Master Course in Data Mining and Knowledge Management
Internship Report Presentation

Cold Start Recommendations: A Non-negative Matrix Factorization Approach

Martin SAVESKI

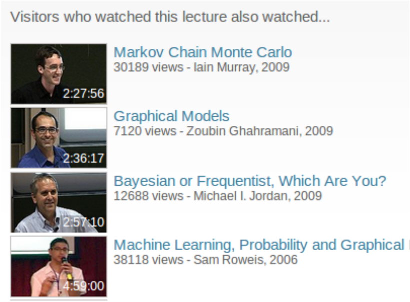
Yahoo! Labs

July 9, 2013

Supervisors: Amin Mantrach, Yahoo! Labs
Daniele Quercia, Yahoo! Labs
Ricard Gavaldà, UPC

Recommender Systems

- Designed to suggest users items of interest
- Widely spread: books, movies, video lectures ...



Two Main Types of Recommender Systems

Content-Based (CB):

- Use the properties of the items
- Build user profiles from the properties of the items liked before
- Suggest items to users with most similar profiles

Collaborative filtering (CF):

- Use the preferences of the community (wisdom of the crowd)
- Recommend items that users with similar tastes like

Two Main Types of Recommender Systems

Content-Based (CB):

- Use the properties of the items
- Build user profiles from the properties of the items liked before
- Suggest items to users with most similar profiles

Collaborative filtering (CF):

- Use the preferences of the community (wisdom of the crowd)
- Recommend items that users with similar tastes like

CF systems achieve the state-of-the-art performance
(Tuzhilin & Adomavicius)

The Cold Start

When a new item/user enters the system and no past information is available, no effective recommendations can be produced

The Cold Start

When a new item/user enters the system and no past information is available, no effective recommendations can be produced

Item cold-start

- No previous ratings are available
- Collaborative filtering is not an option

Common solution: Back-off to content-based recommendations

User cold-start

- Visits from users who are not logged in
- Neither Content-Based, nor Collaborative Filtering is applicable

Common solution: Recommend the top-k items (one-size-fits-all)

The Cold Start

When a new item/user enters the system and no past information is available, no effective recommendations can be produced

Item cold-start

- No previous ratings are available
- Collaborative filtering is not an option

Common solution: Back-off to content-based recommendations

User cold-start

- Visits from users who are not logged in
- Neither Content-Based, nor Collaborative Filtering is applicable

Common solution: Recommend the top-k items (one-size-fits-all)

Practical Importance:

Hundreds of new items and millions of visitors every day

The Cold Start: Challenges

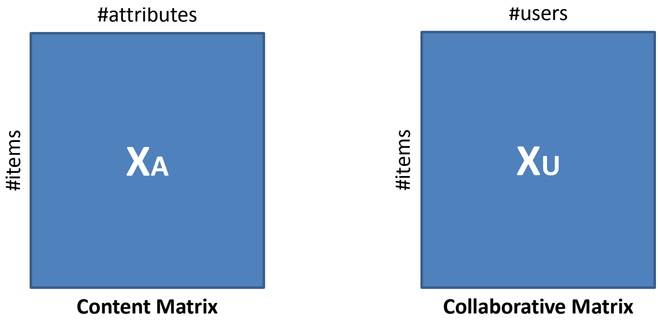
Research Questions:

Item cold-start Can we combine the content and the collaborative information to outperform pure content-based recommenders?

User cold-start Can we use the context of the users, *i.e.*, location and time to perform better then the top-k recommendations?

