

# Collaborative Topic Modeling for Recommending Scientific Articles

Chong Wang and David M. Blei  
Best student paper award at KDD 2011

Computer Science Department, Princeton University

# Outline

- Overview for Recommender Systems
- Methods
  - Collaborative Filtering
  - Topic Modeling
  - Collaborative topic models
- Results
- Conclusions

# Overview for Recommender Systems

- The most widely used Recommender System

# Overview for Recommender Systems

- The most widely used Recommender System

The screenshot displays the Amazon.com homepage. At the top, the Amazon logo is on the left, and the user's name "Hello, Tian Cao" is on the right, along with navigation links for "Today's Deals", "Gifts & Wish Lists", and "Gift Cards". A search bar is positioned below the navigation. On the left side, there is a vertical menu with categories such as "Unlimited Instant Videos", "MP3s & Cloud Player", "Amazon Cloud Drive", "Kindle", "Appstore for Android", "Digital Games & Software", "Audible Audiobooks", "Books", "Movies, Music & Games", "Electronics & Computers", "Home, Garden & Tools", "Grocery, Health & Beauty", "Toys, Kids & Baby", "Clothing, Shoes & Jewelry", "Sports & Outdoors", and "Automotive & Industrial".

The main content area features a banner for "The All-New Kindle Family". Below the banner, three Kindle devices are displayed: the Kindle (price \$79), the Kindle Touch (price \$99), and the Kindle Fire (price \$199). To the right of the Kindle devices, there is a promotional box for "New HDTVs for 2012" and a large "FREE ONE-DAY SHIPPING" banner for "DISCOVER" on select items. Below the shipping banner, there is a link for "Amazon Student Members, Save 50% on a Prime Membership".

# Overview for Recommender Systems

- Type “Digital Camera” in Amazon
- Too many choices to choose from

The screenshot shows the Amazon.com search results page for 'digital camera'. The page is cluttered with various navigation elements and product listings. At the top, there's a search bar with 'digital camera' entered. Below the search bar, there are several navigation tabs: 'All Electronics', 'Brands', 'Best Sellers', 'Audio & Home Theater', 'Camera & Photo', 'Car Electronics & GPS', 'Cell Phones & Accessories', 'Computers', 'MP3 Players', 'TV & Video', and 'Deals'. The left sidebar contains a 'Department' menu with 'Electronics' selected, and a list of sub-categories like 'Camera & Photo', 'Computers & Accessories', etc. The main content area shows the search results for 'digital camera', with a 'Showing 1 - 24 of 470,669 Results' message. The results are sorted by 'Relevance'. The first few results are: 1. Panasonic DMC-FS20K 16.1MP Digital Camera with 8x Wide Angle Image Stabilized Zoom and 2.7 inch LCD (Black) - \$139.99. 2. SanDisk 4GB Secure Digital SD HC Memory Card (S0500-4096, BULK, No Reader) - \$14.99. 3. SanDisk 4GB SDHC card - \$19.99. 4. Canon PowerShot ELPH 300 HS 12.1 MP Digital Camera (Black) - \$179.99. 5. Canon PowerShot SX130HS 12.1 MP Digital Camera with 12x Wide Angle Optical Image Stabilized Zoom with 3.0-inch LCD - \$229.99. 6. Samsung EC-P120 Digital Camera with 14.2 MP and 8x Optical Zoom (Black) - \$99.99. The page also features a 'New in Camera, Photo & Video' banner and a 'Get it by Tuesday, Mar 29' shipping notice.

