Generating Top-N Recommendations from Binary Profile Data

Michael Hahsler
Marketing Research and e-Business Adviser
Hall Financial Group, Frisco, Texas, USA
Hall Wines, St. Helena, California, USA

Berufungsvortrag “Wirtschaftsinformatik”, WU Wien, July 16, 2008.

Outline

1. Motivation
2. Recommender Systems
3. Recommender Systems at Hall Wines
4. Collaborate Filtering Recommendation Techniques using Binary Data
5. Conclusion

Motivation

Motivation

- Hall Wines is a fast growing winery in Napa Valley.
- By 2012 the new landmark visitor center will be finished and attract an estimated 145,000 visitors per year.
- Production and sales will double to about 100,000 cases per year.

Motivation (cont.)

Concentration on direct-to-consumer (DTC) sales:

- 57% of wineries in the US project DTC sales to be the fastest growing channel in 2008 (VinterActive Research, 2008).
- DTC sales generate on average twice the profits per case by bypassing 2 or 3 tiers (distributor, wholesale, retail).

Key components of DTC sales: Tasting room, wine club, Internet, direct mail, phone, events.

To support a large and growing customer base substantial investments in customer relationship management (CRM) are under way.

Part of the analytical CRM initiative are ***Recommender systems.***

Recommender Systems

Recommender Systems

Recommender systems apply statistical and knowledge discovery techniques to the problem of making product recommendations (Sarwar, Karypis, Konstan, and Riedl, 2000).

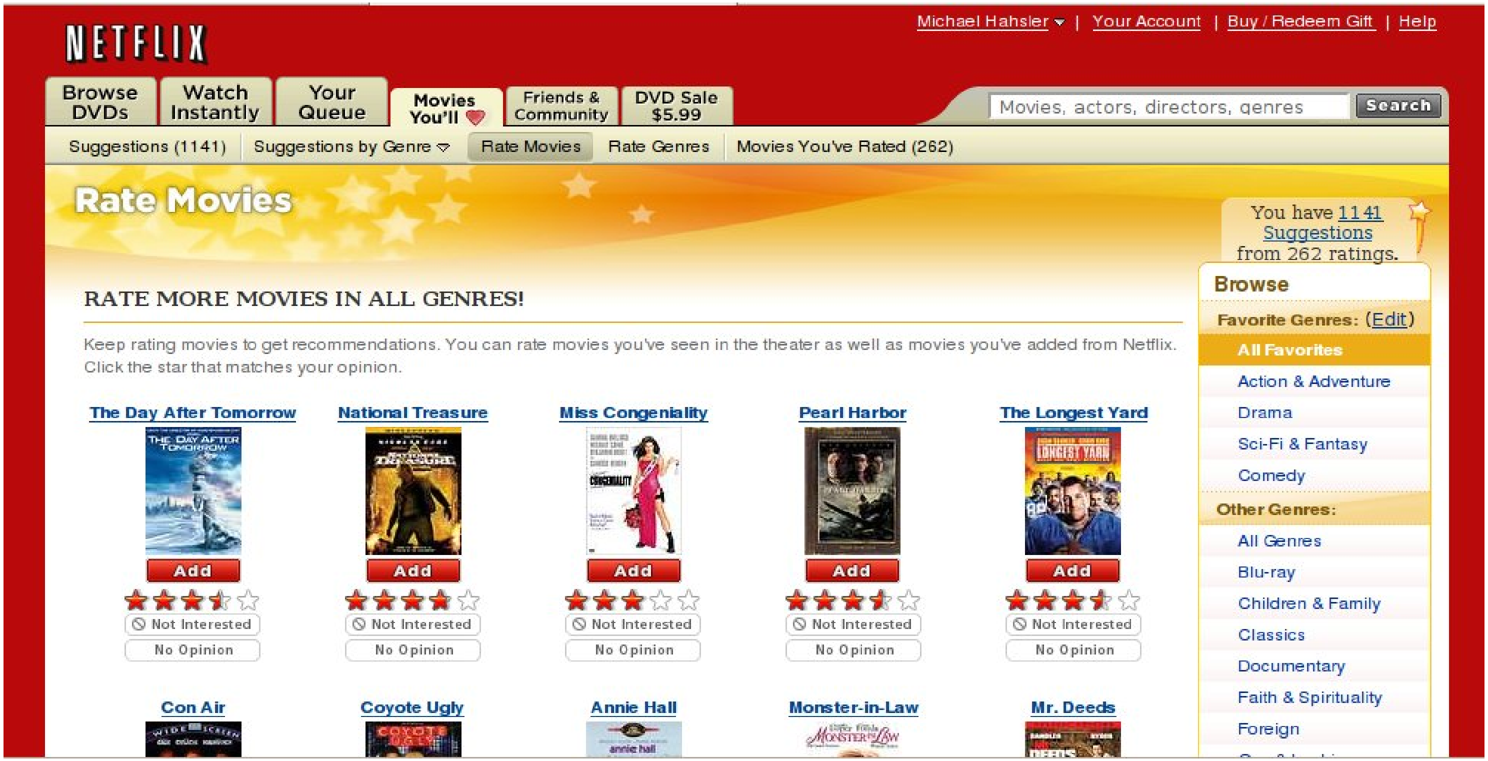
Advantages of recommender systems (Schafer, Konstan, and Riedl, 2001):

- Improve conversion rate: Help customer find a product she/he wants to buy.
- Cross-selling: Suggest additional products.
- Improve loyalty: By creating a value-added relationship.

Types of recommender systems (Ansari, Essegaier, and Kohli, 2000):

- Content filtering: Consumer preferences for product attributes.
- ***Collaborative filtering:*** Mimics word-of-mouth based on analysis of rating/usage/sales data.

Recommender Systems (cont.)