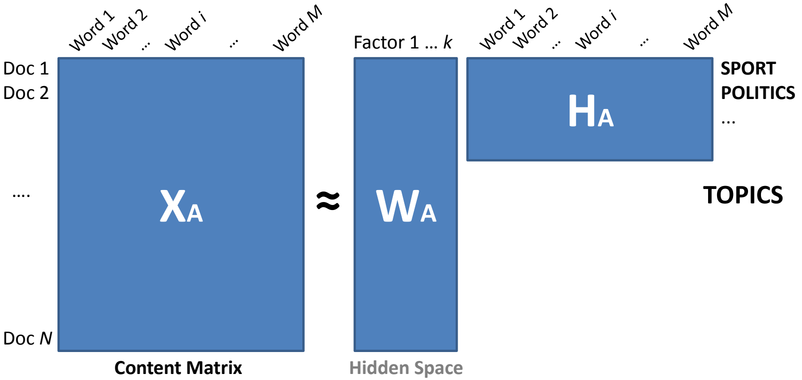
Item Cold-start

Data in Matrix Form



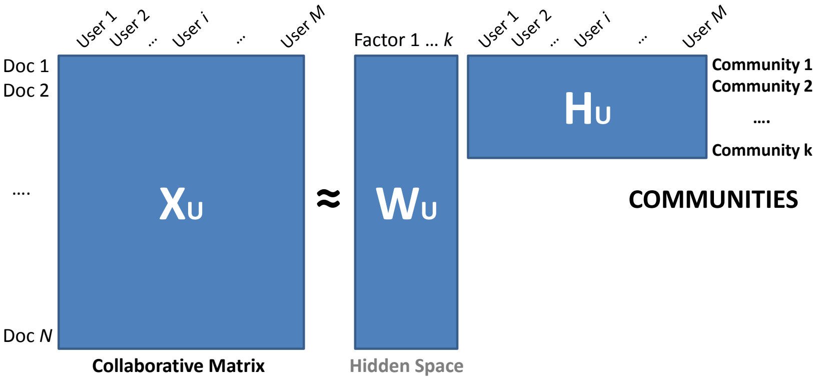
- Both are non-negative matrices
- New item corresponds to an empty row in $\mathbf{X_U}$

Non-negative Matrix Factorization (NMF)



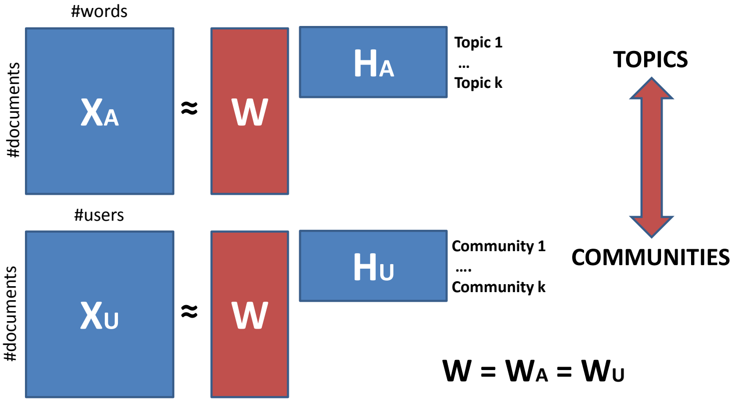
- NMF decomposes $\mathbf{X_A}$ in two lower-rank matrices $\mathbf{W_A}$ and $\mathbf{H_A}$

Non-negative Matrix Factorization (NMF)



- Many CF approaches perform UV decomposition of $\mathbf{X_U}$ (NMF without non-negativity constraints)

Joint Non-negative Matrix Factorization (JNMF)



At test time, given the properties of a new item, q_A:

- Project q_A in w, using $\mathbf{H_A}$
- Infer q_u as: $q_u = w\mathbf{H_U}$

JNMF: Optimization

Non-convex optimization problem:

$$\min : J = \frac{1}{2}(\alpha \underbrace{||\mathbf{X_A} - \mathbf{WH_A}||_F^2}_{\text{Factorization of } \mathbf{X_A}} + (1-\alpha) \underbrace{||\mathbf{X_U} - \mathbf{WH_U}||_F^2}_{\text{Factorization of } \mathbf{X_U}} + \lambda \underbrace{(||\mathbf{W}||_F^2 + ||\mathbf{H_A}||_F^2 + ||\mathbf{H_U}||_F^2)}_{\text{Tikhonov Regularization}})$$

$$s.t. \quad \mathbf{W} \geq 0, \mathbf{H_A} \geq 0, \mathbf{H_U} \geq 0$$

Hyper-parameters:

