

Evaluating Collaborative Filtering Over Time

Neal Lathia

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of the
University of London.

Department of Computer Science
University College London

September 9, 2010

**To my parents
and their fervent passion for education**

Abstract

Recommender systems have become essential tools for users to navigate the plethora of content in the online world. Collaborative filtering—a broad term referring to the use of a variety, or combination, of machine learning algorithms operating on user ratings—lies at the heart of recommender systems’ success. These algorithms have been traditionally studied from the point of view of how well they can *predict* users’ ratings and how *precisely* they rank content; state of the art approaches are continuously improved in these respects. However, a rift has grown between how filtering algorithms are investigated and how they will operate when deployed in real systems. Deployed systems will continuously be queried for personalised recommendations; in practice, this implies that system administrators will iteratively retrain their algorithms in order to include the latest ratings. Collaborative filtering research does not take this into account: algorithms are improved and compared to each other from a *static* viewpoint, while they will be ultimately deployed in a *dynamic* setting. Given this scenario, two new problems emerge: current filtering algorithms are neither (a) designed nor (b) evaluated as algorithms that must account for time. This thesis addresses the divergence between research and practice by examining how collaborative filtering algorithms behave over time. Our contributions include:

1. A fine grained **analysis** of temporal changes in rating data and user/item similarity graphs that clearly demonstrates how recommender system data is dynamic and constantly changing.
2. A **novel methodology** and time-based **metrics** for evaluating collaborative filtering over time, both in terms of accuracy and the diversity of top- N recommendations.
3. A set of **hybrid algorithms** that improve collaborative filtering in a range of different scenarios. These include temporal-switching algorithms that aim to promote either accuracy or diversity; parameter update methods to improve temporal accuracy; and re-ranking a subset of users’ recommendations in order to increase diversity.
4. A set of **temporal monitors** that secure collaborative filtering from a wide range of different temporal attacks by flagging anomalous rating patterns.

We have implemented and extensively evaluated the above using large-scale sets of user ratings; we further discuss how this novel methodology provides insight into dimensions of recommender systems that were previously unexplored. We conclude that investigating collaborative filtering from a temporal perspective is not only more suitable to the context in which recommender systems are deployed, but also opens a number of future research opportunities.

Acknowledgements

Over the past years, I have been very lucky: I have been surrounded by brilliant, intelligent and inspiring people. They contributed to this thesis with their questions, insights, encouragement, and support; I am much indebted to them all. I will never be able to thank my supervisors, Steve Hailes and Licia Capra, enough: being mentored by researchers of this calibre was often all the motivation I needed. Thanks to Cecilia Mascolo, who was the first to suggest that I apply for a PhD (would I be writing this had it not been for that suggestion?); Daniele Quercia, with his unrivalled and contagious passion for research (and blogging); and all of the members of the MobiSys group. Thanks to EPSRC Utiforo, for the financial support, and thanks to all the project partners for the colourful meetings. A special thanks to Torsten Ackemann: the experiments I ran over the past few years would still be running had it not been for his invaluable help with the department's Condor cluster.

A highlight of the recent years is the time I spent in Telefonica I+D's Multimedia Group in Barcelona. A big thanks to Xavier Amatriain, Josep M. Pujol and Jon Froehlich; I not only learned a lot during these summer months, but made some great friends and thoroughly enjoyed my time there. I hope to one day finally manage to go hiking with Jon.

While all those with whom I worked with deserve my utmost thanks, I am even more indebted to my family and friends, who were there to take my mind off of my PhD. Thanks to Paul, Usha, Fergal and Preeya; to Pavle and Justin (we await your return to London), and Viktor (who always turned up at my doorstep at the right time). Thanks to my sisters, Sheila and Anna (who has put up with living with me). A special thanks to Yasmin, who has always been there for me. Lastly, thanks to the bands I have been a part of over these years (The Hartes; Pavle, and The Jukebox Leans; Sean and Duncan), for allowing me to keep nurturing my love for music.

This thesis is dedicated to my parents.

Contents

1	Introduction	17
1.1	Motivating Information Filtering	18
1.2	Brief History of Recommender Systems	19
1.3	Problem Statement and Contributions	20
1.3.1	Timeliness of Research	21
1.4	Publications Related To This Thesis	22
1.5	Summary	23
2	Computing Recommendations With Collaborative Filtering	25
2.1	Ratings And User Profiles	25
2.1.1	Implicit and Explicit Ratings	26
2.2	Collaborative Filtering Algorithms	27
2.2.1	Grouping the Algorithms	28
2.2.2	Baselines	29
2.2.3	k-Nearest Neighbours	29
2.2.4	Matrix Factorisation	32
2.2.5	Hybrid Algorithms	33
2.2.6	Online Algorithms	35
2.2.7	From Prediction to Recommendation	35
2.3	Trust and User Modelling	36
2.3.1	Motivating Trust in Recommender Systems	36
2.3.2	Using Trust For Neighbour Selection	37
2.3.3	Trust-Based Collaborative Filtering	41
2.4	Evaluating Recommendations	42
2.4.1	Rating Datasets	42
2.4.2	Methodology	42
2.4.3	Metrics	43
2.5	Open Problems	45
2.5.1	Ratings: Changing Over Time	45
2.5.2	Methodology & Evaluation	46

