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

Presented by Tian Cao

Outline

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

Overview for Recommender Systems

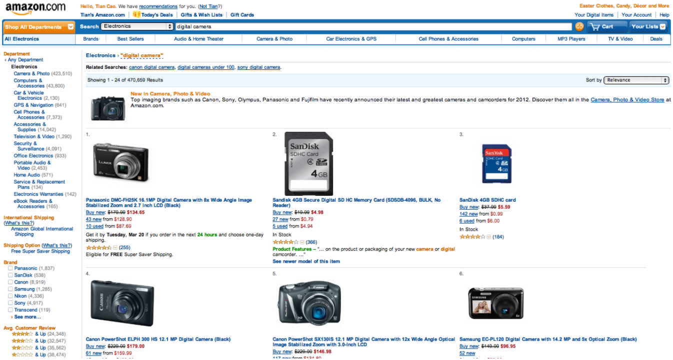
- The most widely used Recommender System

Overview for Recommender Systems

- The most widely used Recommender System

Overview for Recommender Systems

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

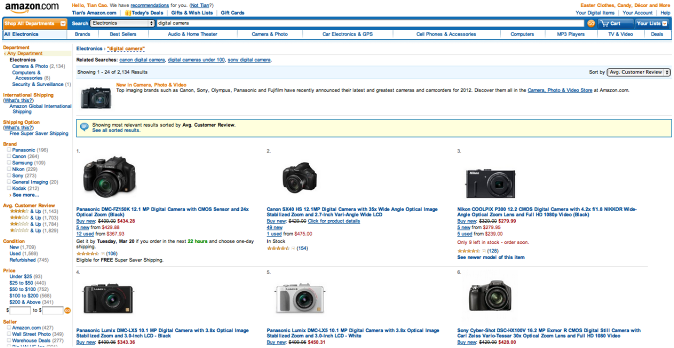


What would you do?

- Read every description yourself
- What do other people say

What would you do?

- Sorted by Avg. Customer Review



More recommender systems

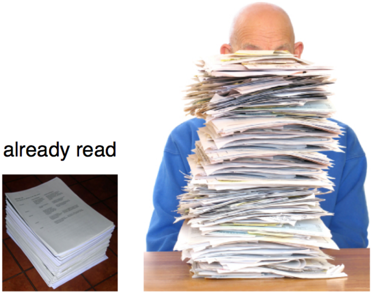
and more .....

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

From Chong Wang's slides

This paper focus on Recommending Scientific artilces

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



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

From Chong Wang's slides

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 exsting 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 exsting 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 exsting 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)

From Chong Wang's slides

Two traditional approaches for recommendation

- Collaborative filtering (CF)
- Topic Modeling
- Combing 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 (lantent 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 λ_u and λ_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 geralize to completely unrated items

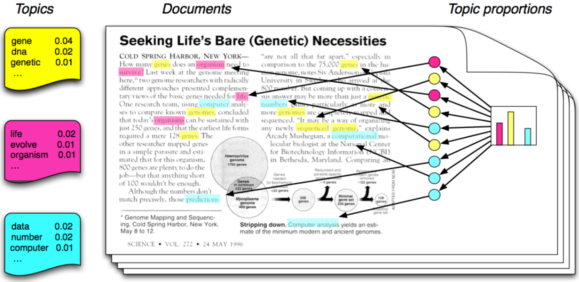
The author's criteria for an article recommender system

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



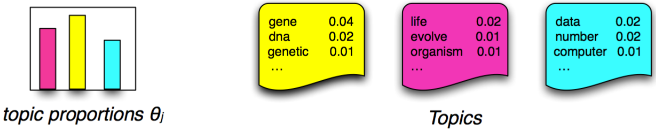
- Each topic is a distribution over words
- Each document is a mixture of topics
- Each word is drawn from one of those topics

From Chong Wang's slides

Latent Dirichlet allcation

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

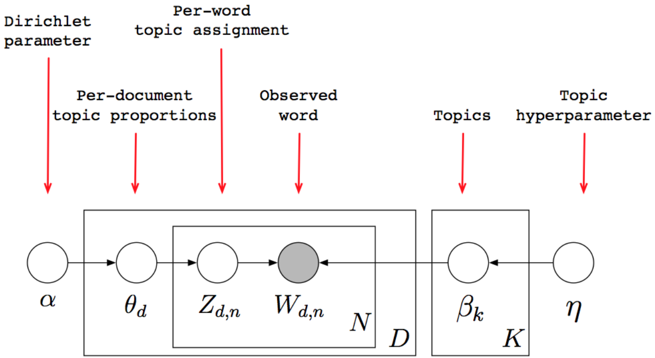
- There are K topics
- For each article, topic proportions $\theta \sim \textit{Dirichlet}(\alpha)$



Note that θ can explain the topics that article talks about!