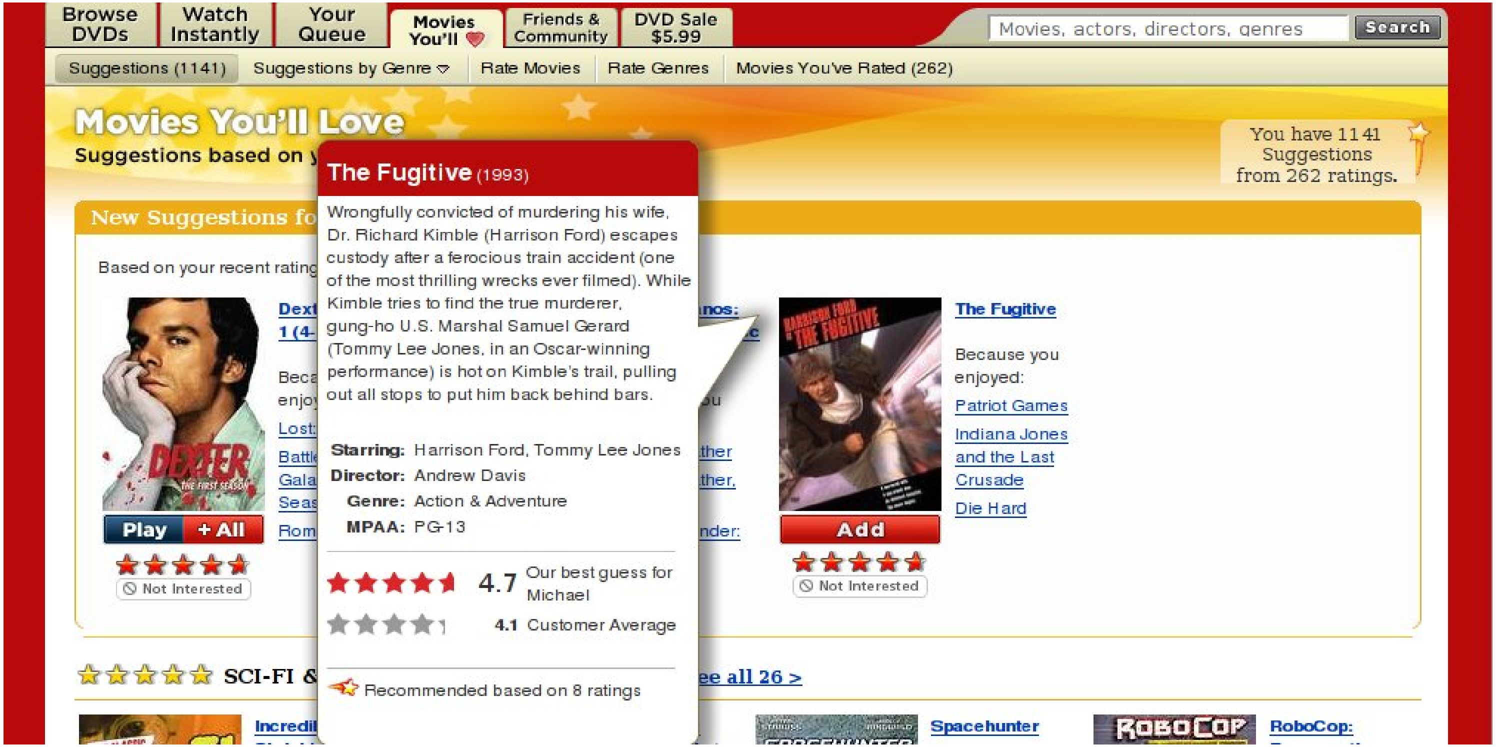
You have 1141 Suggestions from 262 ratings.

RATE MORE MOVIES IN ALL GENRES!

Keep rating movies to get recommendations. You can rate movies you've seen in the theater as well as movies you've added from Netflix. Click the star that matches your opinion.

Input: Typically rating data (here 1-5 stars for movies).

Recommender Systems (cont.)



The Fugitive (1993)

Wrongfully convicted of murdering his wife, Dr. Richard Kimble (Harrison Ford) escapes custody after a ferocious train accident (one of the most thrilling wrecks ever filmed). While Kimble tries to find the true murderer, gung-ho U.S. Marshal Samuel Gerard (Tommy Lee Jones, in an Oscar-winning performance) is hot on Kimble's trail, pulling out all stops to put him back behind bars.

Starring: Harrison Ford, Tommy Lee Jones
Director: Andrew Davis
Genre: Action & Adventure
MPAA: PG-13

★★★★★ 4.7 Our best guess for Michael

★★★★ 4.1 Customer Average

Recommended based on 8 ratings

Output:

- Predicted rating of unrated movies (Breese, Heckerman, and Kadie, 1998)
- A top-N list of unrated (unknown) movies ordered by predicted rating/score (Deshpande and Karypis, 2004)

