MMDS 2012
Stanford University

# Large Scale Machine Learning for Query Document Matching in Web Search*

Hang Li
Huawei Technologies

* Work was done at Microsoft Research, with former colleagues and interns

# Talk Outline

- Motivation
- Regularized Latent Semantic Indexing
- Group Matrix Factorization
- Matching in Latent Space
- Conclusion

# Same Search Intent, Different Query Representations Example = “Distance between Sun and Earth”

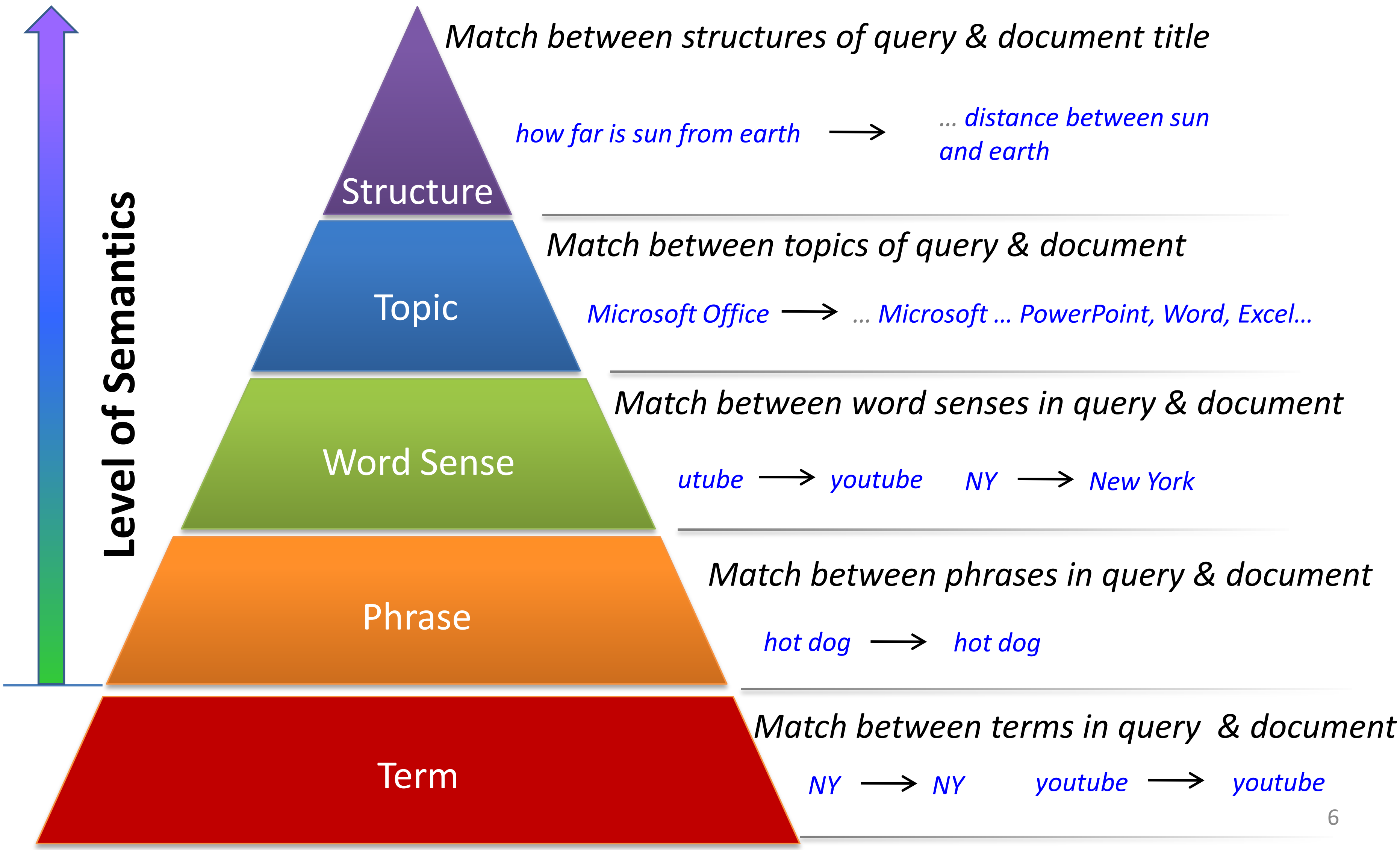
- "how far" earth sun
- "how far" sun
- "how far" sun earth
- average distance earth sun
- average distance from earth to sun
- average distance from the earth to the sun
- distance between earth & sun
- distance between earth and sun
- distance between earth and the sun
- distance from earth to the sun
- distance from sun to earth
- distance from sun to the earth
- distance from the earth to the sun
- distance from the sun to earth
- distance from the sun to the earth
- distance of earth from sun
- distance between earth sun
- how far away is the sun from earth
- how far away is the sun from the earth
- how far earth from sun
- how far earth is from the sun
- how far from earth is the sun
- how far from earth to sun
- how far from the earth to the sun
- distance between sun and earth

# Same Search Intent, Different Query Representations Example = "Youtube"

| | | |
|---|---|---|
| • yutube | yuotube | yuo tube |
| • ytube | youtubr | yu tube |
| • youtubo | youtuber | youtubecom |
| • youtube om | youtube music videos | youtube videos |
| • youtube | youtube com | youtube co |
| • youtub com | you tube music videos | yout tube |
| • youtub | you tube com yourtube | your tube |
| • you tube | you tub | you tube video clips |
| • you tube videos | www you tube com | wwww youtube com |
| • www youtube | www youtube com | www youtube co |
| • yotube | www you tube | www utube com |
| • ww youtube com | www utube | www u tube |
| • utube videos | utube com | utube |
| • u tube com | utub | u tube videos |
| • u tube | my tube | toutube |
| • outube | our tube | toutube |

# Semantic Matching Project: Solving Document Mismatch in Web Search

# Matching at Different Levels



# Query Understanding

Structure Identification

michael jordan: **main phrase**

**Structure**

Topic Identification

michael jordan berkely: **machine learning**

**Topic**

Similar Query Finding

michael I. jordan

michael jordan

**Sense**

Phrase Identification

[michael jordan] berkeley

**Phrase**

Spelling Error Correction

michael jordan berkeley

**Term**

michael jordan berkele

# Document Understanding

Title Structure Identification

Topic Identification

Key Phrase Identification

Phrase Identification

Michael Jordan: ***main phrase in Title*** **Structure**

Michael Jordan is Professor in the Department of Electrical Engineering: ***machine learning*** **Topic**