From Chong Wang's slides

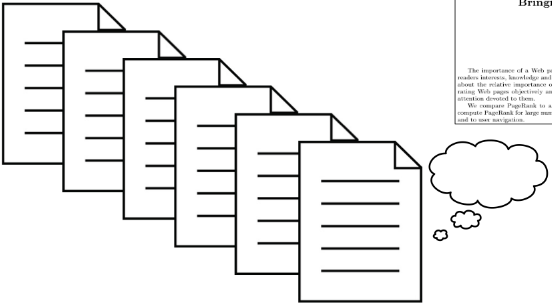
The graphical model

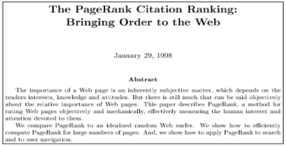


- Vertices denote random variables
- Edges denote dependence between random variables
- Shading denotes observed variables
- Plates denote replicated variables

From Chong Wang's slides

Running a topic model





**The PageRank Citation Ranking:
Bringing Order to the Web**

January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

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

nodes	gene	distribution	learning	relative
wireless	genes	random	machine	importance
protocol	expression	probability	training	give
routing	tissues	distributions	vector	original
protocols	regulation	sampling	learn	respect
node	coexpression	stochastic	machines	obtain
sensor	tissuespecific	markov	kernel	ranking
peertopeer	expressed	density	learned	metric
scalable	tissue	estimation	classifiers	weighted
hoc	regulatory	statistics	classifier	compute

Inferred topic propostions 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

????????????

????????????

????????????

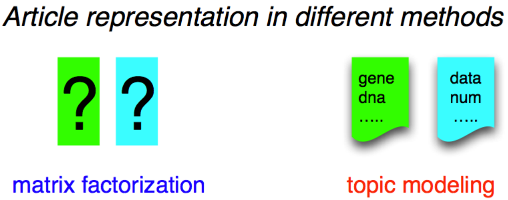
topic modeling

estimate estimates likelihood maximum estimated missing

algorithm signal input signals output exact performs music

distribution random probability distributions sampling stochastic

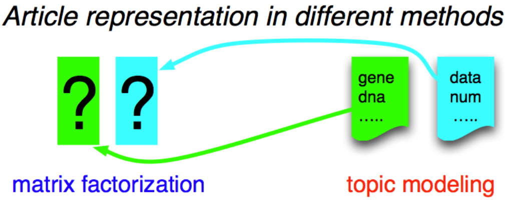
Collabrative topic models: motivations



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

From Chong Wang's slides

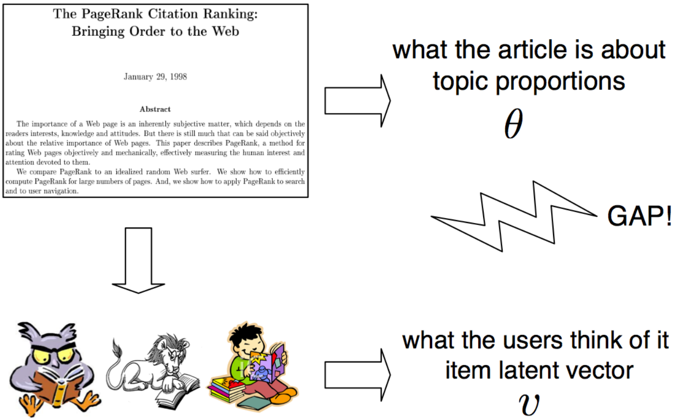
Collabrative topic models: motivations



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

From Chong Wang's slides

Collabrative topic models: motivations



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 irreleant**
- We assume the item latent vector v is close to topic propotions θ, but could diverge from θ if it has to

For an article,

- When there are few ratings, v_j is unlikely to be far from θ_j
- When there are lots of ratings, v_j is likely to diverge from θ_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,

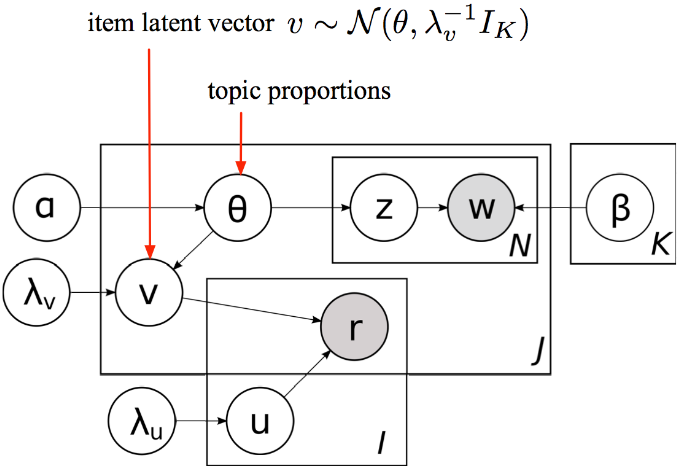
- Draw topic proportions $\theta_i \sim Dirichlet(\alpha)$.
- Draw item latent offset $\epsilon_j \sim N(0, \lambda_v^{-1} I_k)$ and set the item latent vector as $v_j = \theta_j + \epsilon_j$.
- Everything else is the same, the rating becomes,

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

This model is called Collaborative Topic Regression (CTR).