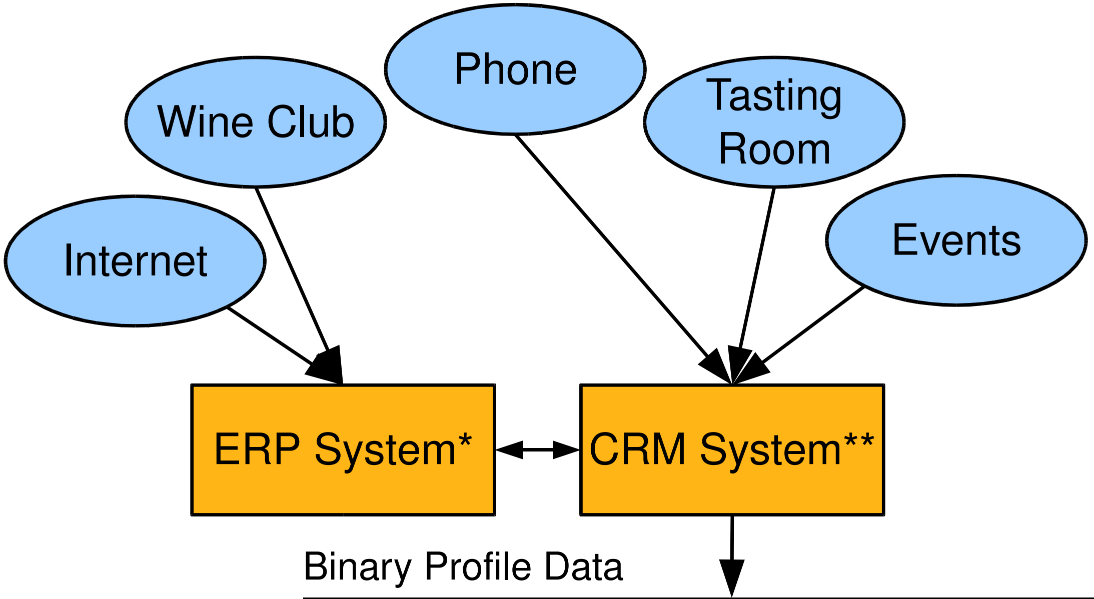
Recommender Systems at Hall Wines

Data Sources for Hall Wine

- 10 core wines produced every year and distributed nationally
- 20 to 30 single vineyard and speciality wines
- 3 vintages are offered and a library for older wines is planned
- Plans for a wine club under a different brand with several hundred wines from California

⟶ 120–500 different wines
⟶ 500,000+ customers

Data Sources (cont.)



Binary Profile Data

Customer ID	Wine A	Wine B	...	Wine Z
1	1	0	...	1
2	0	1	...	0
3	1	0	...	0
4	0	0	...	1

* Enterprise Resource Planning System

** Customer Relationship Management System

Data Sources (cont.)

Reason for binary profile data:

Heterogeneity of collected data

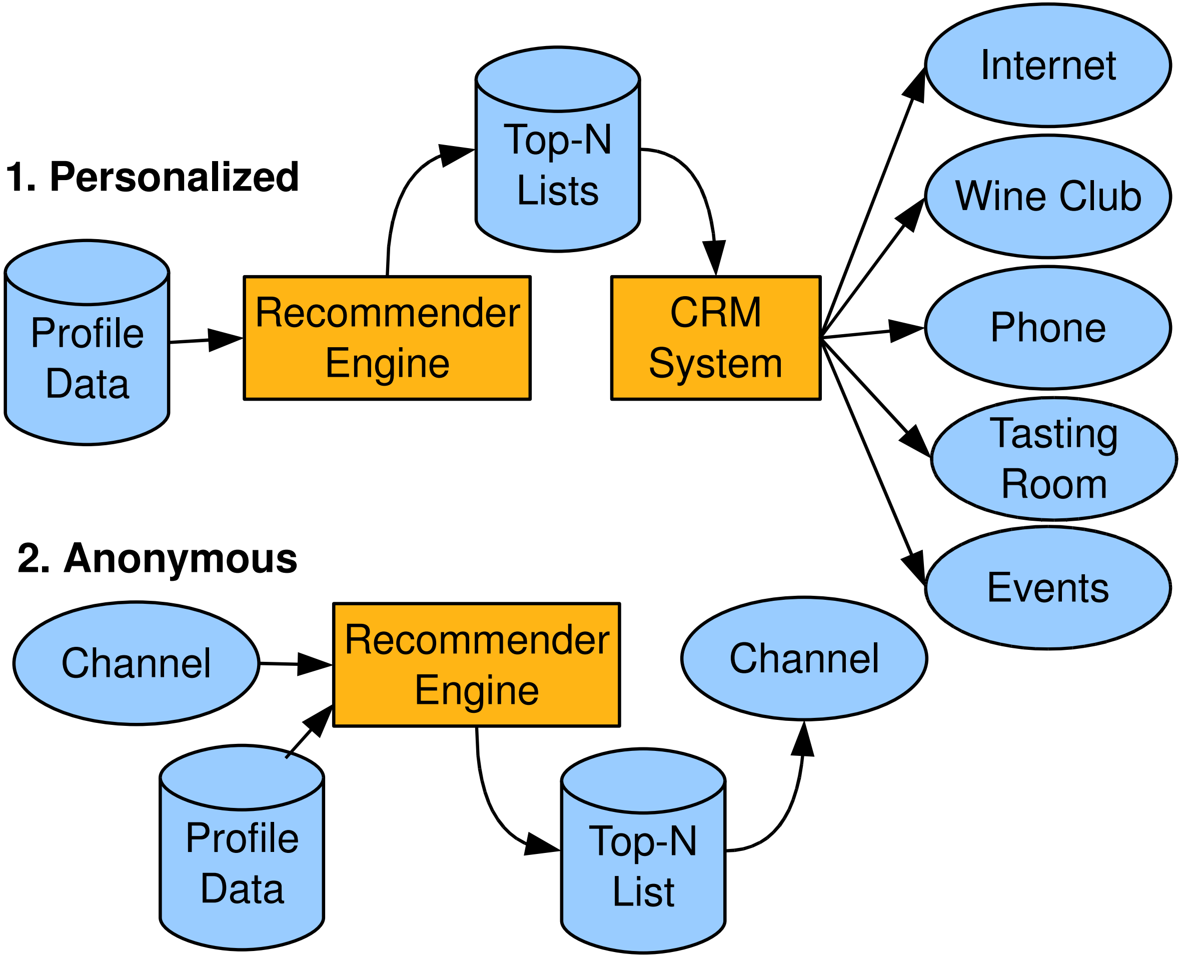
Examples:

- A customer purchases a case of wine after a tasting at the tasting room.
- A customer returns wine she/he bought online.
- A customer indicates wine is his favorite in a wine tasting event but does not buy.
- A wine club member gets her/his monthly shipment of wine.
- A routine call to a customer reveals that she/he did not enjoy the wine she bought a month ago.
- A customer repeatedly visits the description page of a wine on the web site.