- k: number of factors
- α: importance of each factorization
- λ: smoothness of the solution

Optimization Algorithms:

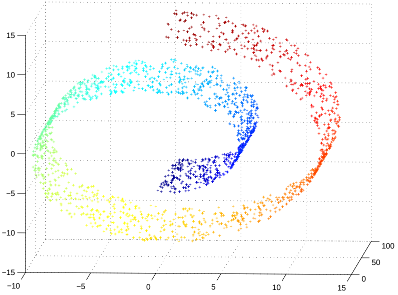
1. Multiplicative Update Rules
2. Alternating Least Squares with projections to 0

Joint NMF with Graph Regularization (JNMF-GR)

Assumption: if two points x_i and x_j are close in the NN graph, then their latent representations w_i and w_j should also be close.

$$S = \frac{1}{2} \sum_{i,j=1}^{n} ||w_i - w_j||^2 [\mathbf{NN}]_{ij} = \mathrm{trace}(\mathbf{W}^T \mathbf{L} \mathbf{W})$$

Controlled with additional hyper-parameter



Evaluation: Experimental Setup

Tasks:

- Item cold-start (implicit feedback)
- Email recipient prediction
- Author prediction

Datasets:

	Type	#documents	#users	|vocabulary|	Span
Yahoo! News	articles&comments	41K	650K	60K	40 days
Enron	email&recipients	36K	5K	12K-56K	10 mailboxes
NIPS	papers&authors	1.7K	2K	13K	13 years

Evaluation: Experimental Setup (cont.)

Baselines:

- Content Based Recommender (CB)
- Content Topic Based Recommender (CTB)
- Author Topic Model (ATM) - learns users' topical profiles

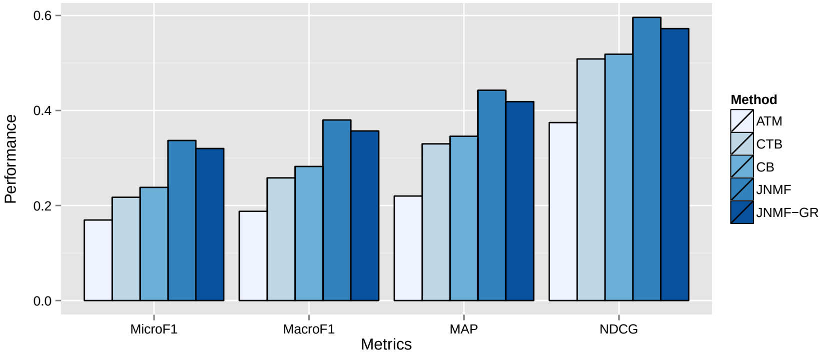
Metrics:

- Macro and Micro F1
- Mean Average Precision (MAP)
- Normalized Discounted Cumulative Gain (NDCG)
- Ranking accuracy (recall oriented, suitable for implicit feedback)

Protocol:

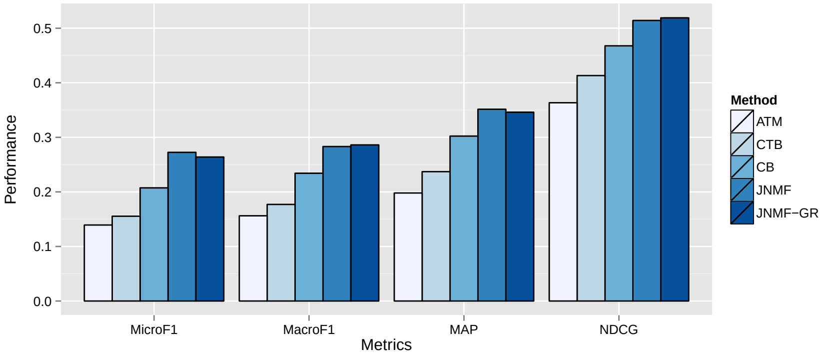
- Divide the data sets in chronological train/test folds
- Hyper-parameters tuned on independent set (part of train set)
- Differences evaluated with paired *t*-test

