

If You Liked 'X', Then You Will Love 'Y'

Ideas from The Netflix Prize Competition

Discussion Leader

Ed Ramsden

Cascade Systems Society

Harder House

November 5, 2010

Netflix Prize Competition

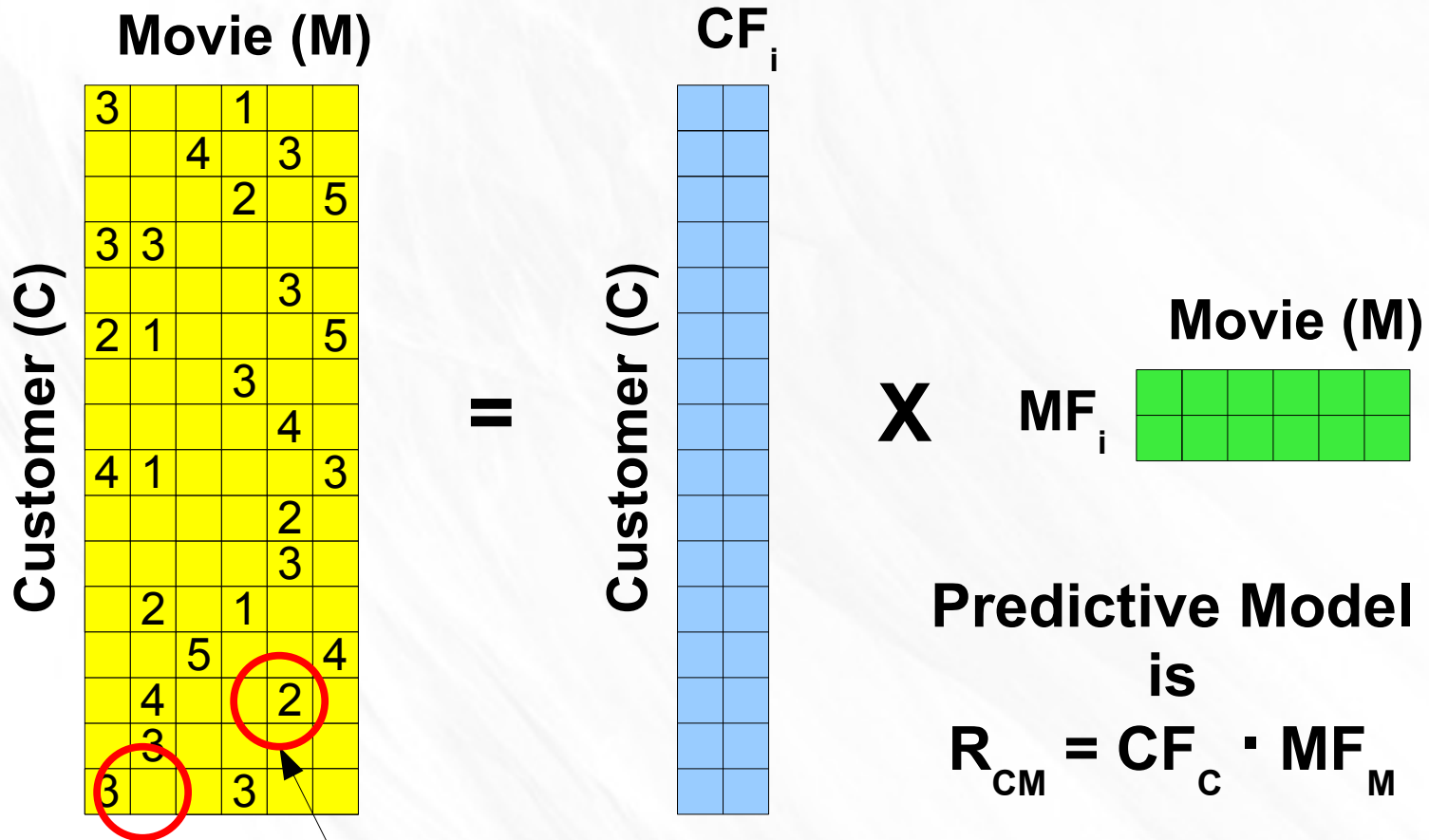
- \$1 Million prize to first team to build a model that predicted how a customer would rate previously unseen movies with an accuracy 10% better than Netflix's internal 'Cinematch' model.
- Contest Ran from October 2006 – July 2009.
- 40,000+ teams downloaded dataset, ~5000 submitted model results.
- Large-scale, publicly released dataset with REAL data!
 - ~480,000 customers
 - 17,770 Movies (with titles)
 - ~100,000,000 individual ratings (movie/customer/rating/date)
- Active Forum where people shared results and techniques relatively freely (considering it **was** a competition!)