Typical situation for many businesses:

- No rating data available
- Extremely heterogeneous data sources
- ***Very limited research on recommender systems based on binary data available.***

Recommender Engine



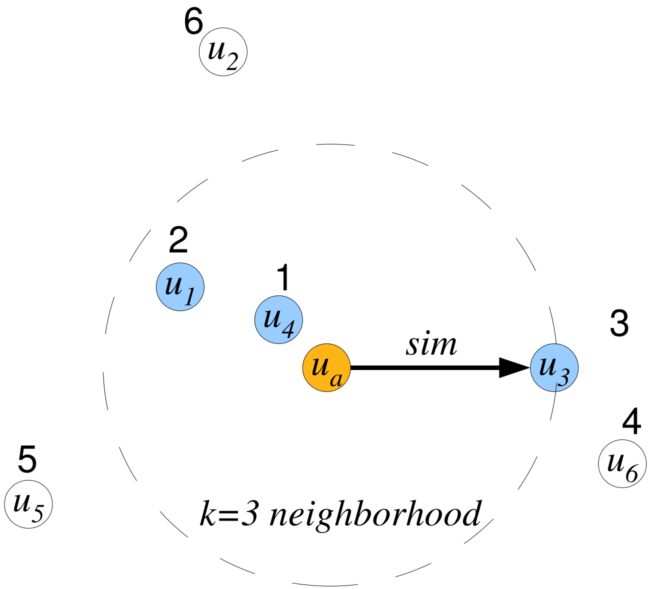
Collaborate Filtering Recommendation Techniques for Binary Data

User-based Collaborative Filtering (CF)

Produce recommendations based on preferences of similar users (Goldberg, Nichols, Oki, and Terry, 1992; Resnick, Iacovou, Suchak, Bergstrom, and Riedl, 1994; Mild and Reutterer, 2001).

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_a	0	0	1	1	0	1	0	1
u_1	0	1	1	1	1	0	0	0
u_2	1	0	0	0	1	1	0	0
u_3	0	0	0	1	0	1	0	1
u_4	0	0	0	1	1	1	0	1
u_5	1	1	0	0	0	0	0	1
u_6	0	1	0	0	1	1	0	1
	0	1	~~1~~	~~3~~	2	~~2~~	0	~~2~~

Recommendation: i_5, i_2



1. Find k nearest neighbors based on similarity between users.
2. Generate recommendation based on the items liked by the k nearest neighbors. E.g., recommend most popular items or use a weighing scheme.

User-based CF (cont.)

Measure similarity between two users u_x and u_y:

- ***Pearson correlation coefficient:***

$$\mathrm{sim}_{\mathrm{Pearson}}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i \in I} x_i y_i - I\bar{\mathbf{x}}\bar{\mathbf{y}}}{(I-1) s_x s_y}$$

- ***Cosine similarity:***

$$\mathrm{sim}_{\mathrm{Cosine}}(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$$

- ***Jaccard index*** (only binary data):

$$\mathrm{sim}_{\mathrm{Jaccard}}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

where $\mathbf{x} = b_{u_x,\cdot}$ and $\mathbf{y} = b_{u_y,\cdot}$ represent the user's profile vectors and X and Y are the sets of the items with a 1 in the respective profile.

Problems: Memory-based. Expensive online similarity computation.

Item-based CF

Produce recommendations based on item similarities (Kitts, Freed, and Vrieze, 2000; Sarwar, Karypis, Konstan, and Riedl, 2001)

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
i_1	~~1~~	~~0.1~~	~~0~~	**0.3**	**0.2**	**0.4**	~~0~~	~~0.1~~
i_2	~~0.1~~	~~1~~	**0.8**	**0.9**	~~0~~	**0.2**	~~0.1~~	~~0~~
i_3	~~0~~	**0.8**	~~1~~	~~0~~	**0.4**	~~0.1~~	~~0.3~~	**0.5**
i_4	**0.3**	**0.9**	~~0~~	~~1~~	~~0~~	**0.3**	~~0~~	~~0.1~~
i_5	**0.2**	~~0~~	**0.4**	~~0~~	~~1~~	**0.1**	~~0~~	~~0~~
i_6	**0.4**	**0.2**	~~0.1~~	**0.3**	~~0.1~~	~~1~~	~~0~~	~~0.1~~
i_7	~~0~~	**0.1**	**0.3**	~~0~~	~~0~~	~~0~~	~~1~~	~~0~~
i_8	**0.1**	~~0~~	**0.5**	**0.1**	~~0~~	~~0.1~~	~~0~~	~~1~~
	~~0.3~~	0	0.9	0.4	~~0.2~~	0.5	0	~~0~~

k=3

$u_a=\{i_1, i_5, i_8\}$

Recommendation: i_3, i_6, i_4

1. Calculate similarities between items and keep for each item only the values for the k most similar items.
2. For each item add the similarities with the active user's items.
3. Remove the items of the active user and recommend the N items with the highest score.

Item-based CF (cont.)

Similarity measures:

- Pearson correlation coefficient, cosine similarity, jaccard index
- ***Conditional probability-based similarity*** (Deshpande and Karypis, 2004):

$$\mathrm{sim}_{\mathrm{Conditional}}(x, y) = \frac{\mathrm{Freq}(xy)}{\mathrm{Freq}(x)} = \hat{P}(y|x)$$

  where x and y are two items, $\mathrm{Freq}(\cdot)$ is the number of users with the given item in their profile.

Properties:

- Models (reduced similarity matrix) is relatively small ($N \times k$) and can be fully precomputed.
- Item-based CF is known to only produce slightly inferior results compared to user-based CF (Deshpande and Karypis, 2004).
- Higher order models which take the joint distribution of sets of items into account are possible (Deshpande and Karypis, 2004).
- Successful application in large scale systems (e.g., Amazon.com)

Association Rules

Produce recommendations based on a dependency model for items given by association rules (Fu, Budzik, and Hammond, 2000; Mobasher, Dai, Luo, and Nakagawa, 2001; Geyer-Schulz, Hahsler, and Jahn, 2002; Lin, Alvarez, and Ruiz, 2002; Demiriz, 2004)

The binary profile matrix $\mathbf{B}$ is seen as a database containing the set of items $\mathcal{I} = \{i_1, i_2, \ldots, i_I\}$. Each user is treated as a transaction.