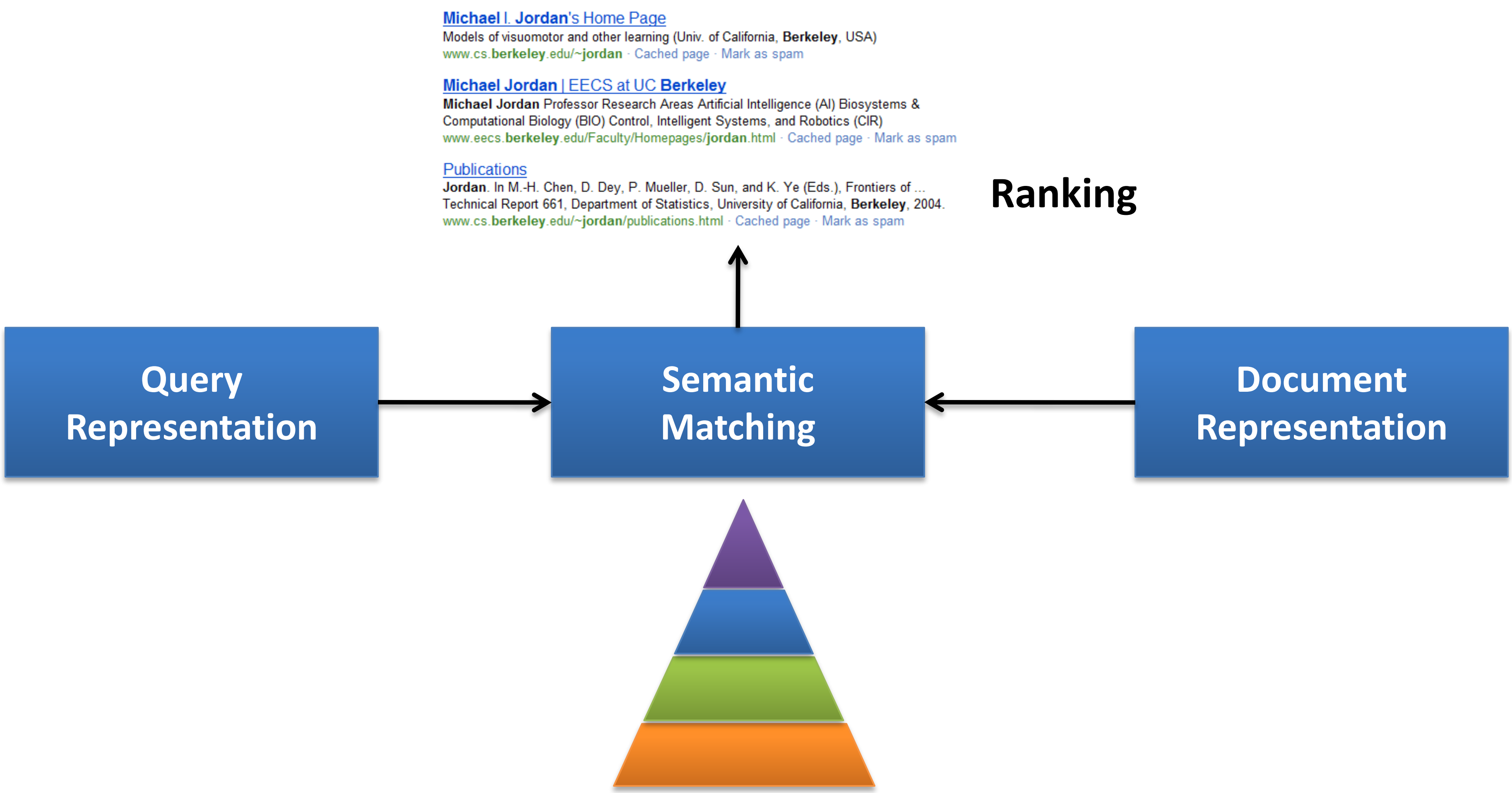
[Michael Jordan], [Professor] [Electrical Engineering]: ***keyphrase*** **Phrase**

[Michael Jordan] is [Professor] in the [Department] of [Electrical Engineering] **Phrase**

Homepage of Michael Jordan

Michael Jordan is Professor in the Department of Electrical Engineering

# Online Matching




**Matching can be conducted at different levels**

# Related Work

- Studied in long history of IR
- Query expansion, pseudo relevance feedback
- Latent Semantic Indexing, Probabilistic Latent Semantic Indexing, Latent Dirichlet Allocation
- … …
- New problem setting
  - Large amount of data available
  - New machine learning techniques

# Matching vs Ranking

In search, first matching and then ranking

| | Matching | Ranking |
|---|---|---|
| Prediction | Matching degree between query and document | Ranking list of documents |
| Model | f(q, d) | f(q,d1), f(q,d2), … f(q,dn) |
| Challenge | Mismatch | Correct ranking on top |

# Matching between Heterogeneous Data is Everywhere

- Matching between user and product (collaborative filtering)
- Matching between text and image (image annotation)
- Matching between people (dating)
- Matching between languages (machine translation)
- Matching between receptor and ligand (drug design)

# Regularized Latent Semantic Indexing

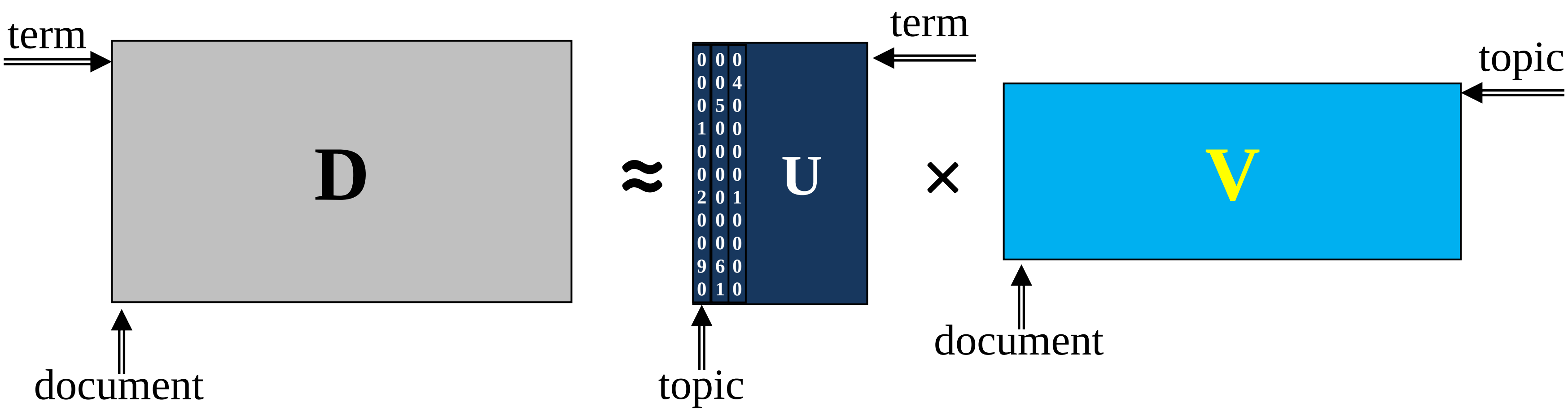
Joint Work with Quan Wang, Jun Xu, and Nick Craswell

SIGIR 2011

# Regularized Latent Semantic Indexing

- Motivation
  - Matching between query and document at topic level
  - Scale up to large datasets (vs. existing methods)
- Approach
  - Matrix Factorization
  - Regularization on topics and documents (vs. Sparse Coding)
  - Learning problem can be easily decomposed
- Results
  - $l_1$ on topics leads to sparse topics and $l_2$ on documents leads to accurate matching
  - Comparable with existing methods in topic discovery and search relevance
  - But can easily scale up to large document sets

