AN ARTICLE RECOMMENDATION SYSTEM BASED ON COLLABORATIVE TOPIC MODELING

Dang Fan, Wang Shuhao, Ding Peng

~~June 4~~, 2012

Overview

- Modeling
- Implementation
- Characteristics and performance

Modeling

Recommendations tasks

Collaborative Filtering

- User i: $u_i \in \mathbb{R}^K$
- Item j: $v_j \in \mathbb{R}^K$
- $\hat{r}_{ij} = u_i^T v_j$

Modeling

Recommendations tasks

Collaborative Filtering

- User i: $u_i \in \mathbb{R}^K$
- Item j: $v_j \in \mathbb{R}^K$
- $\hat{r}_{ij} = u_i^T v_j$

Probabilistic Matrix Factorization

- $u_i \sim \mathcal{N}(0, \lambda_u^{-1} I_K)$
- $v_j \sim \mathcal{N}(0, \lambda_v^{-1} I_K)$
- $r_{ij} \sim \mathcal{N}(u_i^T v_j, c_{ij}^{-1})$

Modeling

Recommendations tasks

Collaborative Filtering

- User i: $u_i \in \mathbb{R}^K$
- Item j: $v_j \in \mathbb{R}^K$
- $\hat{r}_{ij} = u_i^T v_j$

Probabilistic Matrix Factorization

- $u_i \sim \mathcal{N}(0, \lambda_u^{-1} I_K)$
- $v_j \sim \mathcal{N}(0, \lambda_v^{-1} I_K)$
- $r_{ij} \sim \mathcal{N}(u_i^T v_j, c_{ij}^{-1})$

Probabilistic Topic Models

Latent Dirichlet Allocation

- K topics $\beta = \beta_{1:K}$
- For each article w_j
  1. Topic distribution $\theta_j \sim$ Dirichlet(α)
  2. For each word n
     - Topic $z_{jn} \sim$ Multinomial(θ_j)
     - Word $w_{jn} \sim$ Multinomial($\beta_{z_{jn}}$)

Modeling

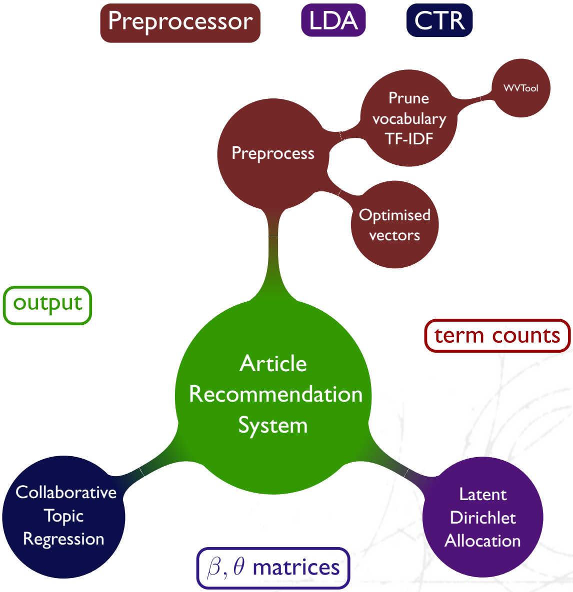
A hybrid approach

Collaborative Topic Regression

1. For each user i, $u_i \sim \mathcal{N}(0, \lambda_u^{-1} I_K)$
2. For each item j,
    a) Topic distribution $\theta_j \sim$ Dirichlet(α)
    b) Item latent offset $\epsilon_j \sim \mathcal{N}(0, \lambda_v^{-1} I_K)$
       $v_j = \epsilon_j + \theta_j$
    c) For each word w_{jn},
        i. Topic $z_{jn} \sim$ Multinomial(θ_j)
        ii. Word $w_{jn} \sim$ Multinomial($\beta_{z_{jn}}$)
3. For each user-item pair (i, j), $r_{ij} \sim \mathcal{N}(u_i^T v_j, c_{ij}^{-1})$

Implementation

The framework



Implementation

Implementation of LDA

Expectation-Maximization Algorithm

- E Step: Calculate the expected value of the log likelihood function
- M Step: Find the parameter that maximizes this quantity

An open-source Java project `lda-j` can be found on the Internet.