- Offset ϵ_j corrects θ_j for the popularity
- Precision parameter λ_v penalizes how much v_j could diverge from θ_j.

The graphical model



From Chong Wang's slides

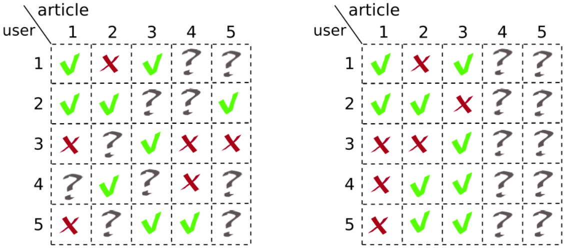
Learning and Prediction

- **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

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

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

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



(a) in-matrix prediction (b) out-of-matrix prediction

Experimental settings

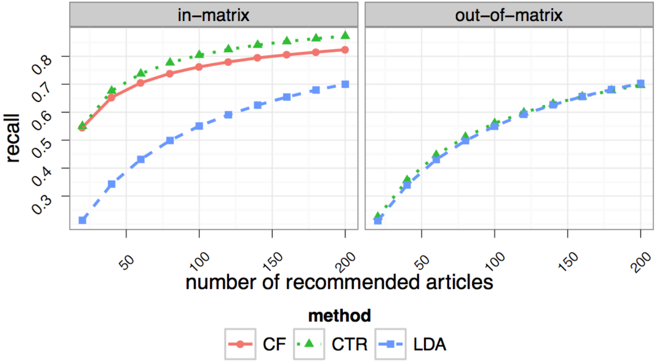
- Data from CiteUlike:
  - 5,551 users, 16,980 articles, and 204,986 bibliography entries. (Sparsity=99.8 %)
  - For each article, concatenate its title and abstract as its content.
  - These articles were added to CiteUlike between 2004 and 2010
- Evaluation: five-fold cross-validation with recall,

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

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

Results

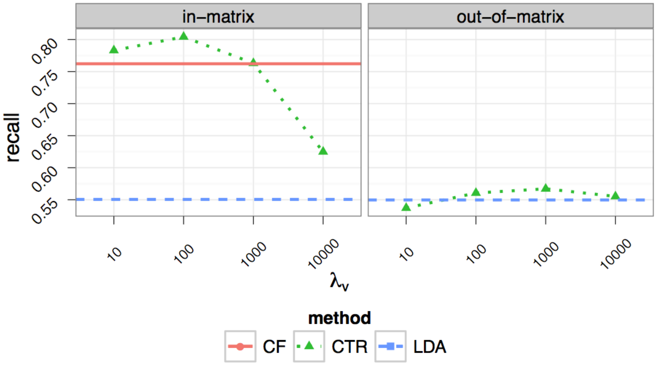
- In-matrix prediction: CTR improves more when number of recommendations gets larger.
- Out-of-matrix prediction: about the same as LDA.



When precision parameter λ_v varies

Recall λ_v penalizes how v could diverge from θ,

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



Interpretation: example user profile I

top topics	1. image, measure, measures, images, motion, matching 2. learning, machine, training, vector, learn, machines 3. sets, objects, defined, categories, representations
top articles	1. Information theory inference learning algorithms (✓) 2. Machine learning in automated text categorization (✓) 3. Artificial intelligence a modern approach (×) 4. Data mining: practical machine learning tools ... (×) 5. Statistical learning theory (×) 6. Modern information retrieval (✓) 7. Pattern recognition and machine learning (✓) 8. Recognition by components: a theory of human ... (×) 9. Data clustering a review (✓) 10. Indexing by latent semantic analysis (✓)

Interpretation: example user profile II

top topics	1. users, user, interface, interfaces, needs, explicit, implicit 2. based, world, real, characteristics, actual, exploring 3. evaluation, collaborative, products, filtering, product
top articles	1. Combining collaborative filtering with personal . . . (×) 2. An adaptive system for the personalized access . . . (✓) 3. Implicit interest indicators (×) 4. Footprints history-rich tools for information foraging (✓) 5. Using social tagging to improve social navigation (✓) 6. User models for adaptive hypermedia and . . . (✓) 7. Collaborative filtering recommender systems (✓) 8. Knowledge tree: a distributed architecture . . . (✓) 9. Evaluating collaborative filtering recommender . . . (✓) 10. Personalizing search via automated analysis . . . (✓)

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