# What would you do?

- Read every description yourself
- What do other people say

## Avg. Customer Review

★★★★☆ & Up (776)

★★★☆☆ & Up (1,045)

★★☆☆☆ & Up (1,090)

★☆☆☆☆ & Up (1,110)

# What would you do?

- Sorted by Avg. Customer Review

The screenshot shows the Amazon.com search results for 'digital camera'. The page is sorted by 'Avg. Customer Review'. The top navigation bar includes the Amazon logo, a search bar with 'digital camera' entered, and links for 'Your Cart' and 'Your Lists'. Below the navigation bar, there are several filters and sorting options. The main content area displays a grid of six digital camera products, each with a thumbnail image, a title, and pricing information. The products are:

- 1. Panasonic DMC-FZ150K 12.1 MP Digital Camera with CMOS Sensor and 3x Optical Zoom (Black)**  
Listed from \$475.00. In Stock. 4-4-4-4 (154).  
Click for FREE Super Saver Shipping.
- 2. Canon SX40 HS 12.1MP Digital Camera with 30x Wide Angle Optical Image Stabilized Zoom and 3.2-inch Var-Angle Hi-Res LCD**  
Listed from \$475.00. In Stock. 4-4-4-4 (154).
- 3. Nikon COOLPIX P900 16.2 MP CMOS Digital Camera with 4.2x 81x WFOVIR Wide-Angle Optical Zoom Lens and Full HD 1080p Video (Black)**  
Listed from \$239.00. Only 8 left in stock - order soon. 4-4-4-4 (128).  
See newer model of this item.
- 4. Panasonic Lumix DMC-LX5 10.1 MP Digital Camera with 3.8x Optical Image Stabilized Zoom and 3.0-inch LCD - Black**  
Listed from \$499.99 to \$343.39.
- 5. Panasonic Lumix DMC-LX5 10.1 MP Digital Camera with 3.8x Optical Image Stabilized Zoom and 3.0-inch LCD - White**  
Listed from \$499.99 to \$445.31.
- 6. Sony Cyber-Shot DSC-HX100V 16.2 MP Exmor R CMOS Digital 961i Camera with Carl Zeiss Vario-Tessar 30x Optical Zoom Lens and Full HD 1080 Video**  
Listed from \$499.99 to \$425.99.

Additional details visible in the screenshot include the 'Avg. Customer Review' section on the left, which shows a star rating of 4.4 out of 5, and the 'Shipping Option' section, which highlights 'Free Super Saver Shipping'.

## More recommender systems



*and more .....*

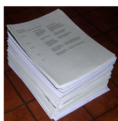
- I am a graduate student and I also do research ...



# This paper focus on Recommending Scientific articles

- A search of “Data Mining” in Google Scholar gives 2,010,000 results.

already read



- If I have read article A, B and C, what should I read next?

# The problem of finding relevant articles

- Finding relevant articles is an important task for researcher

# The problem of finding relevant articles

- Finding relevant articles is an important task for researcher
  - learn about the general idea in an area
  - keep up to the state of art of an area

# The problem of finding relevant articles

- Finding relevant articles is an important task for researcher
  - learn about the general idea in an area
  - keep up to the state of art of an area
- Two popular existing approaches

# The problem of finding relevant articles

- Finding relevant articles is an important task for researcher
  - learn about the general idea in an area
  - keep up to the state of art of an area
- Two popular existing approaches
  - following article references: easily missing relevant citations
  - using keyword search
    - difficult to form queries
    - only good for directed exploration

# The problem of finding relevant articles

- Finding relevant articles is an important task for researcher
  - learn about the general idea in an area
  - keep up to the state of art of an area
- Two popular existing approaches
  - following article references: easily missing relevant citations
  - using keyword search
    - difficult to form queries
    - only good for directed exploration
- The author develop **recommendation algorithms** given online communities sharing referene libraries. ([www.citeulike.org](http://www.citeulike.org))

## Two traditional approaches for recommendation

- Collaborative filtering (CF)
- Topic Modeling
- Combining of the two models

# Collaborative Filtering

Three important elements

- users
- items: article
- ratings: a user likes/dislikes some of the articles

Popular solutions: collaborative filtering (CF)

- matrix factorization: one of the most popular algorithms for recommender system

