

RECOMMENDER SYSTEMS

Petr Ryšavý

Thursday 8th December, 2016

IDA, Dept. of Computer Science, FEE, CTU

RECOMMENDER SYSTEMS

- Goal is to predict "rating" or "preference".
- System propose to user items that the user will be likely interested in.
- Great development in past years.
- Goals: keep attention of user, get more profit, make users more happy with the service
- Many areas: sellers (Amazon), movies (Netflix), music (Pandora), news (NYT) and many others

The screenshot shows a search engine interface with the following elements:

- Search Bar:** Contains the text "boty na běžky|heuréka" and a "Vyhledat" (Search) button.
- Search Results:**
 - A dropdown menu lists suggestions: "boty na běžky", "boty na běžky heuréka", "boty na běžky skol", "boty na běžky atomic", "boty na běžky botas", and "boty na běžky fischer".
 - A main result for "**Boty na bezky**" is displayed, including a description: "Boty na běžky za výhodné ceny. Vyberte u profesionálů online." and a link to "Reklama gregorysport.cz/boty_na_bezky".
 - Below the main result are links: "Boty na běžky - Běžecské lyže - Doplnky pro lyžování - Hůlky".
 - At the bottom, another link "Boty na běžky" is visible.
- Right Sidebar:** Contains a partially visible advertisement for "Salomon Pilot" with a price of "3 300 Kč" and the word "Lyže".
- Left Sidebar:** Shows a vertical list of product images and thumbnails.

Necelých půl roku po odchodu z Downing Street představila manželka...

REKLAMA



Fischer RC5 SKATE 46

Koupit

3 899 Kč

Ježíškovy pošty speciálním
ručním příležitostním
razítkem. Autorkou...



Fischer RC5 SKAT

Koupit

3 899 Kč

NÁZORY



Jan Macháček

**Američtí liberálové až po uši v depresi
a apokalypse**

*Aby bylo jasno, chápu, že v Americe je pořád
lecko v šoku ze Trumpova zvolení. Chápu, že
mnozí ještě lapají po dechu...*



Temný a světlejší scénář pro Evropu

Nejproblematičtější dopady zvolení Donalda

IDNES



**Sudové víno
Zásah je nut**

Víno se nekou
lihové aféře. Je
Hajda (ČSSD), l
zkomplikuje pr









**Octavia už s
vyjde výhod**

Když v roce 20

General setting

- Set of users \mathbf{C} .
- Set of items \mathbf{S} .
- Utility function $u : \mathbf{C} \times \mathbf{S} \mapsto \mathcal{R}$.
- \mathcal{R} is set of ratings (number of stars, liked/disliked).

						
Alice	1			0.8	0.7	0.9
Bob	0.9		0.2		0.8	
Cecil						0.7
David			0.3	0.2	0.2	

Several challenges

- How to get ratings? - ask users explicitly/watch their behavior
- Data are extremely sparse.
- Cold start.

- *Demographic systems* - recommendations based on age, location, gender etc.
- *Content-based systems* - find similar items to items favored by user
- *Collaborative filtering* - recommended items preferred by similar users
- *Hybrid systems* - combines two or more recommendation approaches

DEMOGRAPHIC SYSTEMS

- ✓ No need for history of user ratings
- ✓ Only few observations needed
- ✗ Privacy issues
- ✗ Very stereotypical

CONTENT-BASED SYSTEMS

- Recommend items similar to those that were rated recently by the user.
- Each item is assigned [item profile](#).
- Metadata, content of newspaper articles, reviews by users, tags (artificial/manual), words in newspaper articles
- Normalization, proper scaling of numerical features, higher weight to more rare words

Evaluating similarity







- Jaccard index

$$\frac{|A \cap B|}{|A \cup B|}$$

- Cosine similarity

$$\frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

- Should not depend on items that users have not rated or attributes that were not assigned to none of the items.
- Normalization of ratings.

						
Alice	1			0.8	0.7	0.9
Bob	0.9		0.2		0.8	
Cecil						0.7
David			0.3	0.2	0.2	







- Represent attributes that are favored by the users.
- Weighted average of rated item profiles.
- Classification can be applied to predict rating.
- For example decision trees/random forest/regression.

- ✓ No need for data of other users.
- ✓ Can recommend to users which have unique taste.
- ✓ Avoids "first rater problem"
- ✗ Need to extract features (manually/by special algorithms)
- ✗ Cold start for new users.

COLLABORATIVE FILTERING

Collaborative filtering

- Predict rating based on actions of similar users.