# Regularized Latent Semantic Indexing




term representation of doc $n$

topics

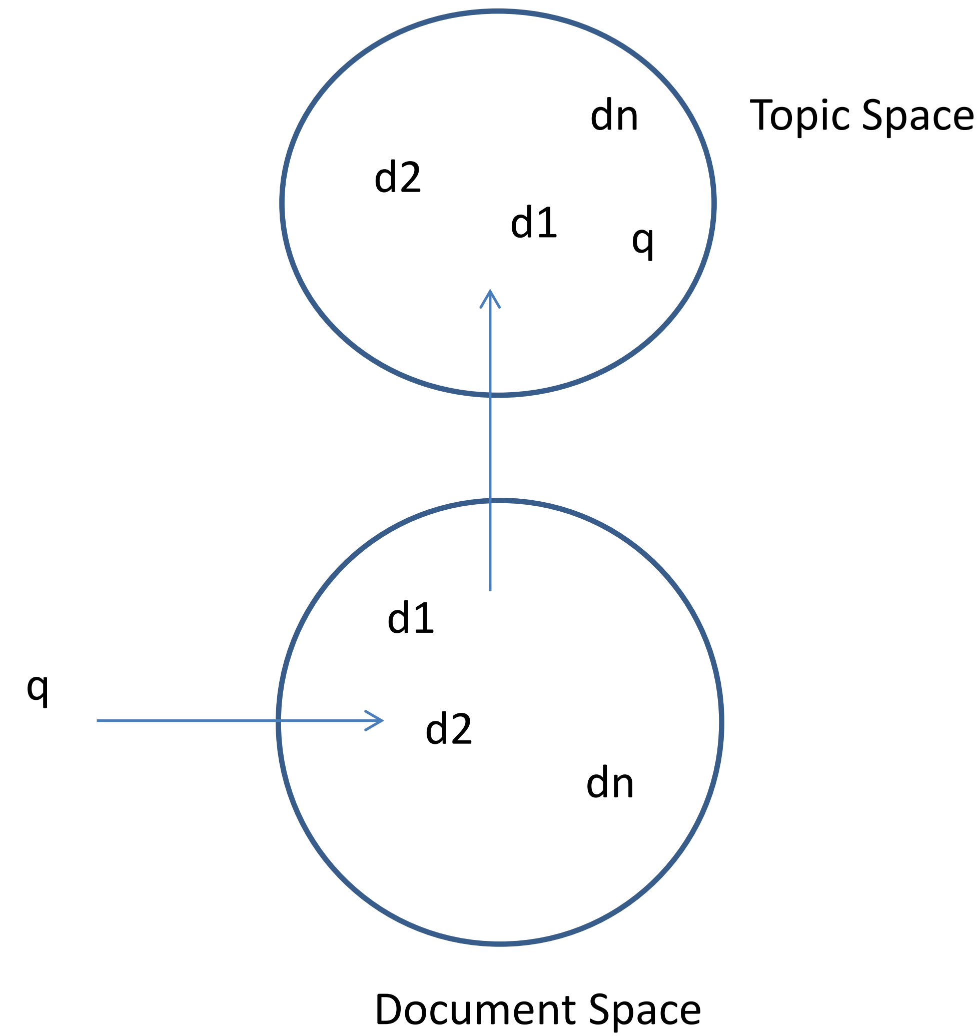
topic representation of doc $n$

documents are smooth

$$\min_{\mathbf{U},\{\boldsymbol{v}_n\}} \sum_{n=1}^{N} \|\boldsymbol{d}_n - \mathbf{U}\boldsymbol{v}_n\|_2^2 + \lambda_1 \sum_{k=1}^{K} \|\boldsymbol{u}_k\|_1 + \lambda_2 \sum_{n=1}^{N} \|\boldsymbol{v}_n\|_2^2$$

topics are sparse

# Query and Document Matching in Topic Space



# Optimization Strategy

$$\min_{\mathbf{U},\{\boldsymbol{v}_n\}} \sum_{n=1}^{N} \|\boldsymbol{d}_n - \mathbf{U}\boldsymbol{v}_n\|_2^2 + \lambda_1 \sum_{k=1}^{K} \|\boldsymbol{u}_k\|_1 + \lambda_2 \sum_{n=1}^{N} \|\boldsymbol{v}_n\|_2^2$$

Coordinate Decent

Update $\mathbf{U}$

$$\min_{\{\bar{\boldsymbol{u}}_m\}} \sum_{m=1}^{M} \left\|\bar{\boldsymbol{d}}_m - \mathbf{V}^T\bar{\boldsymbol{u}}_m\right\|_2^2 + \lambda_1 \sum_{m=1}^{M} \|\bar{\boldsymbol{u}}_m\|_1$$

$$\min_{\bar{\boldsymbol{u}}_m} \left\|\bar{\boldsymbol{d}}_m - \mathbf{V}^T\bar{\boldsymbol{u}}_m\right\|_2^2 + \lambda_1 \|\bar{\boldsymbol{u}}_m\|_1$$

for $m = 1, \cdots, M$

$$u_{mk} = \begin{cases} \dfrac{(r_{mk} - \sum_{l \neq k} s_{kl} u_{ml}) - \frac{1}{2}\lambda_1}{s_{kk}}, & \text{if } u_{mk} > 0 \\ \dfrac{(r_{mk} - \sum_{l \neq k} s_{kl} u_{ml}) + \frac{1}{2}\lambda_1}{s_{kk}}, & \text{if } u_{mk} < 0 \end{cases}$$

Update $\mathbf{V}$

$$\min_{\{\boldsymbol{v}_n\}} \sum_{n=1}^{N} \|\boldsymbol{d}_n - \mathbf{U}\boldsymbol{v}_n\|_2^2 + \lambda_2 \sum_{n=1}^{N} \|\boldsymbol{v}_n\|_2^2$$

$$\min_{\boldsymbol{v}_n} \|\boldsymbol{d}_n - \mathbf{U}\boldsymbol{v}_n\|_2^2 + \lambda_2 \|\boldsymbol{v}_n\|_2^2$$

for $n = 1, \cdots, N$

$$\boldsymbol{v}_n^* = \left(\mathbf{U}^T\mathbf{U} + \lambda_2\mathbf{I}\right)^{-1} \mathbf{U}^T \boldsymbol{d}_n$$

Analytic Solution

# RLSI Algorithm

- Single machine multi core version
- Multiple machine version (MapReduce and MPI)

**Require:** $\mathbf{D} \in \mathbb{R}^{M \times N}$

$\mathbf{V}^{(0)} \in \mathbb{R}^{K \times N} \leftarrow$ random matrix
**for** $t = 1 : T$ **do**
  $\mathbf{U}^{(t)} \leftarrow \text{Update}\mathbf{U}(\mathbf{D}, \mathbf{V}^{(t-1)})$
  $\mathbf{V}^{(t)} \leftarrow \text{Update}\mathbf{V}(\mathbf{D}, \mathbf{U}^{(t)})$
**end for**
**return** $\mathbf{U}^{(T)}, \mathbf{V}^{(T)}$

**Algorithm 2** Update **U**

**Require:** $\mathbf{D} \in \mathbb{R}^{M \times N}$, $\mathbf{V} \in \mathbb{R}^{K \times N}$

$\mathbf{S} \leftarrow \mathbf{V}\mathbf{V}^T$
$\mathbf{R} \leftarrow \mathbf{D}\mathbf{V}^T$
**for** $m = 1 : M$ **do**
  $\bar{\boldsymbol{u}}_m \leftarrow \mathbf{0}$
  **repeat**
    **for** $k = 1 : K$ **do**
      $w_{mk} \leftarrow r_{mk} - \sum_{l \neq k} s_{kl} u_{ml}$
      $u_{mk} \leftarrow \frac{\left(|w_{mk}| - \frac{1}{2}\lambda N\right)_+ \text{sign}(w_{mk})}{s_{kk}}$
    **end for**
  **until** convergence
**end for**
**return** **U**

terms processed in parallel

**Algorithm 3** Update **V**

**Require:** $\mathbf{D} \in \mathbb{R}^{M \times N}$, $\mathbf{U} \in \mathbb{R}^{M \times K}$

$\boldsymbol{\Sigma} \leftarrow \left(\mathbf{U}^T\mathbf{U} + \theta\mathbf{I}\right)^{-1}$
$\boldsymbol{\Phi} \leftarrow \mathbf{U}^T\mathbf{D}$
**for** $n = 1 : N$ **do**
  $\boldsymbol{v}_n \leftarrow \boldsymbol{\Sigma}\boldsymbol{\phi}_n$, where $\boldsymbol{\phi}_n$ is the $n^{th}$ column
**end for**
**return** **V**

docs processed in parallel

# Scalability Comparison

| **algorithm** | **max dataset applied (#docs; #words)** | **# topics** | **# processors used** |
|---|---|---|---|
| PLDA and PLDA+ (by Google) | Wiki-200T(2,112,618; **200,000**) | 1000 | 2, 048 |
| AD-LDA | NY Times (300,000; **102,660**)<br>PubMed (8,200,000; **141,043**) | 200 | 16 |
| RLSI | B01 (1,562,807; **7,014,881**)<br>Bing News (940,702; **500,033**)<br>Wiki-All (3,239,884; **6,043,069**)<br>MSWeb Data (2,635,158; **2,371,146**) | 500 ~ 1000 | single machine, 24 cores |