Results: Author Prediction (NIPS)



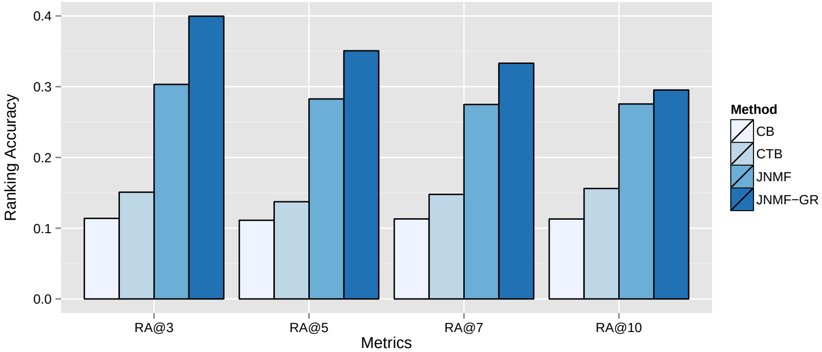
$$\text{JNMF} \approx \text{JNMF-GR} > \text{CB} > \text{CTB} > \text{ATM}$$

Results: Email Recipient Prediction (Enron)



$$\text{JNMF} \approx \text{JNMF-GR} > \text{CB} > \text{CTB} > \text{ATM}$$

Results: Item cold-start (Yahoo! News)



JNMF-GR > JNMF > CTB > CB

User Cold-start

User cold-start

No user history available $\Rightarrow$ top-k recommendations

User cold-start

No user history available $\Rightarrow$ top-k recommendations

User context (IP address derived):

- Location
- Time

User cold-start

No user history available $\Rightarrow$ top-k recommendations

User context (IP address derived):

- Location
- Time

Geography of user engagement in news platforms

- Specific locations $\Leftrightarrow$ specific interests?
- Use these patterns to partially overcome the user cold-start?

User cold-start

No user history available $\Rightarrow$ top-k recommendations

User context (IP address derived):

- Location
- Time

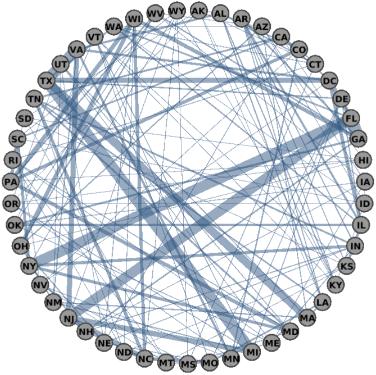
Geography of user engagement in news platforms

- Specific locations $\Leftrightarrow$ specific interests?
- Use these patterns to partially overcome the user cold-start?

Yahoo! News US

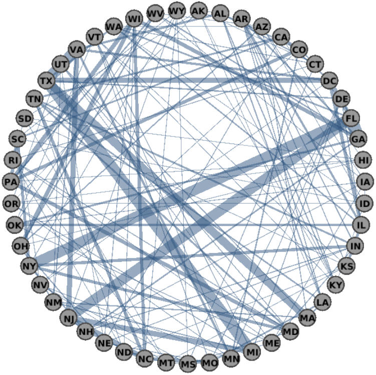
- 2 years
- 200K articles
- 41M comments