Some Models & Their Performance...

- '3 stars': **1.35**
- Average Rating of Given Customer: **1.07**
- Average Rating for Given Movie: **1.05**
- User/Movie Scaling (Z-score) functions: **0.98-1.00**
- Cinematch (Netflix In-house Algorithm): **0.95**
- Knn ('nearest' neighbor analysis): **0.93-0.95**
- Basic Matrix Factorization (SVD, 20 factors): **0.91-0.93**
- Sophisticated Matrix Factorization (SVD+, 1000+ factors): **0.88**
- **WINNING MODEL/GOAL: 0.8567**
- Best Possible: **0.6 – 0.7 ???**

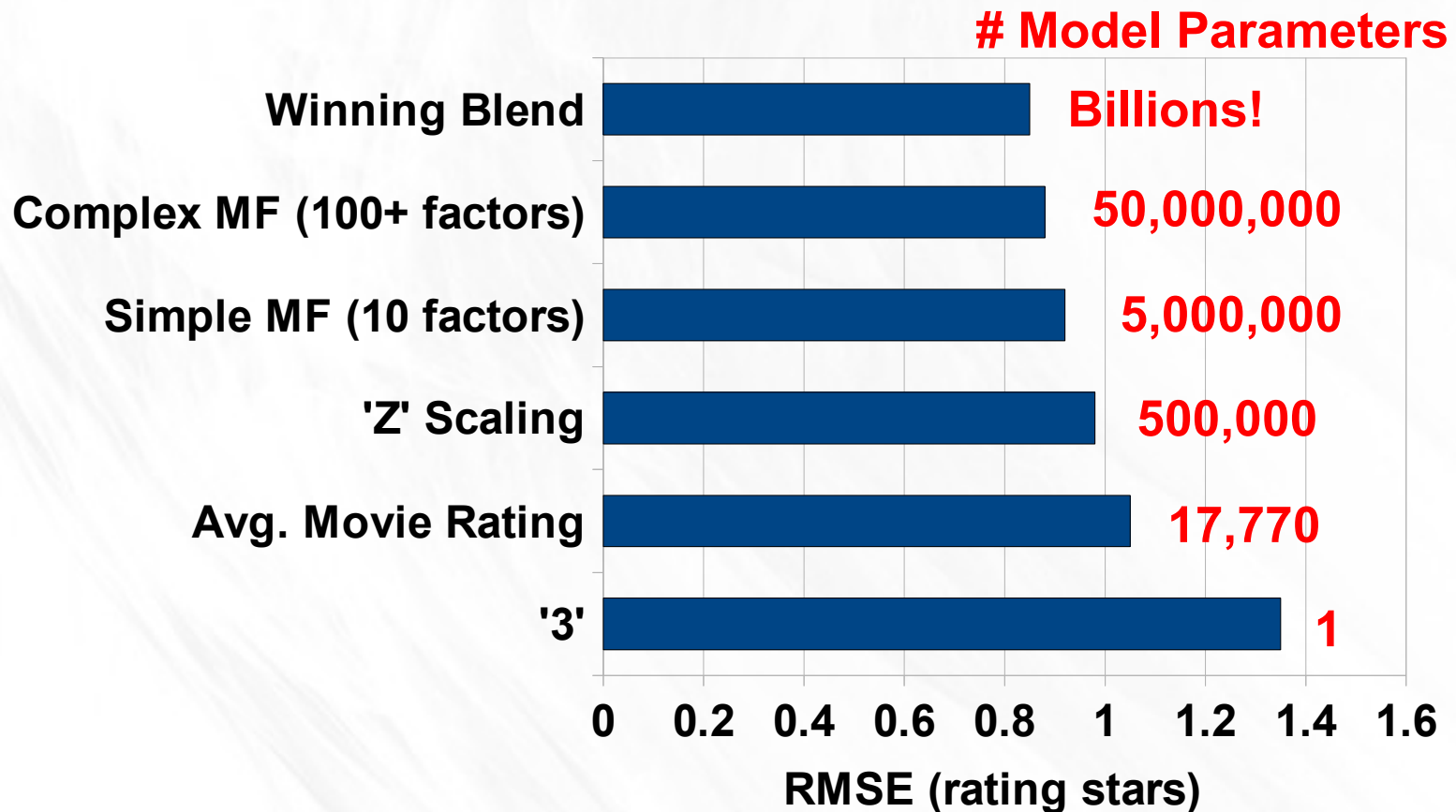
Performance as RMSE on a 1-5 stars scale over ~3 million predictions submitted to Netflix's 'Scoring Oracle'

Matrix Factorization Model (Sparse SVD)



Known Entries (~100 Million)
Unknown Entries (~8.4 Billion)

Diminishing Returns vs. Model Complexity



For noisy and poorly understood systems, predictive accuracy from sophisticated models may not be much better than that of simple models.