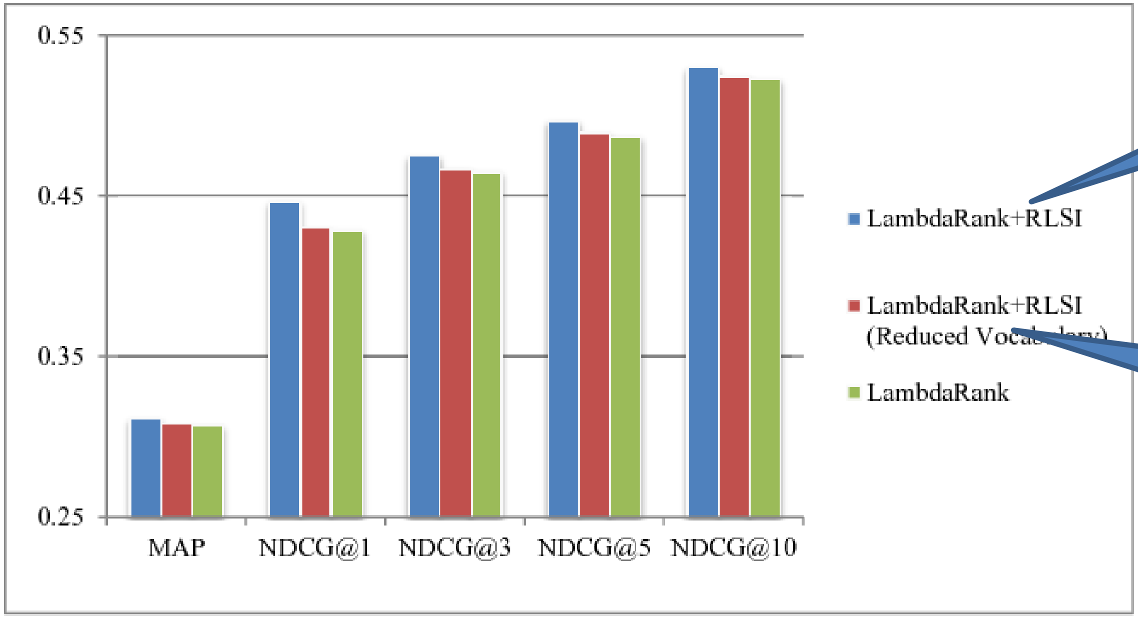
# Experimental Results on Topic Discovery

Topics discovered by RLSI are equally readable compared with LDA, PLSI, LSI

**Table 8: Topics discovered by RLSI, LDA, PLSI, and LSI from AP dataset.**

| | **Topic 1** | **Topic 2** | **Topic 3** | **Topic 4** | **Topic 5** | **Topic 6** | **Topic 7** | **Topic 8** | **Topic 9** | **Topic 10** |
|---|---|---|---|---|---|---|---|---|---|---|
| | opec | africa | aid | school | noriega | percent | plane | israeli | nuclear | bush |
| | oil | south | virus | student | panama | billion | crash | palestinian | soviet | dukakis |
| RLSI | cent | african | infect | teacher | panamanian | rate | flight | israel | treaty | campaign |
| AvgComp = 0.0075 | barrel | angola | test | educate | delval | 0 | air | arab | missile | quayle |
| | price | apartheid | patient | college | canal | trade | airline | plo | weapon | bentsen |
| | soviet | school | dukakis | party | year | water | price | court | air | iran |
| | nuclear | student | democrat | govern | new | year | year | charge | plane | iranian |
| LDA | union | year | campaign | minister | time | fish | market | case | flight | ship |
| AvgComp = 1 | state | educate | bush | elect | television | animal | trade | judge | crash | iraq |
| | treaty | university | jackson | nation | film | 0 | percent | attorney | airline | navy |
| | company | israeli | year | year | bush | court | soviet | year | plane | year |
| | million | iran | state | state | dukakis | charge | treaty | state | flight | state |
| PLSI | share | israel | new | new | democrat | attorney | missile | new | airline | new |
| AvgComp = 0.9534 | billion | palestinian | nation | nation | campaign | judge | nuclear | nation | crash | people |
| | stock | arab | govern | 0 | republican | trial | gorbachev | govern | air | nation |
| | soviet | 567 | 0 | earthquake | drug | 0 | israel | yen | urgent | student |
| | percent | 234 | yen | quake | school | dukakis | israeli | dukakis | oil | school |
| LSI | police | 0 | dollar | richter | test | bush | student | bush | opec | noriega |
| AvgComp = 1 | govern | percent | percent | scale | court | jackson | palestinian | dollar | dukakis | panama |
| | state | 12 | tokyo | damage | dukakis | dem | africa | jackson | cent | teacher |

# Experimental Results on Web Search




RLSI can help improve search relevance

Reducing vocabulary hurts ranking accuracy

# Group Matrix Factorization

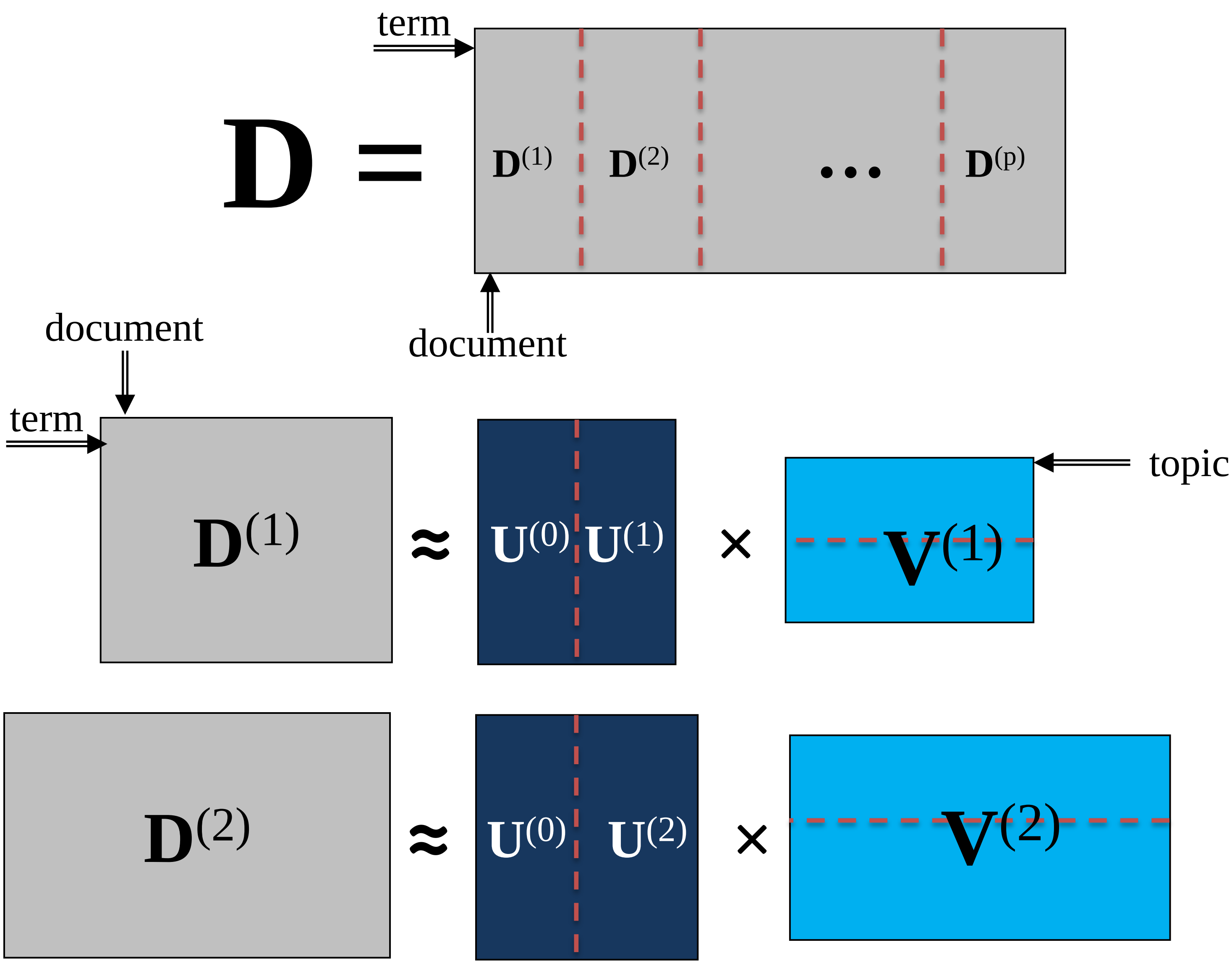
Joint Work with Quan Wang, Zheng Cao, and Jun Xu

