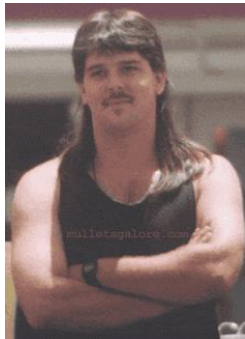


COMP 345: Data Mining Recommender Systems

Slides Adapted From: www.mmds.org (Mining Massive Datasets)

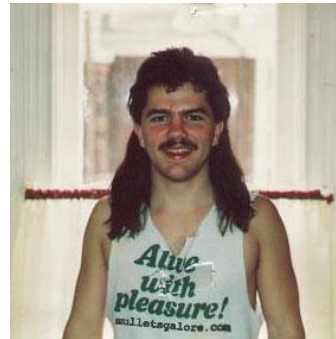


Example: Recommender Systems



■ Customer X

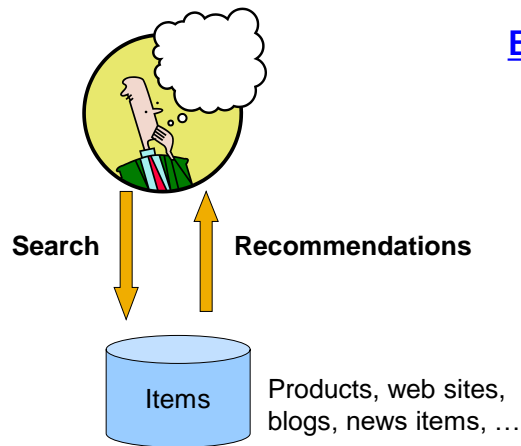
- Buys Metallica CD
- Buys Megadeth CD



■ Customer Y

- Does search on Metallica
- Recommender system suggests Megadeth from data collected about customer X

Recommendations



Examples:

amazon.com.



StumbleUpon



del.icio.us



movielens
helping you find the *right* movies

last.fm
the social music revolution

Google
News

YouTube

XBOX
LIVE

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

3

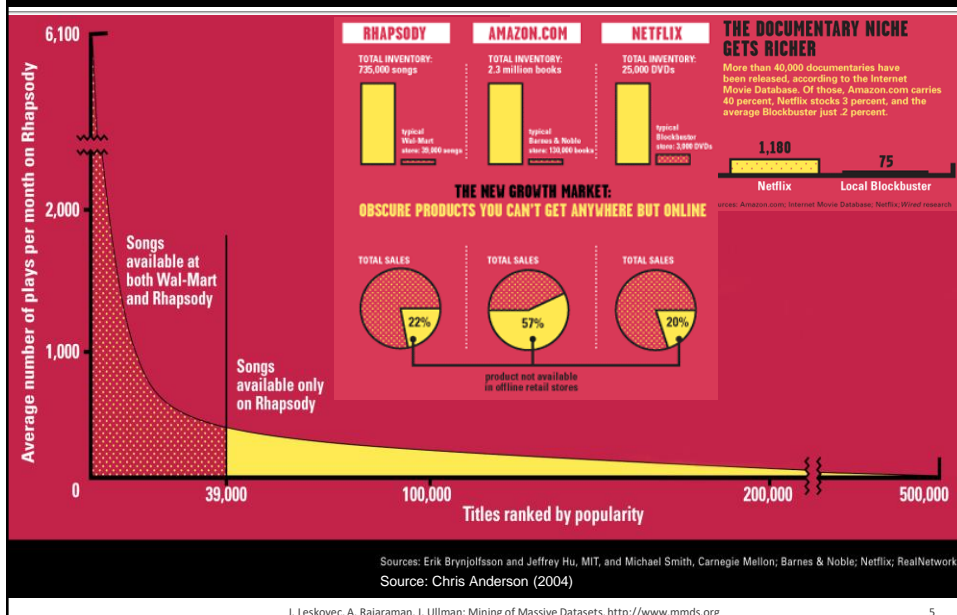
From Scarcity to Abundance

- **Shelf space is a scarce commodity for traditional retailers**
 - Also: TV networks, movie theaters,...
- **Web enables near-zero-cost dissemination of information about products**
 - From scarcity to abundance
- **More choice necessitates better filters**
 - Recommendation engines
 - How **Into Thin Air** made **Touching the Void** a bestseller: <http://www.wired.com/wired/archive/12.10/tail.html>

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

4

Sidenote: The Long Tail



Types of Recommendations

- **Editorial and hand curated**
 - List of favorites
 - Lists of “essential” items
- **Simple aggregates**
 - Top 10, Most Popular, Recent Uploads
- **Tailored to individual users**
 - Amazon, Netflix, ...