Data vs. Insight

Observations:

Additional movie data (IMDB) did not seem to help people get better scores.

'Designed' models (e.g. scaling schemes) seemed to run out of steam around 0.98-0.95.

Would CUSTOMER data have enabled better 'designed' models?

You can find out for yourself – the GroupLens Research project has movie ratings with user demographic data!
<http://www.grouplens.org>

Meaning of Parameters & Structure?

- Variables and relations in a system dynamics model correspond to actual or hypothesized entities.
- Regressors in a regression model correspond to 'measurable' quantities/states.
- What do the factors in a matrix factorization model mean?

Factorization Scale Example (scale extremes, 1 factor)

What do *Schindler's List* and *Sex and the City* have in common?
Battlefield Earth and *Soul Plane*?

1.4795	Sex and the City: Season 6: Part 1	Rating count 33830
1.4795	To Kill a Mockingbird	Rating count 64474
1.4751	Sex and the City: Season 6: Part 2	Rating count 28369
1.4708	Pride and Prejudice	Rating count 19813
1.452	Sex and the City: Season 5	Rating count 33677
1.4466	Sex and the City: Season 4	Rating count 36172
1.4429	Schindler's List	Rating count 101052
1.4331	Sex and the City: Season 3	Rating count 39439
1.4167	Anne of Green Gables	Rating count 8023
1.4118	Sex and the City: Season 2	Rating count 44176
-1.189	House of the Dead	Rating count 5583
-1.1696	Alone in the Dark	Rating count 9862
-1.1508	Dumb and Dumberer: When Harry Met Lloyd	Rating count 9705
-1.1471	The Real Cancun	Rating count 7870
-1.14	Gigli	Rating count 9949
-1.1358	Battlefield Earth	Rating count 9441
-1.1185	My Boss's Daughter	Rating count 18587
-1.1163	Fear Dot Com	Rating count 6826
-1.0908	Soul Plane	Rating count 16065
-1.0735	Child's Play 2: Chucky's Back	Rating count 6718

<http://www.netflixprize.com/community/viewtopic.php?id=1470>

Thread: *What do features mean?*, poster: *Lazy Monkey*, 2009-07-23 13:45:41

Ensemble Methods – The Wisdom of Crowds

If you have lots of good models, you can sometimes make a better model by combining them.

Ensembling is a useful technique for improving model accuracy – analogous to the Wisdom of Crowds idea.

Models need to be somewhat uncorrelated for this to work though!

The Leading Teams got their last ~4% improvement in accuracy from blending the results of *hundreds* of models!

How do You Combine Models?

Results of Simple Blending Methods

The two Netflix Prize Leaders (*BellKor's Pragmatic Chaos* and *The Ensemble*) made the results of over 1000 of their models available for a 'blending' challenge organized by the Australasian Data Mining Conference. Here are the results of some blending methods...

Method	RMSE
Worst Model	994.12
Average of All Models	892.77
Best Model	888.32
Average of Best 10 Models	884.36
Linear Regression on all Models	882.82
Best Combination of 3 Models	882.73
Best Combination of 10 Models	881.58
Linear NN based on all Models	879.10
Best (Team Optibrebs – unknown method)	877.91

AusDM 09 Analytic Challenge 'Ensembling'

<http://www.tiberius.biz/ausdm09/AusDM09EnsemblingChallenge.pdf>

Diffusion of Data to Model Parameters

Observations:

- Matrix Factorization methods do better than KNN or scaling methods.
- MF methods have a higher cross-diffusion of information from data to model parameters.

Hypothesis:

Models that allow the data to 'diffuse' throughout to all model parameters will have better performance than those where predictions are based on smaller subsets of the data.

Part of the success of regression and NN techniques?