Implementation

Implementation of CTR

EM-style algorithm

$$\mathcal{L} = -\frac{\lambda_u}{2}\Sigma_i u_i^T u_i - \frac{\lambda_v}{2}\Sigma_j (v_j - \theta_j)^T (v_j - \theta_j)$$

$$+\Sigma_j \Sigma_n \log(\Sigma_k \theta_{jk}\beta_{k,w_{jn}}) - \Sigma_{i,j}\frac{c_{ij}}{2}(r_{ij} - u_i^T v_j)^2$$

Characteristics

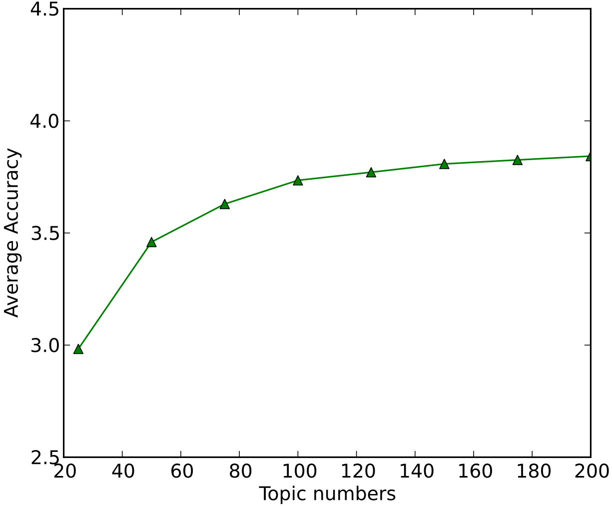
- A hybrid approach
- Good performance

Performance

K(Topics)	Accuracy(of 5)	AP@5	Running Time
25	2.9831	0.7435	40min
50	3.4597	0.8296	54min
75	3.6293	0.8545	1.8h
100	3.7348	0.8717	3.5h
125	3.7709	0.8794	4.3h
150	3.8080	0.8812	7.0h
175	3.8258	0.8873	9.8h
200	3.8427	0.8881	13.2h

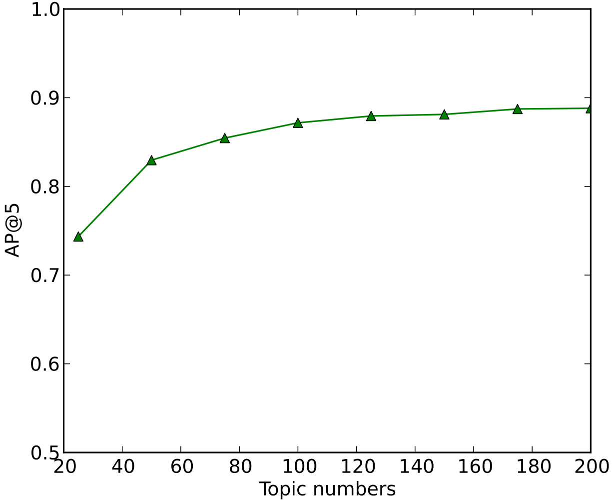
Performance

Accuracy



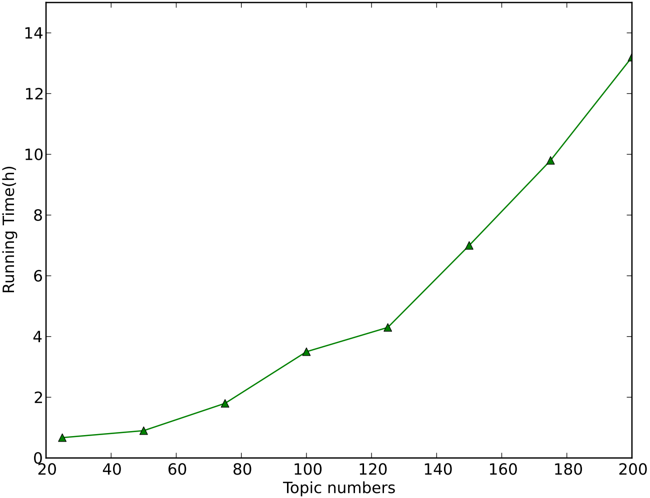
Performance

AP@5



Performance

Running Time



Thanks