State Commenting Graph

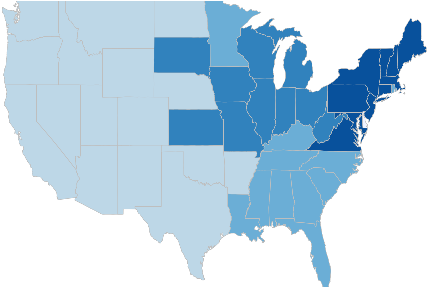


weight(i, j): # of times users from state i and j commented on the same article

State Commenting Graph



weight(i,j): # of times users from state i and j commented on the same article



Four clusters of like-minded states

The Time Zone Effect

Hypothesis:

Users in the same time zone preferentially engage with the same articles

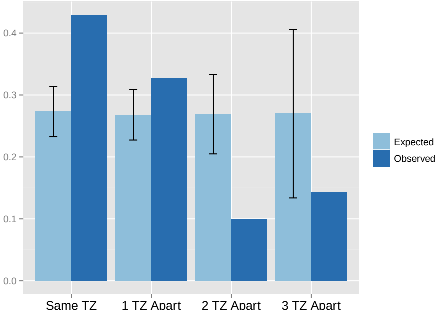
The Time Zone Effect

Hypothesis:

Users in the same time zone preferentially engage with the same articles

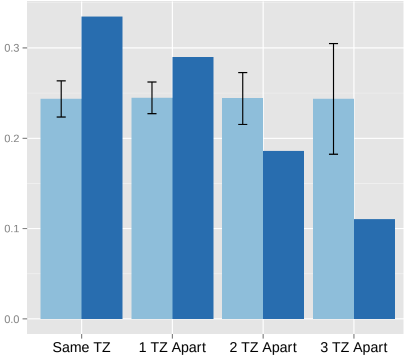
Test:

1. Measuring engagement k time zones apart
2. Random model: shuffling the time zone assignments
3. Comparison:



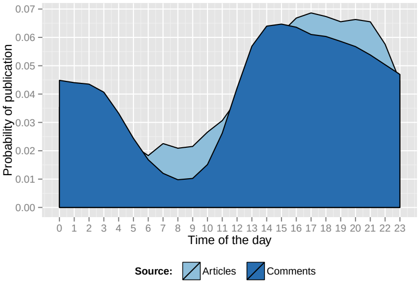
What Makes the Time Zone Effect

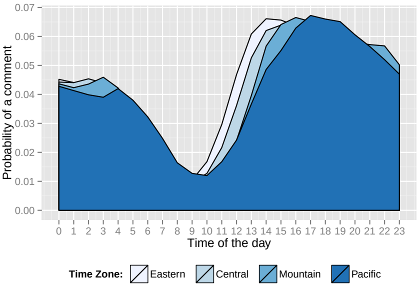
1. Interests in different topics
   - Users in the same time zone are interested in the same topics



What Makes the Time Zone Effect

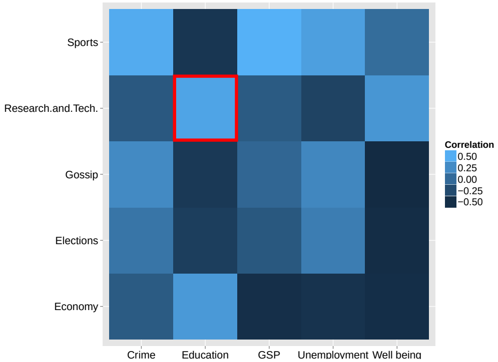
2. Better alignment with the publishing cycle
   - Commenting cycles of some time zones meet the publishing cycle





What Makes the Time Zone Effect

3. Differences in the socio-economical status of the users
   - E.g. Education ⇒ Research&Technology



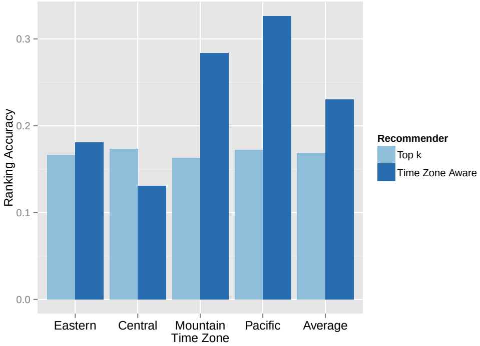
Time-zone-aware recommender for user cold-start