Resources:

Netflix Prize Forum

- <http://www.netflixprize.com//community/>

Various Related Papers/Articles

- Y. Koren et al., *Matrix Factorization Techniques for Recommender Systems*, IEEE Computer, Aug 2009, p.47
- <http://www.customerfusion.co.uk/pdf/kddpaper.pdf>
- http://www.netflixprize.com/assets/ProgressPrize2008_BellKor.pdf
- http://www.netflixprize.com/assets/ProgressPrize2008_BigChaos.pdf
- http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf
- http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf
- http://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf
- <http://sifter.org/~simon/journal/20061211.html> (first proposal of MF methods)

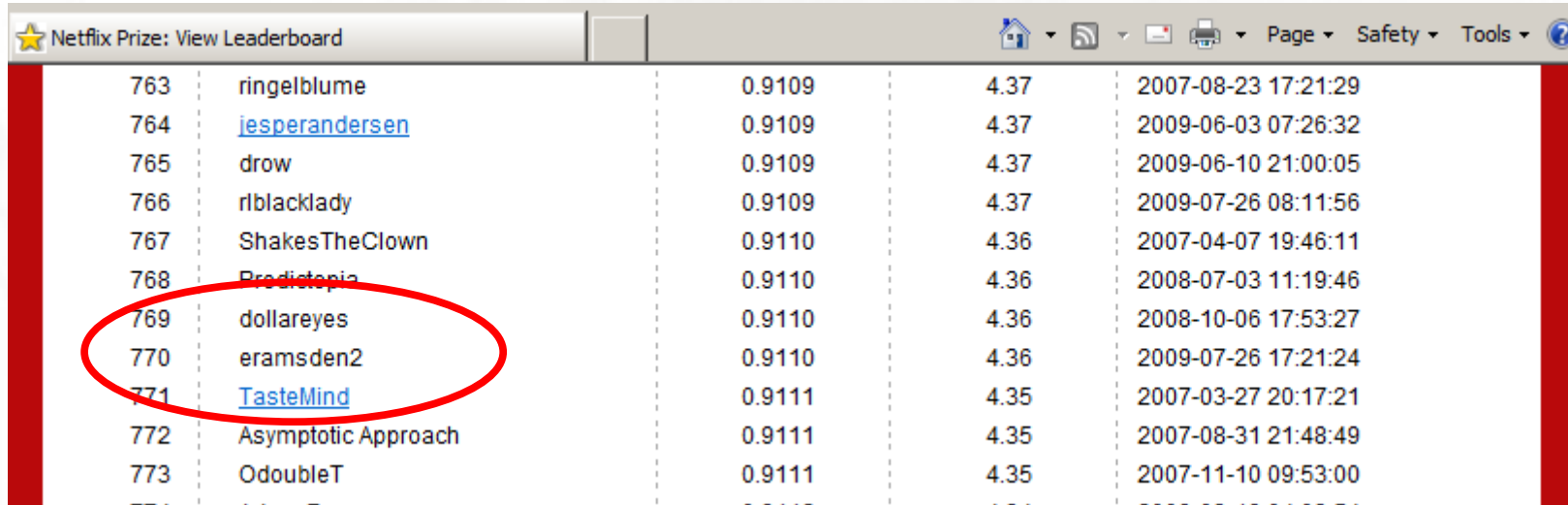
AusDM Ensembling Challenge

- <http://www.tiberius.biz/ausdm09/AusDM09EnsemblingChallenge.pdf>

GroupLens Research (University of Minnesota – other movie data sets)

- <http://www.grouplens.org/>

Conclusion:



The screenshot shows a browser window titled "Netflix Prize: View Leaderboard". The browser's address bar and navigation icons are visible at the top. The main content is a table with five columns: Rank, Username, Rating, Score, and Date. The row for "TasteMind" (rank 771) is circled in red. The table data is as follows:

Rank	Username	Rating	Score	Date
763	ringelblume	0.9109	4.37	2007-08-23 17:21:29
764	jesperandersen	0.9109	4.37	2009-06-03 07:26:32
765	drow	0.9109	4.37	2009-06-10 21:00:05
766	rlblacklady	0.9109	4.37	2009-07-26 08:11:56
767	ShakesTheClown	0.9110	4.36	2007-04-07 19:46:11
768	Prodiotopia	0.9110	4.36	2008-07-03 11:19:46
769	dollareyes	0.9110	4.36	2008-10-06 17:53:27
770	eramsden2	0.9110	4.36	2009-07-26 17:21:24
771	TasteMind	0.9111	4.35	2007-03-27 20:17:21
772	Asymptotic Approach	0.9111	4.35	2007-08-31 21:48:49
773	Odoubt	0.9111	4.35	2007-11-10 09:53:00

Questions?