Rule: $X \rightarrow Y$ where $X, Y \subseteq \mathcal{I}$, $X \cap Y = \emptyset$ and $|Y| = 1$.

Measures of significance and interestingness:

$$\mathrm{support}(X \rightarrow Y) = \mathrm{support}(X \cup Y) = \mathrm{Freq}(X \cup Y)/U > s$$

$$\mathrm{confidence}(X \rightarrow Y) = \mathrm{support}(X \cup Y)/\mathrm{support}(X) = \hat{P}(Y|X) > c$$

Length constraint:

$$|X \cup Y| \leq l$$

Association Rules (cont.)

1. Dependency model: All rules of form $X \rightarrow Y$ with minimum support s, minimum confidence c and satisfying the length constraint l.
2. Find all maching rules $X \rightarrow Y$ for which $X \subseteq u_a$.
3. Recommend N unique right-hand-sides (Y) of the maching rules with the highest confidence.

Properties:

- Model grows in the worst case exponentially with the number of items. Model size can be controlled by l, s and c.
- Model is very similar to item-based CF with conditional probability-based similarity (with higher order effects).

Comparison – MovieLense

MovieLens data set:

Rating matrix $\mathbf{R} = (r_{u,i})$ where $r_{u,i}$ is the rating (1–5 stars or "not rated") by user $u \in 1, \ldots, U$ for item $i \in 1, \ldots, I$. $U = 943$ users and $I = 1682$ movies.

Creating binary data and preprocessing:

1. Conversions to binary profile matrix $\mathbf{B} = (b_{u,i})$ where
$$b_{u,i} = \begin{cases} 1 & \text{if } r_{u,i} \geq 3, \\ 0 & \text{otherwise.} \end{cases}$$
2. Remove duplicated movies and movies without name in name file.
3. Remove users with less than 10 items in profile.

Used data set:

Binary profile matrix $\mathbf{B}$ with $U = 941$ users times $I = 1559$ items containing 81984 ones (density = 0.056).
Average items per profile: 86.97

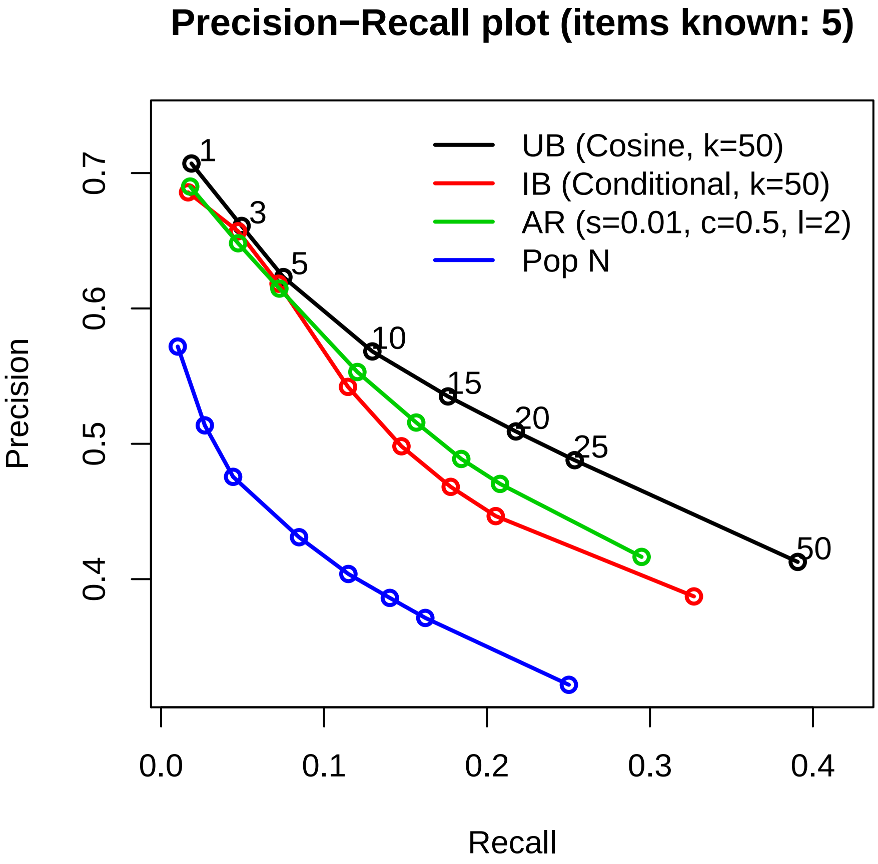
Evaluation Setup

- 4-fold evaluation with 75% training data and 25% test data.
- 1 or 5 items for users in test data know.
- Generate top-N recommendation lists. N is varied between 1 and 50.
- How well they can predict the remaining items?
  Evaluation with averaged precision/recall plots.

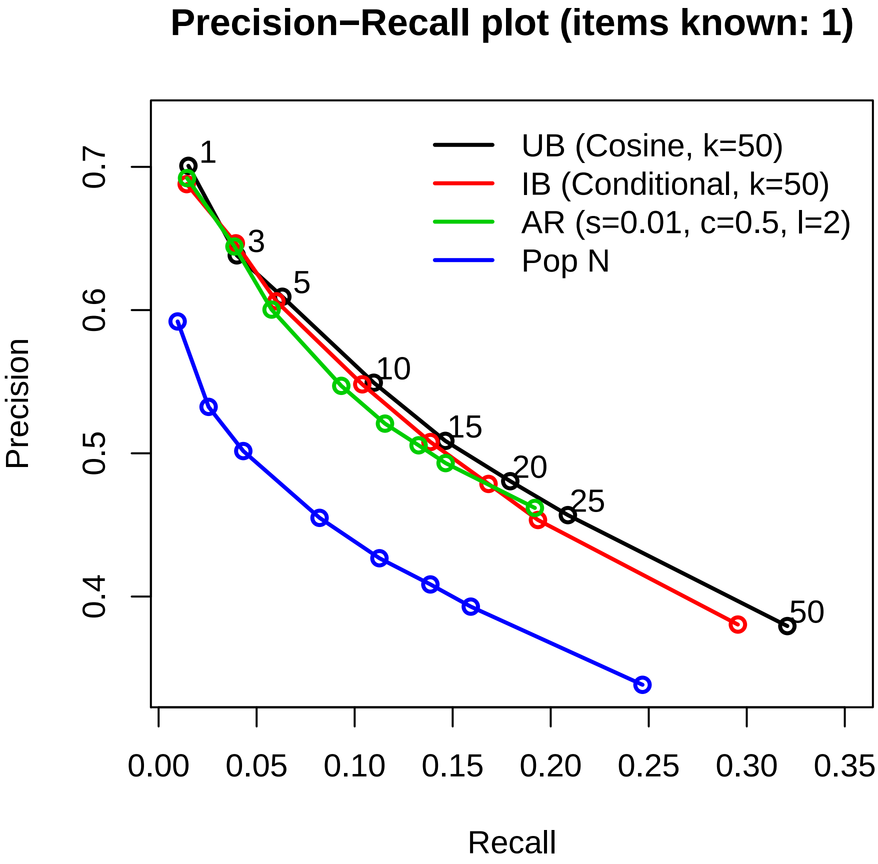
$$\text{precision} = \frac{tp}{tp + fp} = \frac{\text{\# correctly predicted items}}{N}$$