The user-item matrix

user \ item	1	2	3
1	✓	✗	?
2	✓	?	✗
3	?	✓	✗

 $\Rightarrow$ 
$$\begin{bmatrix} 1 & 0 & ? \\ 1 & ? & 0 \\ ? & 1 & 0 \end{bmatrix}$$



# Matrix factorization

- Users and items are represented in a shared but unknown latent space (latent factor model)
  - user  $i$  –  $u_i \in R^k$
  - item  $j$  –  $v_j \in R^k$
- Each dimension of the latent space is assumed to represent some kind of *unknown factors*
- The rating of item  $j$  by user  $i$  is achieved by the dot product,

$$r_{ij} = u_i^T v_j,$$

where  $r_{ij} = 1$  indicates *like* and 0 *dislike*. In the matrix form,

$$R = U^T V.$$

# Learning and Prediction

- Learning the latent vectors for users and items

$$\min_{U, V} \sum_{i, j} (r_{ij} - u_i^T v_j)^2 + \lambda_u \|u_i\|^2 + \lambda_v \|v_j\|^2,$$

where  $\lambda_u$  and  $\lambda_v$  are regularization parameters.

- Prediction for user  $i$  on item  $j$  (not rated by user  $i$  before),

$$r_{ij} \approx u_i^T v_j.$$

How do we understand these latent vectors for users and items?

# Disadvantages for matrix factorization

Two main disadvantages to matrix factorization for recommendation

- learnt latent space is not easy to interpret
- only uses information from the users-cannot to generalize to completely unrated items

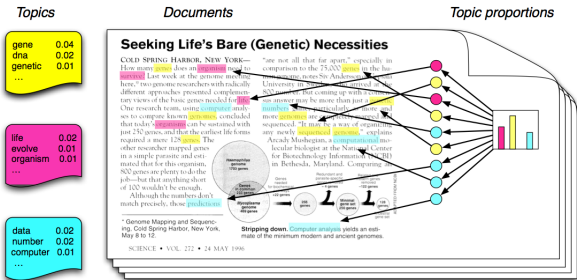
# The author's criteria for an article recommender system

It should be able to

- recommend old articles (already rated, easy)
- recommend new articles (not rated before, not that easy, but doable)
- provide the interpretability - not just a list of items (challenging)

The goal is not only to improve the performance, but also the interpretability.

# Topic modeling



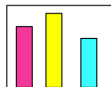
- Each topic is a distribution over words
- Each document is a mixture of topics
- Each word is drawn from one of those topics

From Chong Wang's slides

# Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a popular topic model. It assumes

- There are  $K$  topics
- For each article, topic proportions  $\theta \sim \text{Dirichlet}(\alpha)$



topic proportions  $\theta_j$

gene	0.04
dna	0.02
genetic	0.01
...	

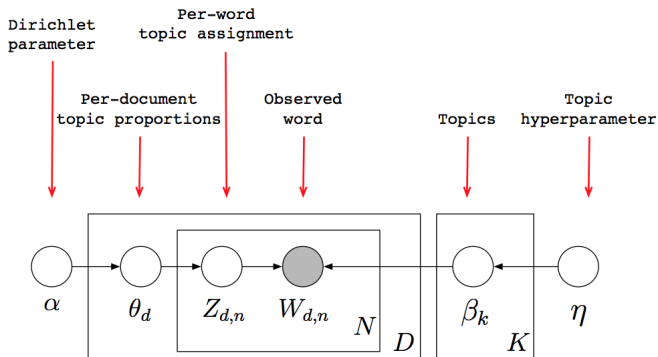
life	0.02
evolve	0.01
organism	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Topics

Note that  $\theta$  can explain the topics that article talks about!