Recommend what is most popular in user's time zone

Evaluation: top-k vs. time-zone-aware recommendations

Subset of 1 month:

- >6M comments
- >35K articles
- >500K users



To sum up

Item cold-start

- Proposed two hybrid recommenders based on NMF
- Combining content + collaborative information outperforms pure content based recommendations

User cold-start

- Studied the geography of user engagement in news platforms
- Making time-zone-aware recommendations leads to improved recommendation accuracy

Future Work

Time aware extension of the hybrid models

- Our tastes change over time

Future Work

Time aware extension of the hybrid models

- Our tastes change over time

Exploiting power law behaviours

- Both words and users are power law distributed

Future Work

Time aware extension of the hybrid models

- Our tastes change over time

Exploiting power law behaviours

- Both words and users are power law distributed

Decoupling the geographic locality effect

- How much the time zone effect is solely due to time zones?

Future Work

Time aware extension of the hybrid models

- Our tastes change over time

Exploiting power law behaviours

- Both words and users are power law distributed

Decoupling the geographic locality effect

- How much the time zone effect is solely due to time zones?

References

- Alexander Tuzhilin & Gediminas Adomavicius. *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, 2005.

Cold Start Recommendations:
A Non-negative Matrix Factorization Approach

Martin SAVESKI

Yahoo! Labs

July 9, 2013

Related Work

Hybrid recommender: Content-Boosted Collaborative Filtering

- Enriches the collaborative profiles by adding content-based scores
- In cold-start boils down to content-based recommendations

Author Topic Model:

- Extension of LDA
- Learns topical profiles for each author

We compare with the ATM in the evaluation

Joint NMF: Multiplicative Update Rules

$$
\begin{aligned}
[\mathbf{W}]_{ij} &\leftarrow [\mathbf{W}]_{ij} \cdot \frac{[\alpha \mathbf{X_A} \mathbf{H_s}^\top + (1-\alpha)\mathbf{X_U}\mathbf{H_u}^\top]_{ij}}{[\alpha \mathbf{W}\mathbf{H_A}\mathbf{H_s}^\top + (1-\alpha)\mathbf{W}\mathbf{H_U}\mathbf{H_u}^\top + \lambda_W \mathbf{W}]_{ij}}, \\
[\mathbf{H_A}]_{ij} &\leftarrow [\mathbf{H_A}]_{ij} \cdot \frac{[\alpha \mathbf{W}^\top \mathbf{X_A}]_{ij}}{[\alpha \mathbf{W}^\top \mathbf{W}\mathbf{H_A} + \lambda_{H_s} \mathbf{H_A}]_{ij}}, \\
[\mathbf{H_U}]_{ij} &\leftarrow [\mathbf{H_U}]_{ij} \cdot \frac{[(1-\alpha)\mathbf{W}^\top \mathbf{X_U}]_{ij}}{[(1-\alpha)\mathbf{W}^\top \mathbf{W}\mathbf{H_U} + \lambda_{H_u} \mathbf{H_U}]_{ij}}.
\end{aligned}
$$

Joint NMF: Alternating Least Squares

Require: k, α, λ_W, λ_{H_s}, λ_{H_u}

Initialize $\mathbf{W}$ with random positive values

repeat

Solve for $\mathbf{H_A}$: $\alpha \mathbf{W}^\mathsf{T} \mathbf{X_A} = \left(\alpha \mathbf{W}^\mathsf{T} \mathbf{W} + \lambda_{H_s} \mathbf{I}\right) \mathbf{H_A}$