$$\text{recall} = \frac{tp}{tp + fn} = \frac{\text{\# correctly predicted items}}{\text{\# items to be predicted}}$$

Comparison – MovieLense



Comparison – MovieLense (cont.)



Comparison – MovieLense (cont.)

Top-N recommendation lists can be considered rankings and can be represented as N-ary order relations ($\leq$) on the set of all items in the lists (items which do not occur in all lists are added to the end of the corresponding order).

Average distance between methods: E.g., the cardinality of the symmetric difference of two relations (the number of tuples contained in exactly one of two relations) (Hornik and Meyer, 2008)

Total number of tuples in relations: 8649 (for 93 items)

```
        UB   IB   AR Pop N
UB       0 2555 2634  3317
IB    2555    0 1611  2052
AR    2634 1611    0  2636
Pop N 3317 2052 2636     0
```

Comparison – MovieLense (cont.)

u_a = {Toy Story (1995), Air Force One (1997),
      Cop Land (1997), Michael (1996), Blade Runner (1982)}

	UB	IB	AR	Top N	Consensus*
Contact (1997)	1	4	25	4	4
Liar Liar (1997)	2	18	32	10	8
Conspiracy Theory (1997)	3	NA	NA	NA	NA
Star Wars (1977)	4	1	1	1	1
Return of the Jedi (1983)	5	2	4	2	2
English Patient, The (1996)	6	26	41	6	12
Saint, The (1997)	7	NA	NA	NA	NA
Fargo (1996)	8	3	13	3	3
Mr. Holland's Opus (1995)	9	34	30	39	17
Scream (1996)	10	20	35	7	10

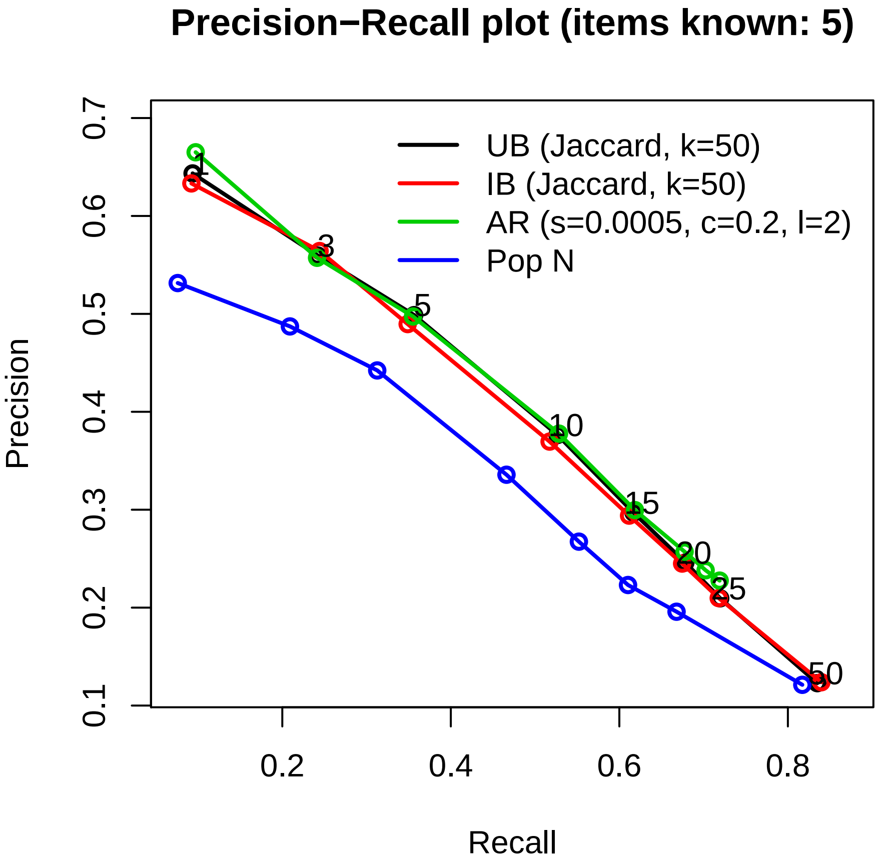
* Consensus method: Condorcet (minimizes the weighted sum of symmetric difference distance)