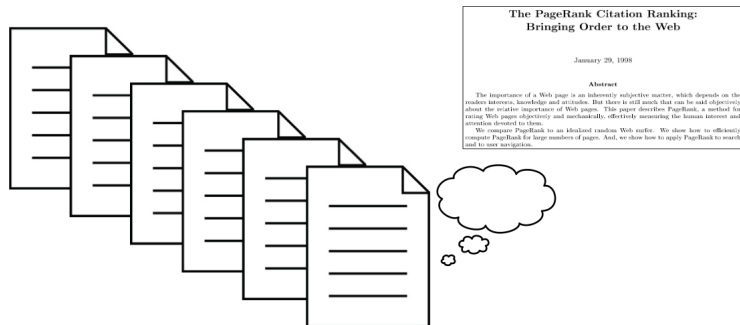
# The graphical model



- Vertices denote random variables
- Edges denote dependence between random variables
- Shading denotes observed variables
- Plates denote replicated variables

From Chong Wang's slides

# Running a topic model



- **Data:** article titles + abstracts from CiteUlike
  - 16,980 articles
  - 1.6M words
  - 8K unique terms
- **Model:** 200-topic LDA model with variational inference



nodes  
wireless  
protocol  
routing  
protocols  
node  
sensor  
peertopeer  
scalable  
hoc

gene  
genes  
expression  
tissues  
regulation  
coexpression  
tissuespecific  
expressed  
tissue  
regulatory

distribution  
random  
probability  
distributions  
sampling  
stochastic  
markov  
density  
estimation  
statistics

learning  
machine  
training  
vector  
learn  
machines  
kernel  
learned  
classifiers  
classifier

relative  
importance  
give  
original  
respect  
obtain  
ranking  
metric  
weighted  
compute

# Inferred topic proportions for article

## Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

*Harvard University and Educational Testing Service*

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 8th, 1976, Professor S. D. SILVEY in the Chair]

### SUMMARY

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.

## topic proportions



estimate estimates likelihood maximum estimated missing



algorithm signal input signals output exact performs music



distribution random probability distributions sampling stochastic

# Comparison of the article representation

## Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

*Harvard University and Educational Testing Service*

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 8th, 1976, Professor S. D. SILVEY in the Chair]

### SUMMARY

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.

## matrix factorization

████████████████████ ??????????????  
████████████████████ ???????????????  
████████████████████ ???????????????

## topic modeling

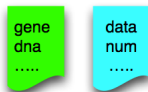
████████████████████ estimate estimates likelihood maximum estimated missing  
████████████████████ algorithm signal input signals output exact performs music  
████████████████████ distribution random probability distributions sampling stochastic

# Collabrative topic models: motivations

## *Article representation in different methods*



matrix factorization

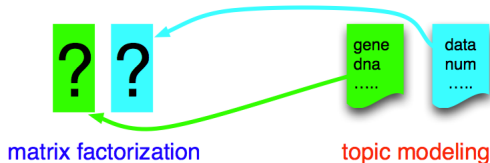


topic modeling

- In matrix factorization, an article has a latent representation  $v$  in some *unknown latent space*
- In topic modeling, an article has topic proportions  $\theta$  in the *learned topic space*

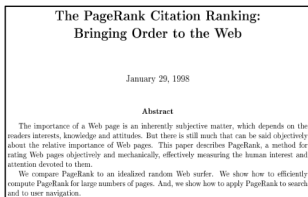
# Collabrative topic models: motivations

*Article representation in different methods*



If we simply fix  $v = \theta$ , we seem to find a way to explain the unknown space using the topic space.

# Collabrative topic models: motivations



what the article is about  
topic proportions

$\theta$



GAP!



what the users think of it  
item latent vector

$v$

The author proposed an approach to fill the gap.

From Chong Wang's slides

# The basic idea

- What the users think of an article might be **different** from what the article is actually about, but **unlikely entirely irrelevant**
- We assume the item latent vector  $v$  is close to topic proportions  $\theta$ , but could diverge from  $\theta$  if it has to

For an article,

- When there are few ratings,  $v_j$  is unlikely to be far from  $\theta_j$
- When there are lots of ratings,  $v_j$  is likely to diverge from  $\theta_j$ . It actually generates or removes some topics to cater the users

## The proposed model

For each user  $i$ ,

- Draw user latent vector  $u_i \sim N(0, \lambda_u^{-1} I_k)$ .

