 40
5.2 Collaborative Filtering ..... 42
5.3 Hybrid Model ..... 46

**Chapter 6 Design of GA-based Recommendation System** ..... **50**

6.1 Pre-process Data ..... 50
*6.1.1 User database* ..... *50*
*6.1.2 Movie database* ..... *50*
6.2 System architecture ..... 51

*6.2.1 Phase I - Collect User Information* ..... *52*
6.2.1.1 Profile Capture ..... 52
6.2.1.2 Collect Movie Ratings ..... 53
*6.2.2 Phase II - Create User Profile* ..... *54*
6.2.2.1 Collect Movie Info ..... 54
6.2.2.2 Create New User Profile features ..... 55
*6.2.3 Phase III - GA Recommender* ..... *58*
6.2.3.1 Feature Selection ..... 58
6.2.3.2 Feature Weighting ..... 59
6.2.3.3 Similarity Measure ..... 59
6.2.3.4 Repair Algorithm ..... 62
6.2.3.5 Create Neighbor Set ..... 62
6.2.3.6 Recommendations ..... 63
6.3 Fitness Measure ..... 64

**Chapter 7 Experiments ..... 65**

7.1 Experimental Setup ..... 65
*7.1.1 Data preparation* ..... *65*
7.1.1.1 MovieLens ..... 65
7.1.1.2 IMDb ..... 66
*7.1.2 Pearson Coefficient* ..... *67*
*7.1.3 GA Parameter* ..... *68*
7.2 Experiment Evaluation ..... 68
*7.2.1 Experiment I - Features Weights* ..... *68*
*7.2.2 Experiment II - Fitness Accuracy* ..... *68*
7.2.2.1 Using Pearson Algorithm (Pearson) ..... 69
7.2.2.2 Using Genetic Algorithm with 22 features (GA) ..... 69
7.2.2.3 Using Genetic Algorithm with 29 features (GA Merge) ..... 69
7.2.2.4 Using Genetic Algorithm with 37 features (GA User Profile 37) ..... 70
7.2.2.5 Using Genetic Algorithm with 12 features (GA User Profile 12) ..... 71
7.2.2.6 Using Genetic Algorithm with 7 features (GA User Profile 7) ..... 71
*7.2.3 Experiment III - Performance on Neighbor Set* ..... *72*

**Chapter 8 Results of Experiments ..... 73**

8.1 Experiment I - Feature Weights ..... 73
8.2 Experiment II - Fitness Accuracy ..... 74
*8.2.1 Fitness Accuracy for fixed 50 users* ..... *74*
*8.2.2 Fitness Accuracy for random 50 users* ..... *75*
*8.2.3 Process time for fixed 50 users* ..... *76*

*8.2.4 Process time for random 50 users* ..... 77
8.3 Experiment III – Performance on Neighbor Set ..... 78
*8.3.1 Process time vs. Neighbor Set* ..... 78
*8.3.2 Fitness vs. Neighbor Set* ..... 79

**Chapter 9 Analysis of Results** ..... 80

**Chapter 10 Conclusion** ..... 82

**References** ..... 83

**Appendix** ..... 88

Appendix A - MovieLens Database Structure ..... 88
Appendix B - IMDb Database Structure ..... 89
Appendix C1 - Feature Weight Table ..... 90
Appendix C2 - Fitness for 50 Fixed Users Table ..... 91
Appendix C3 - Fitness for 50 Random Users table ..... 92
Appendix C4 - Process Time for 50 Fixed Users Table ..... 93
Appendix C5 - Process Time for 50 Random Users Table ..... 94
Appendix C6 - Performance vs. Neighbor Set ..... 95
Appendix C7 - Fitness vs. Neighbor Set ..... 96
Appendix D - Sample of Result generated by GA Recommender ..... 97
Appendix E - Sample of Result generated by COFE ..... 100