Comparison – MSWeb

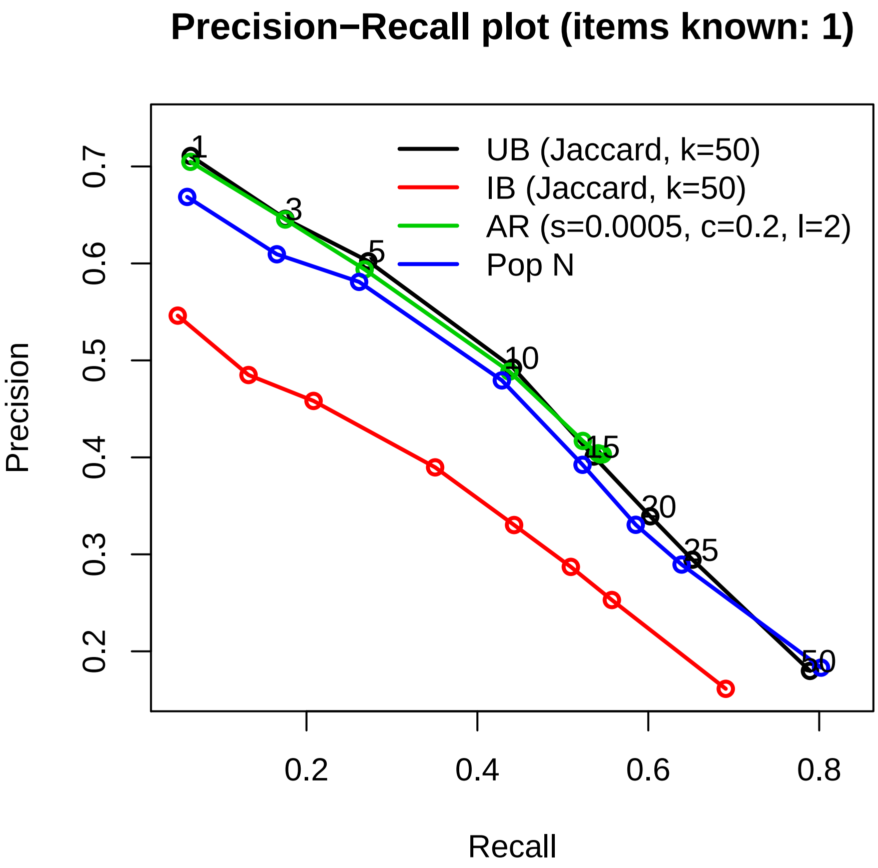
Anonymous web data from www.microsoft.com. The data records the use of www.microsoft.com by 38,000 anonymous, randomly-selected users. For each user, the data lists all the areas of the web site (Vroots) that user visited in a one week time frame (Breese et al., 1998).

Used data:
4151 users with > 5 items and 285 items.
Avg. items per profile: 8.16

Comparison – MSWeb (cont.)



Comparison – MSWeb (cont.)