For each article  $j$ ,

- Draw topic proportions  $\theta_j \sim \text{Dirichlet}(\alpha)$ .
- Draw item latent offset  $\epsilon_j \sim N(0, \lambda_v^{-1} I_k)$  and set the item latent vector as  $v_j = \theta_j + \epsilon_j$ .
- Everything else is the same, the rating becomes,

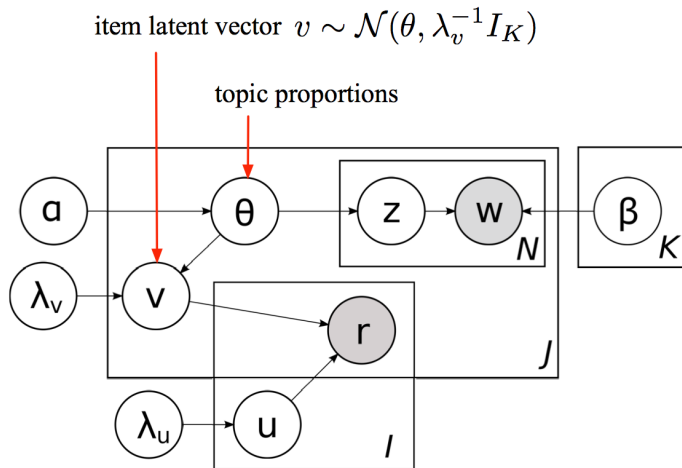
$$E[r_{ij}] = u_i^T v_j = u_i^T (\theta_j + \epsilon_j).$$

This model is called Collaborative Topic Regression (CTR).

- Offset  $\epsilon_j$  corrects  $\theta_j$  for the popularity
- Precision parameter  $\lambda_v$  penalizes how much  $v_j$  could diverge from  $\theta_j$ .



# The graphical model



# Learning and Prediction

- **Learning:** use a standard EM algorithm to learn the maximum a posteriori (MAP) estimates.
- **Prediction:** consider two scenarios,
  - In-matrix prediction: items have been rated before

$$r_{ij}^* \approx (u_i^*)^T (\theta_j^* + \epsilon_j^*).$$

- Out-of-matrix prediction: items have never been rated

$$r_{ij}^* \approx (u_i^*)^T \theta_j^*.$$

user \ article	1	2	3	4	5
1	✓	✗	✓	?	?
2	✓	✓	?	?	✓
3	✗	?	✓	✗	✗
4	?	✓	?	✗	?
5	✗	?	✓	✓	?

(a) in-matrix prediction

user \ article	1	2	3	4	5
1	✓	✗	✓	?	?
2	✓	✓	✗	?	?
3	✗	✗	✓	?	?
4	✗	✓	✓	?	?
5	✗	✓	✓	?	?

(b) out-of-matrix prediction

# Experimental settings

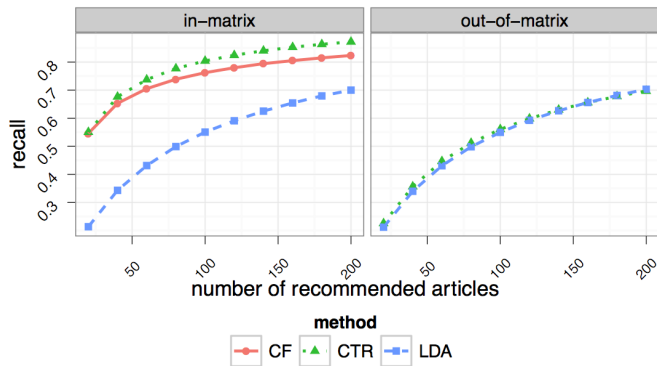
- Data from CiteUlike:
  - 5,551 users, 16,980 articles, and 204,986 bibliography entries. (Sparsity=99.8 %)
  - For each article, concatenate its title and abstract as its content.
  - These articles were added to CiteUlike between 2004 and 2010
- Evaluation: five-fold cross-validation with recall,

$$\text{recall}@M = \frac{\text{number of articles the user likes in top } M}{\text{total number of article the user likes}}$$

- Comparison: matrix factorization for collaborative filter (CF), text-based method (LDA).