SIGIR 2012

# Group Matrix Factorization

- Motivation
  - Matching between query and document at topic level
  - Having even better scalability
- Approach
  - Matrix Factorization (RLSI and NMF)
  - Assuming documents have been classified into classes
  - Class specific topics and shared topics
- Results
  - Comparable with existing methods in topic discovery and search relevance
  - Can scale up to even large document sets

# Group Matrix Factorization



# Group Matrix Factorization (cont')

$$\min_{\{\boldsymbol{u}_k^{(0)}\},\{\boldsymbol{u}_k^{(p)}\},\{\boldsymbol{v}_n^{(p)}\}} \sum_{p=1}^{P}\sum_{n=1}^{N_p} \mathcal{L}\left(\boldsymbol{d}_n^{(p)} \| \tilde{\mathrm{U}}_p \boldsymbol{v}_n^{(p)}\right) + \theta_1 \sum_{k=1}^{K_s} \mathcal{R}_1\left(\boldsymbol{u}_k^{(0)}\right)$$

$$+ \theta_2 \sum_{p=1}^{P}\sum_{k=1}^{K_c} \mathcal{R}_2\left(\boldsymbol{u}_k^{(p)}\right) + \theta_3 \sum_{p=1}^{P}\sum_{n=1}^{N_p} \mathcal{R}_3\left(\boldsymbol{v}_n^{(p)}\right)$$

$$\begin{aligned} s.t. \quad & \boldsymbol{u}_k^{(0)} \in C_1, \quad k = 1, \cdots, K_s, \\ & \boldsymbol{u}_k^{(p)} \in C_2, \quad k = 1, \cdots, K_c, p = 1, \cdots, P, \\ & \boldsymbol{v}_n^{(p)} \in C_3, \quad n = 1, \cdots, N_p, p = 1, \cdots, P, \end{aligned}$$

# Group Regularized Latent Semantic Indexing

$$\min_{\{\boldsymbol{u}_k^{(0)}\},\{\boldsymbol{u}_k^{(p)}\},\{\boldsymbol{v}_n^{(p)}\}} \sum_{p=1}^{P} \sum_{n=1}^{N_p} \left\| \boldsymbol{d}_n^{(p)} - \tilde{\mathrm{U}}_p \boldsymbol{v}_n^{(p)} \right\|_2^2 + \lambda_1 \sum_{k=1}^{K_s} \left\| \boldsymbol{u}_k^{(0)} \right\|_1$$

$$+ \lambda_1 \sum_{p=1}^{P} \sum_{k=1}^{K_c} \left\| \boldsymbol{u}_k^{(p)} \right\|_1 + \lambda_2 \sum_{p=1}^{P} \sum_{n=1}^{N_p} \left\| \boldsymbol{v}_n^{(p)} \right\|_2^2$$

# G-RLSI Algorithm

- Single machine multi core version
- Multiple machine version (MapReduce and MPI)

**Algorithm 1** Group RLSI

**Require:** $\mathbf{D}_1, \cdots, \mathbf{D}_P$

**for** $p = 1 : P$ **do**
  $\mathbf{U}_p \leftarrow$ zero matrix
  $\mathbf{V}_p \leftarrow$ random matrix
**end for**
**repeat**
  $\mathbf{U}_0 \leftarrow \text{Update}U_0\left(\{\mathbf{D}_p\}, \{\mathbf{U}_p\}, \{\mathbf{V}_p\}\right)$
  **for** $p = 1 : P$ **do**
    $\mathbf{U}_p \leftarrow \text{Update}U_p\left(\mathbf{D}_p, \mathbf{U}_0, \mathbf{V}_p\right)$
    $\mathbf{V}_p \leftarrow \text{Update}V_p\left(\mathbf{D}_p, \mathbf{U}_0, \mathbf{U}_p\right)$
  **end for**
**until** convergence
**return** $\mathbf{U}_0, \mathbf{U}_1, \cdots, \mathbf{U}_P, \mathbf{V}_1, \cdots, \mathbf{V}_P$

**Algorithm 2** Update$U_0$

**Require:** $\mathbf{D}_1, \cdots, \mathbf{D}_P, \mathbf{U}_1, \cdots, \mathbf{U}_P, \mathbf{V}_1, \cdots, \mathbf{V}_P$

$\mathbf{S}_0 \leftarrow \sum_{p=1}^{P} \mathbf{H}_p \mathbf{H}_p^T$
$\mathbf{R}_0 \leftarrow \sum_{p=1}^{P} \mathbf{D}_p \mathbf{H}_p^T - \sum_{p=1}^{P} \mathbf{U}_p \mathbf{W}_p \mathbf{H}_p^T$
**for** $m = 1 : M$ **do**
  $\bar{u}_m^{(0)} \leftarrow \mathbf{0}$
  **repeat**
    **for** $k = 1 : K_s$ **do**
      $x_{mk} \leftarrow r_{mk}^{(0)} - \sum_{l \neq k} s_{kl}^{(0)} u_{ml}^{(0)}$
      $u_{mk}^{(0)} \leftarrow \frac{\left(|x_{mk}| - \frac{1}{2}\lambda_1\right)_+ \text{sign}(x_{mk})}{s_{kk}^{(0)}}$
    **end for**
  **until** convergence
**end for**
**return** $\mathbf{U}_0$

**Algorithm 3** Update$U_p$

**Require:** $\mathbf{D}_p, \mathbf{U}_0, \mathbf{V}_p$

$\mathbf{S}_p \leftarrow \mathbf{W}_p \mathbf{W}_p^T$
$\mathbf{R}_p \leftarrow \mathbf{D}_p \mathbf{W}_p^T - \mathbf{U}_0 \mathbf{H}_p \mathbf{W}_p^T$
**for** $m = 1 : M$ **do**
  $\bar{u}_m^{(p)} \leftarrow \mathbf{0}$
  **repeat**
    **for** $k = 1 : K_c$ **do**
      $x_{mk} \leftarrow r_{mk}^{(p)} - \sum_{l \neq k} s_{kl}^{(p)} u_{ml}^{(p)}$
      $u_{mk}^{(p)} \leftarrow \frac{\left(|x_{mk}| - \frac{1}{2}\lambda_1\right)_+ \text{sign}(x_{mk})}{s_{kk}^{(p)}}$
    **end for**
  **until** convergence
**end for**
**return** $\mathbf{U}_p$

**Algorithm 4** Update$V_p$

**Require:** $\mathbf{D}_p, \mathbf{U}_0, \mathbf{U}_p$

$\mathbf{\Sigma}_p \leftarrow \left(\tilde{\mathbf{U}}_p^T \tilde{\mathbf{U}}_p + \lambda_2 \mathbf{I}\right)^{-1}$
$\mathbf{\Phi}_p \leftarrow \tilde{\mathbf{U}}_p^T \mathbf{D}_p$
**for** $n = 1 : N_p$ **do**
  $v_n^{(p)} \leftarrow \mathbf{\Sigma}_p \boldsymbol{\phi}_n^{(p)}$
**end for**
**return** $\mathbf{V}_p$