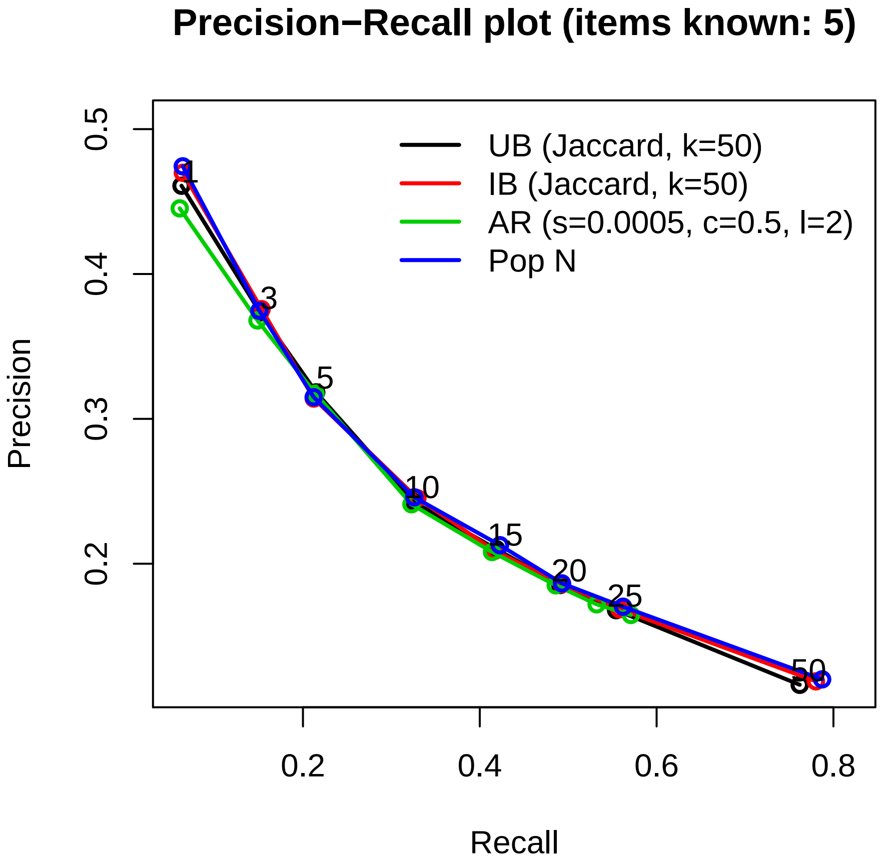


Comparison – Groceries

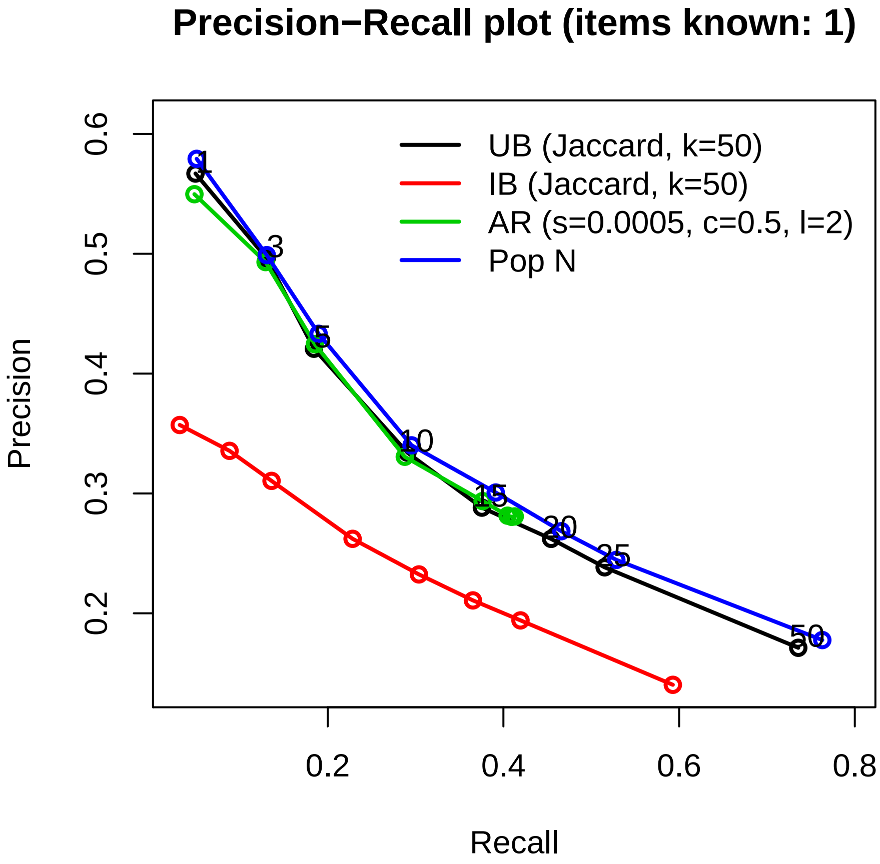
Contains 1 month (30 days) of real-world point-of-sale transaction data from a typical local grocery outlet. The data set contains 9835 transactions and the items are aggregated to 169 categories (Hahsler, Gruen, and Hornik, 2005).

Used data:
2874 transactions with > 5 categories and 169 items.
Avg. items per profile: 8.95

Comparison – Groceries (cont.)



Comparison – Groceries (cont.)



Comparison (cont.)

Challenges with CF and binary data:

- Sparsity of data has an adverse effect on similarity calculation/neighborhood formation.
- The meaning of 0 in the data can be either of “unknown” or “dislike”. Measures like the Jaccard index focus on 1s.
- Quality of recommendations is extremely data set dependent.

Bias of evaluation method:

- Complete set of “liked” items is unknown. Measured precision is only a lower bound to real precision.

Conclusion & Future Research

- Many companies face heterogeneous data sources which can best be aggregated into binary data.
- There is only limited research on CF recommender systems using binary data (except (Breese et al., 1998; Mild and Reutterer, 2001, 2003; Demiriz, 2004)) available. More research is needed.

Future research

- Provide an open source research toolbox for recommendations based on binary data with the R code used here.
- Model “dislike” in binary data (Mobasher et al., 2001).
- Investigate recommender engines based on consensus rankings.

References

A. Ansari, S. Essegaier, and R. Kohli. Internet recommendation systems. *Journal of Marketing Research*, 37:363–375, 2000.

J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 43–52, 1998.

A. Demiriz. Enhancing product recommender systems on sparse binary data. *Data Minining and Knowledge Discovery*, 9(2):147–170, 2004. ISSN 1384-5810.

M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms. *ACM Transations on Information Systems*, 22(1):143–177, 2004. ISSN 1046-8188.

X. Fu, J. Budzik, and K. J. Hammond. Mining navigation history for recommendation. In *IUI '00: Proceedings of the 5th international conference on Intelligent user interfaces*, pages 106–112. ACM, 2000. ISBN 1-58113-134-8.

A. Geyer-Schulz, M. Hahsler, and M. Jahn. A customer purchase incidence model applied to recommender systems. In R. Kohavi, B. Masand, M. Spiliopoulou, and J. Srivastava, editors, *WEBKDD 2001 - Mining Log Data Across All Customer Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001, Revised Papers*, Lecture Notes in Computer Science LNAI 2356, pages 25–47. Springer-Verlag, July 2002.

D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992. ISSN 0001-0782.

M. Hahsler, B. Gruen, and K. Hornik. arules – A computational environment for mining association rules and frequent item sets. ***Journal of Statistical Software***, 14(15):1–25, October 2005. ISSN 1548-7660. URL `http://www.jstatsoft.org/v14/i15/`.

K. Hornik and D. Meyer. ***relations: Data Structures and Algorithms for Relations***, 2008. R package version 0.3-3.

B. Kitts, D. Freed, and M. Vrieze. Cross-sell: a fast promotion-tunable customer-item recommendation method based on conditionally independent probabilities. In ***KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining***, pages 437–446. ACM, 2000. ISBN 1-58113-233-6. doi: http://doi.acm.org/10.1145/347090.347181.

W. Lin, S. A. Alvarez, and C. Ruiz. Efficient adaptive-support association rule mining for recommender systems. ***Data Mining and Knowledge Discovery***, 6(1):83–105, 2002. ISSN 1384-5810.

A. Mild and T. Reutterer. Collaborative filtering methods for binary market basket data analysis. In ***AMT '01: Proceedings of the 6th International Computer Science Conference on Active Media Technology***, pages 302–313, London, UK, 2001. Springer-Verlag. ISBN 3-540-43035-0.

A. Mild and T. Reutterer. An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. ***Journal of Retailing and Consumer Services***, 10(3):123–133, 2003.

B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In ***Proceedings of the ACM Workshop on Web Information and Data Management (WIDM01), Atlanta, Georgia***, 2001.

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In ***CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work***, pages 175–186. ACM, 1994. ISBN 0-89791-689-1. doi: http://doi.acm.org/10.1145/192844.192905.

B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In ***EC '00: Proceedings of the 2nd ACM conference on Electronic commerce***, pages 158–167. ACM, 2000. ISBN 1-58113-272-7.

B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In ***WWW '01: Proceedings of the 10th international conference on World Wide Web***, pages 285–295. ACM, 2001. ISBN 1-58113-348-0.

J. B. Schafer, J. A. Konstan, and J. Riedl. E-commerce recommendation applications. ***Data Mining and Knowledge Discovery***, 5(1/2):115–153, 2001.

VinterActive Research. VinQuest 2008: U.S. consumer direct wine sales trends, March 2008.