# Results

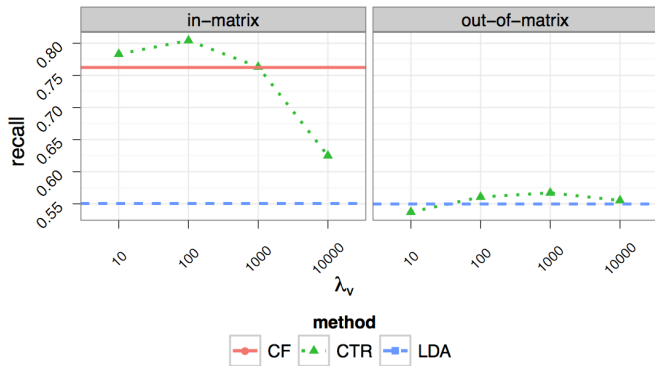
- In-matrix prediction: CTR improves more when number of recommendations gets larger.
- Out-of-matrix prediction: about the same as LDA.



## When precision parameter $\lambda_v$ varies

Recall  $\lambda_v$  penalizes how  $v$  could diverge from  $\theta$ ,

- When  $\lambda_v$  is small, CTR behaves more like CF.
- When  $\lambda_v$  increases, CTR brings in both ratings and content.
- When  $\lambda_v$  is large, CTR behaves more like LDA.



## Interpretation: example user profile I

top topics	<ol style="list-style-type: none"><li>1. image, measure, measures, images, motion, matching</li><li>2. learning, machine, training, vector, learn, machines</li><li>3. sets, objects, defined, categories, representations</li></ol>
top articles	<ol style="list-style-type: none"><li>1. Information theory inference learning algorithms (✓)</li><li>2. Machine learning in automated text categorization (✓)</li><li>3. Artificial intelligence a modern approach (×)</li><li>4. Data mining: practical machine learning tools ... (×)</li><li>5. Statistical learning theory (×)</li><li>6. Modern information retrieval (✓)</li><li>7. Pattern recognition and machine learning (✓)</li><li>8. Recognition by components: a theory of human ... (×)</li><li>9. Data clustering a review (✓)</li><li>10. Indexing by latent semantic analysis (✓)</li></ol>

## Interpretation: example user profile II

top topics	<ol style="list-style-type: none"><li>1. users, user, interface, interfaces, needs, explicit, implicit</li><li>2. based, world, real, characteristics, actual, exploring</li><li>3. evaluation, collaborative, products, filtering, product</li></ol>
top articles	<ol style="list-style-type: none"><li>1. Combining collaborative filtering with personal ... (×)</li><li>2. An adaptive system for the personalized access ... (✓)</li><li>3. Implicit interest indicators (×)</li><li>4. Footprints history-rich tools for information foraging (✓)</li><li>5. Using social tagging to improve social navigation (✓)</li><li>6. User models for adaptive hypermedia and ... (✓)</li><li>7. Collaborative filtering recommender systems (✓)</li><li>8. Knowledge tree: a distributed architecture ... (✓)</li><li>9. Evaluating collaborative filtering recommender ... (✓)</li><li>10. Personalizing search via automated analysis ... (✓)</li></ol>

# Conclusions

- develop an algorithm to recommend scientific articles to users of an online community
- combines the merits of traditional collaborative filtering and probabilistic topic modeling
- provides an interpretable latent structure for users and items
- can form recommendation about both existing and newly published articles