Set all negative elements in $\mathbf{H_A}$ to 0

Solve for $\mathbf{H_U}$: $(1-\alpha) \mathbf{W}^\mathsf{T} \mathbf{X_U} = \left[(1-\alpha) \mathbf{W}^\mathsf{T} \mathbf{W} + \lambda_{H_u} \mathbf{I}\right] \mathbf{H_U}$

Set all negative elements in $\mathbf{H_U}$ to 0

Solve for $\mathbf{W}$: $\alpha \mathbf{X_A} \mathbf{H_s}^\mathsf{T} + (1-\alpha) \mathbf{X_U} \mathbf{H_u}^\mathsf{T} =$
$\mathbf{W} \left[\alpha \mathbf{H_A} \mathbf{H_s}^\mathsf{T} + (1-\alpha) \mathbf{H_U} \mathbf{H_u}^\mathsf{T} + \lambda_W \mathbf{I}\right)$

Set all negative elements in $\mathbf{W}$ to 0

until stopping condition

JNMF with Graph Regularization: Details

New optimization problem:

$$\min : J = \frac{1}{2}(\alpha||\mathbf{X_A} - \mathbf{WH_A}||_F^2 + (1-\alpha)||\mathbf{X_U} - \mathbf{WH_U}||_F^2 + \beta \mathrm{trace}(\mathbf{W}^T\mathbf{LW}) + \lambda(||\mathbf{W}||_F^2 + ||\mathbf{H_A}||_F^2 + ||\mathbf{H_U}||_F^2)$$

$$s.t. \quad \mathbf{W} \geq 0, \mathbf{H_A} \geq 0, \mathbf{H_U} \geq 0$$

Additional hyper-parameters:

- β: extent to which locality is imposed
- p: number of nearest neighbours

Optimization Algorithms:

1. Multiplicative Update Rules
2. Alternating Least Squares with projections to 0
   – Updating **W** requires solving Sylvester Equation

JNMF-GR: Multiplicative Update Rules

$$
\begin{aligned}
[\mathbf{W}]_{ij} &\leftarrow [\mathbf{W}]_{ij} \cdot \frac{[\alpha \mathbf{X_A} \mathbf{H_s}^\mathsf{T} + (1-\alpha)\mathbf{X_U}\mathbf{H_u}^\mathsf{T} + \beta \mathbf{AW}]_{ij}}{[\alpha \mathbf{W}\mathbf{H_A}\mathbf{H_s}^\mathsf{T} + (1-\alpha)\mathbf{W}\mathbf{H_U}\mathbf{H_u}^\mathsf{T} + \beta \mathbf{DW} + \lambda_W \mathbf{W}]_{ij}} \\
[\mathbf{H_A}]_{ij} &\leftarrow [\mathbf{H_A}]_{ij} \cdot \frac{[\alpha \mathbf{W}^\mathsf{T}\mathbf{X_A}]_{ij}}{[\alpha \mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{H_A} + \lambda_{H_s}\mathbf{H_A}]_{ij}} \\
[\mathbf{H_U}]_{ij} &\leftarrow [\mathbf{H_U}]_{ij} \cdot \frac{[(1-\alpha)\mathbf{W}^\mathsf{T}\mathbf{X_U}]_{ij}}{[(1-\alpha)\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{H_U} + \lambda_{H_u}\mathbf{H_U}]_{ij}}
\end{aligned}
$$