2.5.3	System Robustness	47
2.6	Summary	47
3	Temporal Analysis of Rating Datasets	49
3.1	Rating Datasets	49
3.2	Ratings Over Time	50
3.2.1	Dataset Growth	50
3.2.2	Changing Summary Statistics	54
3.2.3	Temporal User Behaviour	57
3.2.4	Daily and Weekly Trends	58
3.3	Similarity Over Time	59
3.3.1	Similarity Measures	59
3.3.2	Static Similarity	60
3.3.3	Temporal Similarity	63
3.4	Summary	72
4	Temporal Accuracy of Collaborative Filtering	75
4.1	Measuring Temporal Performance	75
4.1.1	Simulating Temporal Updates	75
4.1.2	Metrics: Sequential, Continuous, Windowed	76
4.1.3	Case Study	76
4.1.4	Methodology	80
4.2	Results	82
4.2.1	Sequential Results	82
4.2.2	Time-Averaged Results	83
4.2.3	Discussion	84
4.3	Adaptive Temporal Collaborative Filtering	85
4.3.1	Adaptive CF	85
4.3.2	Adaptive kNN	86
4.3.3	Adaptive SVD	89
4.4	Related Work	89
4.5	Summary	90
5	Temporal Diversity in Recommender Systems	93
5.1	Why Temporal Diversity?	93
5.1.1	Changes Over Time	93
5.1.2	User Survey	94
5.2	Evaluating for Diversity	98
5.2.1	From Predictions to Rankings	98
5.2.2	Methodology	98

5.2.3	Measuring Diversity Over Time	99
5.2.4	Results and Analysis	101
5.2.5	Diversity vs. Profile Size	101
5.2.6	Diversity vs. Ratings Input	103
5.2.7	Diversity and Time Between Sessions	103
5.2.8	Lessons Learned	104
5.3	Promoting Temporal Diversity	104
5.3.1	Temporal Switching	104
5.3.2	Temporal User-Based Switching	105
5.3.3	Re-Ranking Frequent Visitors' Lists	106
5.4	Discussion	107
5.5	Summary	108
6	Temporal Defences for Robust Recommendations	109
6.1	Problem Setting	109
6.2	Defeating Non-Temporal Attacks	110
6.3	Temporal Attack Models	112
6.3.1	Measuring Attacks	113
6.4	A Temporal Defence	114
6.4.1	Global Thresholding	114
6.4.2	User Monitoring	116
6.4.3	Item Monitoring	118
6.5	Adaptive Attack Models	120
6.5.1	The Ramp-Up Attack	121
6.6	Discussion & Related Work	122
6.7	Summary	124
7	Conclusion	125
7.1	Thesis Contributions	125
7.2	Future Work	126
7.2.1	Using a Temporal Methodology	127
7.2.2	Beyond Temporal Collaborative Filtering	128
	Appendices	129
A	Diversity Surveys	131
A.1	Pre-Survey Instructions and Demographics	131
A.2	Movie Recommendations	132
A.2.1	Recommendation Structure	133
A.2.2	Survey 1: No Diversity	134

A.2.3	Survey 2: Diversified Popular Movies	134
A.2.4	Survey 3: Diversified Random Movies	135
A.3	Post-Survey Questions	136
Bibliography		136

List of Figures

3.1	Number of Users Over Time (ML-1, ML-2, Netflix)	51
3.2	Number of Movies Over Time (ML-1, ML-2, Netflix)	51
3.3	Number of Total Ratings Over Time (ML-1, ML-2, Netflix)	51
3.4	Non-Cumulative Netflix Daily Growth: the spikes represent days when a lot of users/movies/ratings were added	52
3.5	Non-Cumulative ML-1 Daily Growth	52
3.6	Sparsity Over Time For Each Dataset: Netflix is the most sparse dataset	53
3.7	Rating Distribution Over Time Of Each Dataset: Netflix is the only dataset with no consistent ordering between the rating values	54
3.8	Datasets' Global Rating Mean Over Time, Again highlighting the stop in ML-2's growth	54
3.9	Datasets' Global Rating Variance Over Time	55
3.10	Netflix Rating Median and Mode Over Time	55
3.11	Users Binned By Profile Size Over Time	56
3.12	Average User and Item Mean Rating Over Time	57
3.13	Standard Deviation of Ratings Per User Per Day	57
3.14	MovieLens: Average Number of Ratings Per Week (With Standard Deviation)	58
3.15	MovieLens: Average Number of Ratings Per Hour (With Standard Deviation)	59
3.16	ML-1 PCC, Weighted-PCC & Constrained-PCC Similarity Distribution	61
3.17	ML-1 Jaccard & Cosine Similarity Distribution	61
3.18	Similarity Between User 1 and 30: Similarity depends on how you measure it	65
3.19	Evolution of Similarity for the Jaccard, w PCC, Cosine and PCC Similarity Measures, Comparing User 1 to All Other Users in the System	66
3.20	ML-1 User 1: New Relationships Left Over Time	68
3.21	In-degree long tail of w PCC- k NN $k = 100$ ML-1 Graph	70
3.22	Results When Excluding or Exclusively Using Power Users	72
4.1	User 407: Three Views of Temporal Error	77
4.2	ML-1 Dataset: Three Views of Temporal Error	78
4.3	Temporal Experiments With a Static Test Set (User/Item Mean)	79
4.4	Temporal Experiments With a Static Test Set (kNN/SVD)	79