# Efficiency Comparison

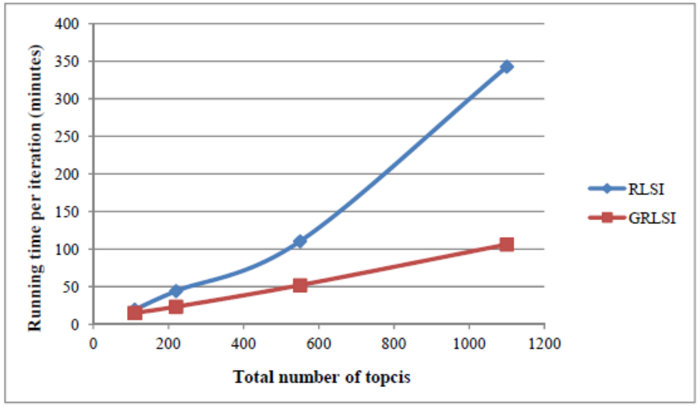



**Figure 2: Execution time of RLSI and GRLSI on Wikipedia.**

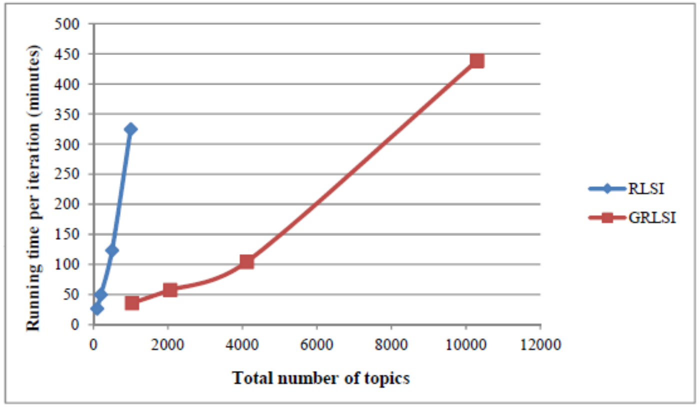



**Figure 4: Execution time of RLSI and GRLSI on Web-I.**

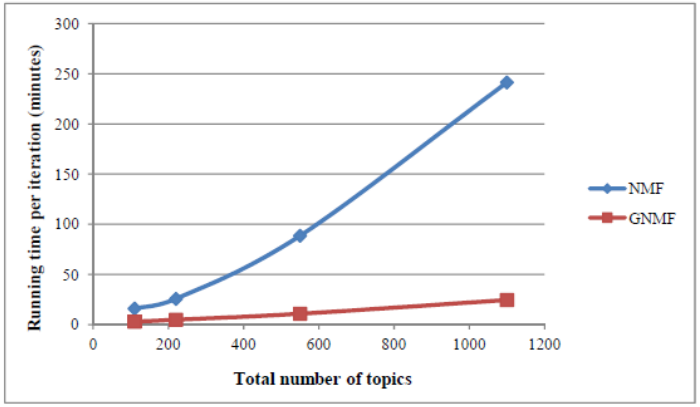



**Figure 3: Execution time of NMF and GNMF on Wikipedia.**

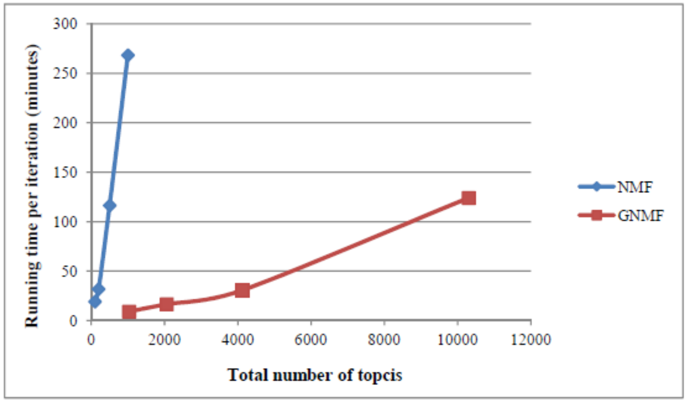



**Figure 5: Execution time of NMF and GNMF on Web-I.**

# Experimental Results on Topic Discovery

**Table 10: Topics discovered by GRLSI (top) and GNMF (bottom) on Wikipedia.**

| | Shared topics | | | Arts | | | Geography | | | Politics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GRLSI | commune | state | political | album | rock | groups | province | municipality | communes | elections | states | kingdom |
| | communes | highways | party | albums | american | musical | state | municipalities | commune | election | congressional | political |
| | department | route | colour | singers | musicians | music | village | gmina | department | weapon | delegations | parties |
| | places | highway | india | musicians | singers | rappers | villages | voivodeship | france | party | elections | country |
| | france | india | canada | track | country | metal | highways | population | departments | parties | united | party |
| | populated | brazil | australia | listing | english | heavy | united | germany | places | political | senate | fascism |
| | municipality | oregan | parties | pop | pop | created | states | spain | county | results | tennessee | submarine |
| GNMF | places | new | language | album | groups | rappers | village | district | department | elections | war | military |
| | populated | york | japanese | albums | rock | musicians | villages | germany | commune | election | world | country |
| | village | city | films | track | american | american | england | districts | communes | results | poland | units |
| | azerbaijan | zealand | cast | listing | musical | singers | india | town | france | members | weapons | formations |
| | population | jersey | chinese | released | metal | singles | population | administrative | departments | parties | conflict | army |
| | municipality | routes | english | band | musicians | wiley | central | towns | home | held | union | infantry |
| | census | south | spanish | records | pop | blues | within | cities | western | general | force | established |

**Table 11: Topics discovered by GRLSI (top) and GNMF (bottom) on Web-I.**

| | Shared topics | | | Arts/literature | | | Business/healthcare | | | Computers/internet | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GRLSI | video | business | games | poems | harry | book | dental | healthcare | care | chat | facebook | web |
| | phone | services | game | poetry | potter | chapter | dentist | practice | medical | teen | people | hosting |
| | mobile | company | cheats | poem | books | summary | care | test | health | online | connect | design |
| | tv | service | xbox | poets | rowling | books | dentistry | management | equipment | people | sign | website |
| | cell | products | ign | love | series | analysis | dentists | exam | ppo | friends | web | domain |
| | phones | management | pc | poet | children | author | health | patient | supplies | join | password | internet |
| | iphone | jobs | updates | american | deathly | study | teeth | jobs | company | video | friends | services |
| GNMF | www | products | day | poems | harry | books | dentist | healthcare | medical | google | facebook | design |
| | http | product | october | quotes | potter | children | dentists | management | equipment | maps | people | web |
| | org | quality | september | shakespear | rowling | read | dentistry | patient | supplies | blog | connect | website |
| | website | buy | july | william | series | reading | dr | hospital | surgical | gmail | sign | development |
| | net | accessories | june | poetry | deathly | list | dental | solutions | patient | map | friends | marketing |
| | html | store | august | poets | hallows | readers | cosmetic | nursing | hospital | engine | password | graphic |
| | web | supplies | april | poem | stone | fiction | teeth | hospitals | device | web | share | logo |

# Experimental Results on Web Search

**Table 12: Relevance performance of RLSI families on Web-II.**

| Method | MAP | NDCG@1 | NDCG@3 | NDCG@5 | NDCG@10 |
|---|---|---|---|---|---|
| BM25 | 0.3006 | 0.3043 | 0.3490 | 0.3910 | 0.4805 |
| BM25+RLSI | 0.3050 | 0.3076 | 0.3539 | 0.3943 | 0.4858 |
| BM25+CRLSI | 0.3027 | 0.3051 | 0.3509 | 0.3927 | 0.4840 |
| BM25+GRLSI | 0.3039 | 0.3066 | 0.3520 | 0.3934 | 0.4855 |

**Table 13: Relevance performance of NMF families on Web-II.**

| Method | MAP | NDCG@1 | NDCG@3 | NDCG@5 | NDCG@10 |
|---|---|---|---|---|---|
| BM25 | 0.3006 | 0.3043 | 0.3490 | 0.3910 | 0.4805 |
| BM25+NMF | 0.3057 | 0.3091 | 0.3546 | 0.3960 | 0.4895 |
| BM25+CNMF | 0.3033 | 0.3055 | 0.3512 | 0.3934 | 0.4869 |
| BM25+GNMF | 0.3046 | 0.3080 | 0.3530 | 0.3955 | 0.4887 |