JNMF-GR: Alternating Least Squares

Require: $\mathbf{A}$, k, α, β, λ_W, λ_{H_s}, λ_{H_u}

Compute the diagonal matrix $\mathbf{D}$ as: $\mathbf{D}_{ii} = \sum_i \mathbf{A}_{ij}$
Compute the Laplacian matrix $\mathbf{L}$ as: $\mathbf{L} = \mathbf{D} - \mathbf{A}$
Initialize $\mathbf{W}$ with random positive values
repeat
  Solve for $\mathbf{H_A}$: $\alpha \mathbf{W}^\mathsf{T} \mathbf{X_A} = (\alpha \mathbf{W}^\mathsf{T} \mathbf{W} + \lambda_{H_s} \mathbf{I})\ \mathbf{H_A}$
  Set all negative elements in $\mathbf{H_A}$ to 0
  Solve for $\mathbf{H_U}$: $(1-\alpha)\mathbf{W}^\mathsf{T} \mathbf{X_U} = [(1-\alpha)\mathbf{W}^\mathsf{T} \mathbf{W} + \lambda_{H_u} \mathbf{I}]\ \mathbf{H_U}$
  Set all negative elements in $\mathbf{H_U}$ to 0
  $\mathbf{C} = \alpha \mathbf{X_A} \mathbf{H_s}^\mathsf{T} + (1-\alpha) \mathbf{X_U} \mathbf{H_u}^\mathsf{T}$
  $\mathbf{B} = \alpha(\mathbf{H_A} \mathbf{H_s}^\mathsf{T} \otimes \mathbf{I}) + (1-\alpha)(\mathbf{H_U} \mathbf{H_u}^\mathsf{T} \otimes \mathbf{I}) + \beta(\mathbf{I} \otimes \mathbf{L}) + \lambda_W \mathbf{I}$
  Solve for $\mathrm{vec}(\mathbf{W})$: $\mathrm{vec}(\mathbf{C}) = \mathbf{B} \cdot \mathrm{vec}(\mathbf{W})$
  $\mathbf{W}$: $\mathbf{W} = \mathit{reshape}(\mathrm{vec}(\mathbf{W}), n, k)$
  Set all negative elements in $\mathbf{W}$ to 0
until stopping condition

Percentile-ranking/Ranking accuracy

Recall based measure and is suitable when no reliable feedback:

- $rank_{u,i}$ is 0%, if most desirable article for u
- $rank_{u,i}$ is 100%, if least least desirable

$$\overline{rank} = \frac{\sum_{u,i} comment_{u,i} \cdot rank_{u,i}}{\sum_{u,i} comment_{u,i}}$$

We covert the percentile ranking into ranking accuracy:

$$accuracy = \frac{50\% - \overline{rank}}{50\%}$$

- 1: best/ideal predictions
- 0: random predictions

Normalized Discounted Cumulative Gain (NDCG)

Based on two assumptions:

1. highly relevant items are more useful
2. the lower the ranked position of a relevant item the less useful it is for the user

$$NDCG = \frac{DCG}{iDCG}, \qquad DCG = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i},$$

- rel_i is the relevance of the document at rank i
- $iDCG$ is the ideal DCG- the prefect ranking of the results

Joint NMF with different hyper-parameters

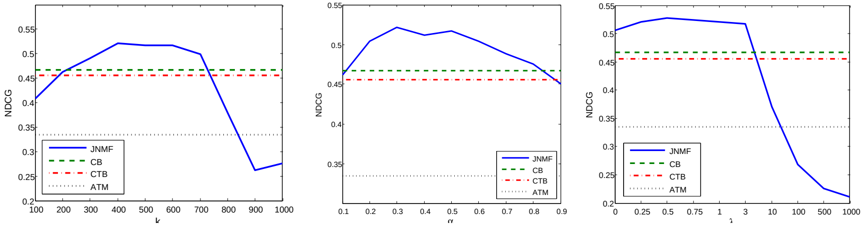


Figure: Behavior of the JNMF hyper-parameters.

Time Zone Effect. Time Advantage

- Quantifying the advantage A a time zone might have within the next n hours
- The probability of commenting in that time zone given the availability of articles:

$$A = \sum_{i=0}^{n} \sum_{t=0}^{23} (p_t^a \cdot p_{t+i}^c) \ ,$$

- p_t^a is the probability of a new article being published at time t (with $t \in [0, 23]$)
- p_{t+i}^c is the probability of a comment at time $t + i$.