4.5	Temporal Experiment Test Sets' Characteristics: Size, and Distribution of Users Who Rate Items First and Items that Are Rated First	81
4.6	Sequential RMSE Results for User Bias Model and SVD	82
4.7	Sequential RMSE Results for k NN Algorithm With $k \in \{20, 50\}$	82
4.8	Time-Averaged RMSE for User Bias Model and SVD	83
4.9	Time-Averaged RMSE for k NN Algorithm and Users With Fewer Than 10 Ratings . . .	84
4.10	Time-Averaged RMSE Comparing $k = 50$, the Bias Model, and Adaptive CF; Proportions of Users Who Selected Each Algorithm Over Time, and Proportions of Users Who Changed Method At Each Interval	86
4.11	Time-Averaged RMSE Comparing $k = 50$ and Adaptive ($k = \alpha$) k NN, Proportions of Users Who Selected Each k Value Over Time, and Proportions of Users whose k Value Changed At Each Interval	87
4.12	Time-Averaged RMSE Gain of Adaptive-SVD With Different Subsets of Parameters . .	88
4.13	Time-Averaged RMSE of k NN With Limited History	90
5.1	Survey Results for (S1) Popular Movies With No Diversity (S2) Popular Movies With Diversity and (S3) Randomly Selected Movies	95
5.2	Boxplots of Each Week's Ratings for the Three Surveys	96
5.3	Top-10 and 20 Temporal Diversity for Baseline, k NN and SVD CF	100
5.4	Top-10 and 20 Temporal Novelty for Baseline, k NN and SVD CF	100
5.5	Profile Size vs. Top-10 Temporal Diversity for Baseline, k NN and SVD CF	102
5.6	Ratings Added vs. Top-10 Temporal Diversity for Baseline, k NN and SVD CF	102
5.7	Time Passed vs. Top-10 Temporal Diversity for Baseline, k NN and SVD CF	102
5.8	Comparing Accuracy with Diversity	103
5.9	Diversity (a) and Accuracy (b) of Temporal Switching Method	105
5.10	Temporal Diversity and Accuracy vs. Diversity With User-Based Temporal Switching . .	106
5.11	Temporal Diversity and Accuracy vs. Diversity When Re-Ranking Frequent Visitors' Lists	106
6.1	Time-Averaged RMSE Of One-Shot Attack, and Prediction Shift When Pruning New-comer's Ratings, and Injecting Attacks Over Varying Time Windows	111
6.2	Attack Types and Impact With No Defences	113
6.3	Netflix Ratings Per User Per Week; Global Thresholding Precision and Recall	115
6.4	Global Thresholding Impact	116
6.5	Example Ratings Per User (1 Week), Proportion of Ratings Per High Volume Raters and High Volume Raters Over Time	117
6.6	User Monitor/Combined Impact Results, and Proportion of High Volume Raters Who Have Been In The Group for Varying Lengths of Time	118
6.7	Item Monitor: Average Precision & Recall	120

6.8 Example Ramp-Up Attack: How it Affects the Monitor’s Values, the Optimal Ratings
Per Sybil and Prediction Shift 121

A.1 Example User Instructions from Survey 1 131

A.2 Demographic Data Questions: Gender, Age, Average Movies Per Month, Familiarity
and Use of Recommender Systems 132

A.3 Example Screen Shot: Survey 1, Week 1 133

A.4 Example Screen Shot: Survey 1, Buffer Screen 1 133

A.5 Example Screen Shot: Survey 1, Final Questions 137

List of Tables

2.1	A Sample of Open Problems in Recommender Systems	45
3.1	Users, Items, Ratings in Each Dataset	50
3.2	MAE Prediction Error, MovieLens u1 Subset	62
3.3	MAE Prediction Error For All MovieLens Subsets	62
3.4	Average Unique Recommenders in Users' Neighbourhoods	67
3.5	w PCC- k NN Graph Properties	69
3.6	Unused Proportions of the Dataset	71
5.1	ANOVA P-Values and Pairwise T-Test Values For The 5 Weeks	96
A.1	S1 (All 5 Weeks): All Time Worldwide Box Office Ranking (December 2009)	134
A.2	S2 (Weeks 1, 2): Diversified All Time Worldwide Box Office Ranking	134
A.3	S2 (Weeks 3, 4, 5): Diversified All Time Worldwide Box Office Ranking	135
A.4	S3 (Weeks 1, 2, 3, 4, 5): Randomly Selected Movies	136