# Matching in Latent Space

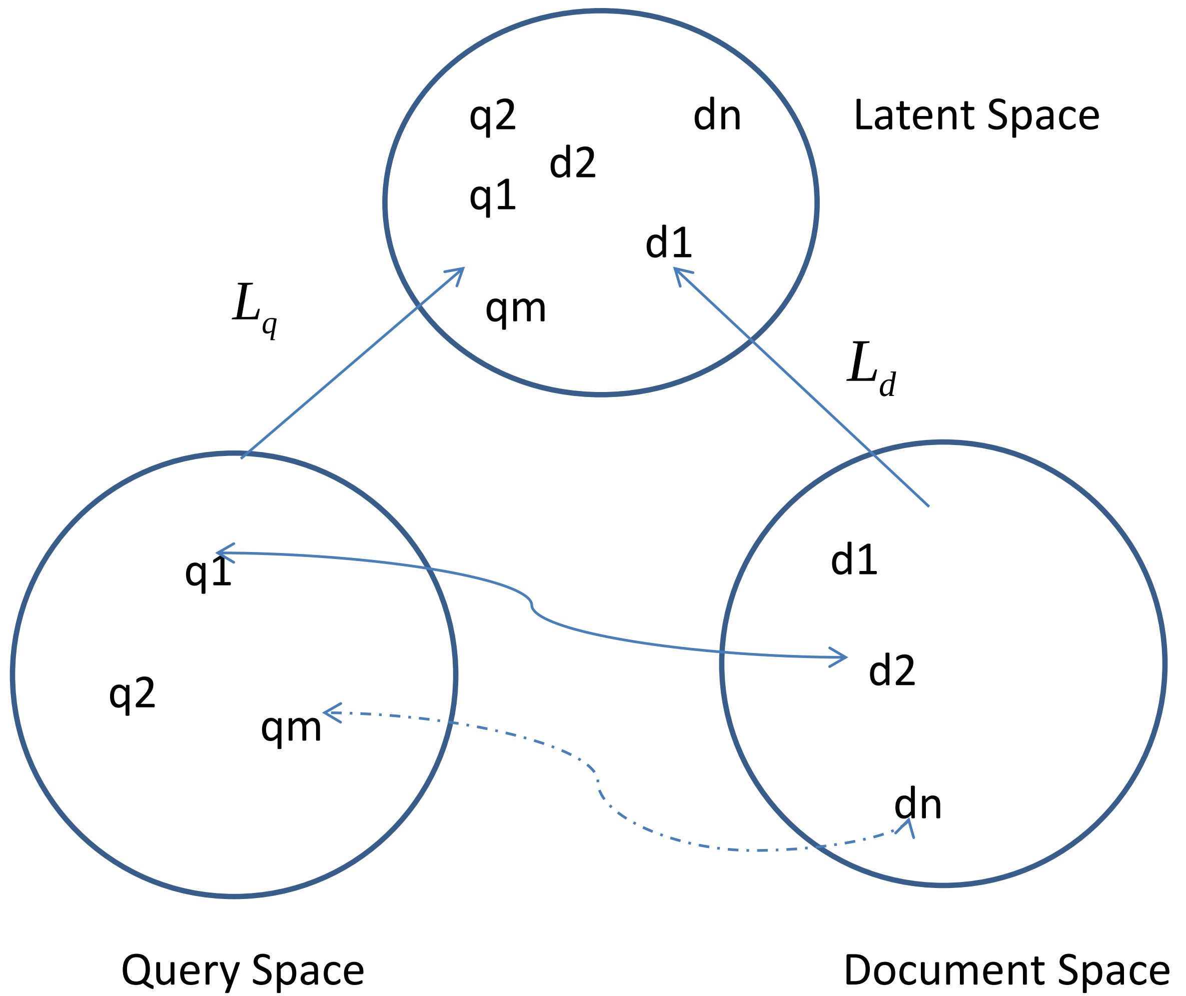
Joint Work with Wei Wu, Zhengdong Lv

Under review

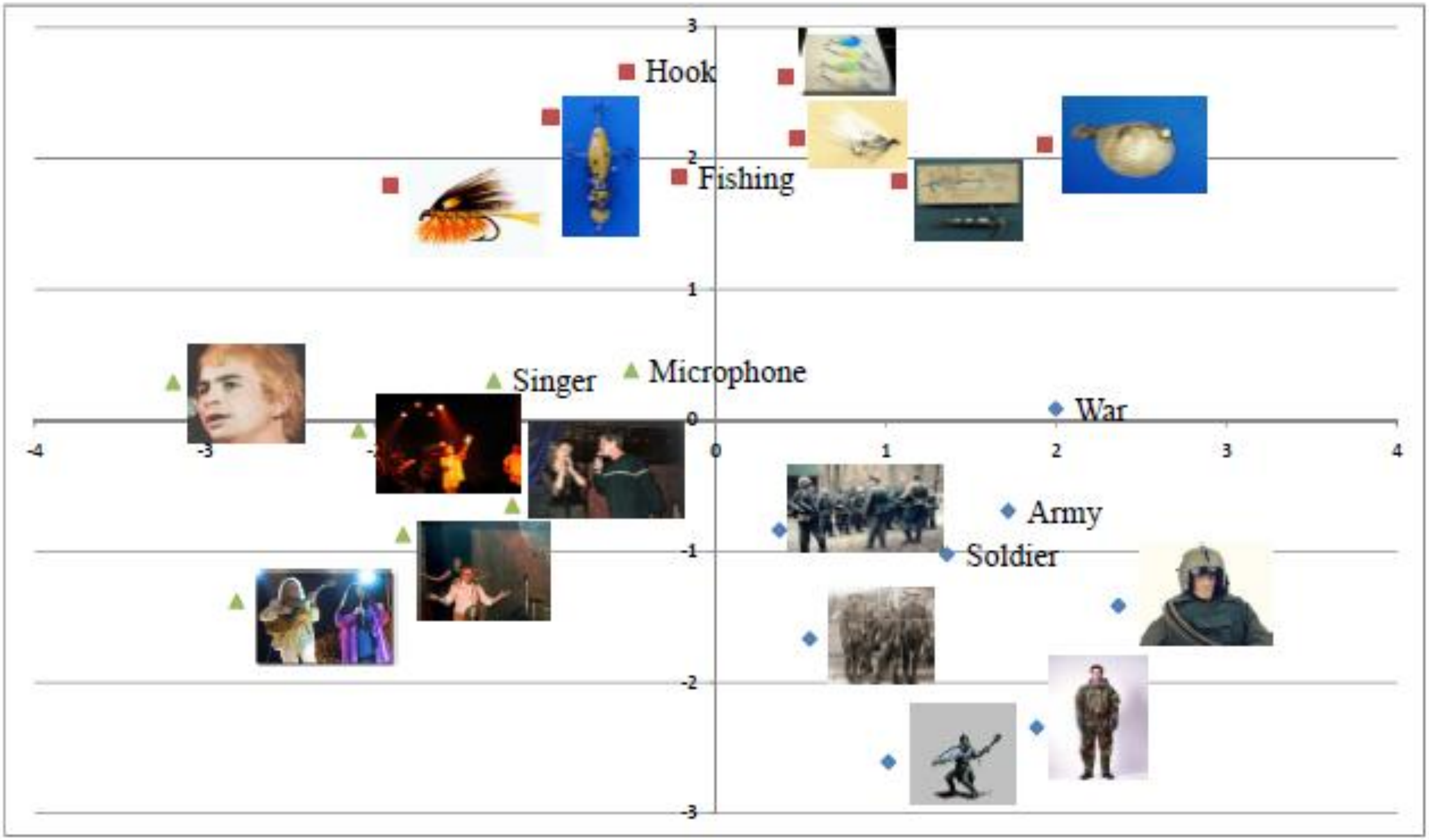
# Matching in Latent Space

- Motivation
  - Matching between query and document in latent space
- Assumption
  - Queries have similarity
  - Document have similarity
  - Click-through data represent “similarity” relations between queries and documents
- Approach
  - Projection to latent space
  - Regularization or constraints
- Results
  - Significantly enhance accuracy of query document matching