Formal Model

- X = set of **Customers**
- S = set of **Items**
- **Utility function** $u: X \times S \rightarrow R$
 - R = set of ratings
 - R is a totally ordered set
 - e.g., **0-5 stars**, real number in **[0,1]**

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmids.org>

7

Utility Matrix

	Avatar	LOTR	Matrix	Pirates
Alice	1		0.2	
Bob		0.5		0.3
Carol	0.2		1	
David				0.4

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmids.org>

8

Key Problems

- **(1) Gathering “known” ratings for matrix**
 - How to collect the data in the utility matrix
- **(2) Extrapolate unknown ratings from the known ones**
 - Mainly interested in high unknown ratings
 - We are not interested in knowing what you don’t like but what you like
- **(3) Evaluating extrapolation methods**
 - How to measure success/performance of recommendation methods

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmids.org>

9

(1) Gathering Ratings

- **Explicit**
 - Ask people to rate items
 - Doesn’t work well in practice – people can’t be bothered
- **Implicit**
 - Learn ratings from user actions
 - E.g., purchase implies high rating
 - What about low ratings?

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmids.org>

10

(2) Extrapolating Utilities

- **Key problem:** Utility matrix U is **sparse**
 - Most people have not rated most items
 - **Cold start:**
 - New items have no ratings
 - New users have no history
- **Three approaches to recommender systems:**
 - 1) Content-based
 - 2) Collaborative
 - 3) Latent factor based

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

11

Content-based Recommender Systems

Content-based Recommendations

- **Main idea:** Recommend items to customer x similar to previous items rated highly by x

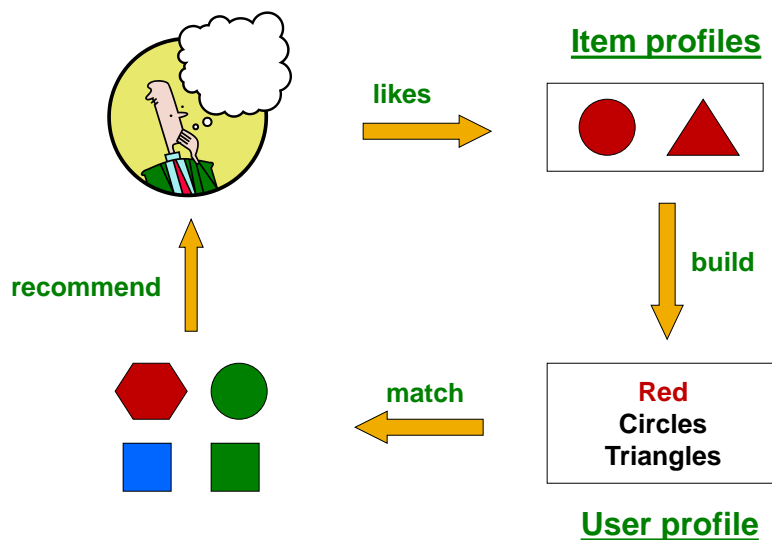
Example:

- **Movie recommendations**
 - Recommend movies with same actor(s), director, genre, ...
- **Websites, blogs, news**
 - Recommend other sites with “similar” content

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmids.org>

13

Plan of Action



J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmids.org>

14

Item Profiles

- For each item, create an **item profile**
- **Profile is a set of features**
 - **Movies:** author, title, actor, director,...
 - **Images, videos:** metadata or tags
 - **People:** set of friends
- Convenient to think of item profile as a vector
 - One entry per feature (each actor, director, etc...)
 - Vector might be Boolean or real-valued

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

15

User Profiles and Prediction

- **User profile possibilities:**
 - Weighted average of rated item profiles
 - **Variation:** weight by difference from average rating for item
 - ...
- **Prediction heuristic:**
 - Given user profile \mathbf{x} and item profile \mathbf{i} , estimate

$$u(\mathbf{x}, \mathbf{i}) = \cos(\mathbf{x}, \mathbf{i}) = \frac{\mathbf{x} \cdot \mathbf{i}}{||\mathbf{x}|| \cdot ||\mathbf{i}||}$$

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

16

Pros: Content-based Approach

- **+: No need for data on other users**
 - No cold-start or sparsity problems
- **+: Able to recommend to users with unique tastes**
- **+: Able to recommend new & unpopular items**
 - No first-rater problem
- **+: Able to provide explanations**
 - Can provide explanations of recommended items by listing content-features that caused an item to be recommended

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

17

Cons: Content-based Approach

- **–: Finding the appropriate features is hard**
 - E.g., images, movies, music
- **–: Recommendations for new users**
 - **How to build a user profile?**
- **–: Overspecialization**
 - Never recommends items outside user's content profile
 - People might have multiple interests
 - **Unable to exploit quality judgments of other users**

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

18