						
Alice	1			0.8	0.7	0.9
Bob	0.9		0.2		0.8	
Cecil						0.7
David			0.3	0.2	0.2	

- For similarity calculations we can use cosine similarity or Jaccard index.
- We need to normalize.

- Way how to deal with sparsity - allows us to estimate blanks.
- There is low probability that two users liked the same item of the same kind.
- Hierarchically cluster items and/or users.

- Another way to estimate blank fields in utility matrix.
- UV-decomposition

$$\mathbf{M} = \mathbf{U} \times \mathbf{V} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ \vdots & \vdots \\ u_{|U|1} & u_{|U|2} \end{bmatrix} \times \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1|S|} \\ v_{21} & v_{22} & \cdots & v_{2|S|} \end{bmatrix}.$$

- We minimize *Root-Mean-Square-Error* of non-blanks in utility matrix \mathbf{M} .

$$\mathbf{M} = \mathbf{U} \times \mathbf{V} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ \vdots & \vdots \\ u_{|U|1} & u_{|U|2} \end{bmatrix} \times \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1|S|} \\ v_{21} & v_{22} & \cdots & v_{2|S|} \end{bmatrix}.$$

- Gradient descent.
- Minimization of RMSE w.r.t. selected parameter u_{ij} or v_{ij} .
- Many local optima. Random restarts and different order of parameters selected for minimization used.

- ✓ No need to represent items as list of features.
- ✓ Scalable - no human interaction.
- ✗ "first rater problem"
- ✗ Cannot work for users with unique taste.
- ✗ Whole database needs to be processed.

HYBRID RECOMMENDER SYSTEMS

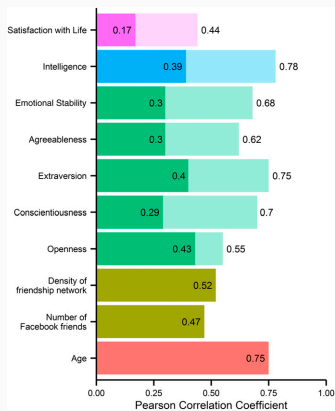
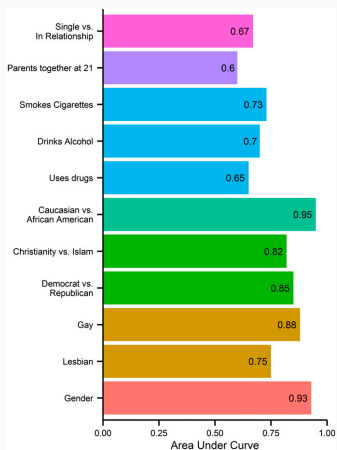
- Combine two or more recommendation approaches.
- Most popular are combinations of content-based systems and collaborative filtering.
- Provide the best results.
- Avoid weaknesses of individual approaches.

CONCLUSION

- In 2006 Netflix provided 100M ratings that 480k users gave to 18k movies.
- First team to provide algorithm 10% more accurate (RMSE) than Netflix algorithm gets \$1M.
- Progress prizes.
- Most accurate algorithm in 2007 used ensemble of 107 different algorithmic approaches. (k -NN, Restricted Boltzman Machines, ...) [2]
- Awarded in September 2009. [7]
- Netflix algorithm only 3% better than trivial baseline.
- 20k teams from 150 countries.
- Cancelled sequel due to a lawsuit against a Netflix user.

Privacy issues [4]

- Plenty of data publicly available.
- For example major US retail network used customer shopping record to predict pregnancies of female customers.



THANK YOU FOR YOUR ATTENTION.
TIME FOR QUESTIONS!



Gediminas Adomavicius and Alexander Tuzhilin.

Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.

IEEE transactions on knowledge and data engineering, 17(6):734–749, 2005.



Robert M Bell, Yehuda Koren, and Chris Volinsky.

The bellkor solution to the netflix prize, 2007.



Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez.

Recommender systems survey.

Knowledge-Based Systems, 46:109–132, 2013.



Michal Kosinski, David Stillwell, and Thore Graepel.

Private traits and attributes are predictable from digital records of human behavior.

Proceedings of the National Academy of Sciences, 110(15):5802–5805, 2013.



Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman.

Mining of massive datasets.

Cambridge University Press, 2014.



RVVSV Prasad and V Valli Kumari.

A categorical review of recommender systems.

International Journal of Distributed and Parallel Systems, 3(5):73, 2012.



Andreas Töscher, Michael Jahrer, and Robert M Bell.

The bigchaos solution to the netflix grand prize.

Netflix prize documentation, pages 1–52, 2009.