# Matching in Latent Space



# Example: Projecting Keywords and Images into Latent Space



# Partial Least Square (PLS)

- Setting
  - Two spaces: $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{Y} \subset \mathbb{R}^n$.
- Input
  - Training data: $\{(x_i, y_i, r_i)\}_{1 \le i \le N}$, $r_i \in \{+1, -1\}$
- Output
  - Similarity function $f(x, y)$
- Assumption
  - Two linear (and orthonormal) transformations $L_\mathcal{X}$ and $L_\mathcal{Y}$
  - Dot product as similarity function $\langle {L_\mathcal{X}}^T x,\ {L_\mathcal{Y}}^T y \rangle = x^T L_\mathcal{X} {L_\mathcal{Y}}^T y$
- Optimization

$$argmax_{L_\mathcal{X}, L_\mathcal{Y}} \sum_{r_i=+1} {x_i}^T L_\mathcal{X} {L_\mathcal{Y}}^T y_i - \sum_{r_i=-1} {x_i}^T L_\mathcal{X} {L_\mathcal{Y}}^T y_i$$

$$subject\ to\ {L_\mathcal{X}}^T L_\mathcal{X} = I_{k \times k}, {L_\mathcal{Y}}^T L_\mathcal{Y} = I_{k \times k}$$

# Solution of Partial Least Square

- Non-convex optimization
- Can prove that global optimal solution exists
- Global optimal can be found by solving SVD (Singular Value Decomposition)
- SVD of Matrix $M_s - M_D = U\Sigma V^T$

# Regularized Mapping to Latent Space (RMLS)

- Setting
  - Two spaces: $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{Y} \subset \mathbb{R}^n$.
- Input
  - Training data: $\{(x_i, y_i, r_i)\}_{1 \le i \le N}$, $r_i \in \{+1, -1\}$
- Output
  - Similarity function $f(x, y)$
- Assumption
  - L1 and L2 regularization on $L_\mathcal{X}$ and $L_\mathcal{Y}$ (sparse transfromations)
  - Dot product as similarity function $\langle {L_\mathcal{X}}^T x, \ {L_\mathcal{Y}}^T y \rangle = x^T L_\mathcal{X} {L_\mathcal{Y}}^T y$
- Optimization

$$argmax_{L_\mathcal{X}, L_\mathcal{Y}} \ \sum_{r_i=+1} {x_i}^T L_\mathcal{X} {L_\mathcal{Y}}^T y_i - \sum_{r_i=-1} {x_i}^T L_\mathcal{X} {L_\mathcal{Y}}^T y_i$$

$$subject\ to \ |lx| \le \vartheta x, \ |ly| \le \vartheta y, \ \| lx \| \le \lambda x, \ \| ly \| \le \lambda y,$$

# Solution of Regularized Mapping to Latent Space

- Coordinate Descent
- Repeat
  - Fix $Lx$, update$Ly$
  - Fix $Ly$, update$Lx$
- Update can be parallelized by rows

# Comparison

| | PLS | RMLS |
|---|---|---|
| Assumption | Orthogonal | L1 and L2 Regularization |
| Optimization Method | Singular Value Decomposition | Coordinate Descent |
| Optimality | Global optimum | Local optimum |
| Efficiency | Low | High |
| Scalability | Low | High |

# Experimental Results

Enterprise Search

| | NDCG@1 | NDCG@3 | NDCG@5 |
|---|---|---|---|
| $MPLS_{Com}$ | **0.715** | **0.733** | **0.747** |
| $MPLS_{Conca}$ | 0.700 | 0.728 | 0.742 |
| $MPLS_{Word}$ | 0.688 | 0.718 | 0.739 |
| $MPLS_{Bipar}$ | 0.659 | 0.684 | 0.705 |
| BM25 | 0.653 | 0.657 | 0.663 |
| RW | 0.654 | 0.683 | 0.700 |
| RW+BM25 | 0.664 | 0.688 | 0.705 |
| LSI | 0.656 | 0.676 | 0.695 |
| LSI+BM25 | 0.692 | 0.701 | 0.712 |

Web Search

| | NDCG@1 | NDCG@3 | NDCG@5 |
|---|---|---|---|
| $MPLS_{Com}$ | **0.681** | **0.731** | **0.739** |
| $MPLS_{Conca}$ | 0.676 | 0.728 | 0.736 |
| $MPLS_{Word}$ | 0.674 | 0.726 | 0.732 |
| $MPLS_{Bipar}$ | 0.612 | 0.680 | 0.693 |
| BM25 | 0.637 | 0.690 | 0.690 |
| RW | 0.655 | 0.704 | 0.704 |
| RW+BM25 | 0.671 | 0.718 | 0.716 |
| LSI | 0.588 | 0.665 | 0.676 |
| LSI+BM25 | 0.649 | 0.705 | 0.706 |

- RMLS and PLS work better than BM25, Random Walk, Latent Semantic Indexing
- RMLS works equally well as PLS, with higher learning efficiency and scalability

# Conclusion

# Conclusion

- Large scale topic modeling techniques
  - Regularized Latent Semantic Indexing
  - Group Matrix Factorization
- Large scale matching techniques
  - Matching in Latent Space
- Comparable with existing methods in terms of accuracy, much better in terms of efficiency and scalability
- Useful for web search

# Publications of the Project

- **Quan Wang, Zheng Cao, Jun Xu, Hang Li, Group Matrix Factorization for Scalable Topic Modeling, In Proceedings of the 35th Annual International ACM SIGIR Conference (SIGIR'12), to appear, 2012.**
- Xiaobing Xue, Yu Tao, Daxin Jiang and Hang Li, Automatically Mining Question Reformulation Patterns from Search Log Data, In Proceedings of the 50th Annual Meeting of Association for Computational Linguistics (ACL'12), to appear, 2012.
- Fan Bu, Hang Li, Xiaoyan Zhu, String Re-Writing Kernel, In Proceedings of the 50th Annual Meeting of Association for Computational Linguistics (ACL'12), to appear, 2012.
- Chen Wang, Keping Bi, Yunhua Hu, Hang Li, and Guihong Cao. Extracting Search-Focused Key N-Grams for Relevance Ranking in Web Search. In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'12), 343-352, 2012.
- Wei Wu, Jun Xu, Hang Li, and Satoshi Oyama, Learning A Robust Relevance Model for Search Using Kernel Methods, Journal of Machine Learning Research, 12, 1429-1458. 2011.
- **Quan Wang, Jun Xu, Hang Li, Nick Craswell, Regularized Latent Semantic Indexing, In Proceedings of the 34th Annual International ACM SIGIR Conference (SIGIR'11), 685-694, 2011.**
- Ziqi Wang, Gu Xu, Hang Li and Ming Zhang, A Fast and Accurate Method for Approximate String Search, In Proceedings of the 49th Annual Meeting of Association for Computational Linguistics: Human Language Technologies (ACL-HLT'11), 52-61, 2011.
- Jun Xu, Hang Li, Chaoliang Zhong, Relevance Ranking Using Kernels, In Proceedings of the 6th Asian Information Retrieval Societies Symposium (AIRS'10), Best Paper Award, 1-12, 2010.
- Jiafeng Guo, Gu Xu, Hang Li, Xueqi Cheng. A Unified and Discriminative Model for Query Refinement. In Proceedings of the 31st Annual International ACM SIGIR Conference (SIGIR'08), 379-386, 2008.
- **Wei Wu, Zhengdong Lv, Hang Li, Regularized Mapping to Latent Structures and Its Application to Web Search, under review.**

# Thank You!

Contact: hangli.hl@huawei.com