ONE LINE SUMMARY OF THE WORK

Incorporating the latent topic similarity feature in the regularized topic space helps improve standard learning-to-rank results.

This work will be presented at the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015) on October 20, 2015 in Melbourne, Australia.
TABLE OF CONTENTS

1 Background

2 Introduction

3 Related Topic Models
   - Unsupervised Latent Dirichlet Allocation (LDA)
   - Supervised Latent Dirichlet Allocation (sLDA)
   - Maximum Margin Models
   - Maximum Margin Supervised Topic Models (MedLDA)

4 Our Pairwise Maximum Margin Learning to Rank Model
   - Posterior Regularization
   - Bayesian Posterior Regularization
   - LDA as a Regularized Bayesian Model
   - Regularized Bayesian Learning to Rank
   - Parameter Estimation

5 Experiments and Results
   - Comparative Models and Evaluation Metric
   - NDCG Results

6 Conclusions
Background: Ad-Hoc Retrieval

- Document Collection
- Indexer
- Index
- Processing
- Query
- Query Formulation
- Searcher

Diagram arrows indicate the flow of information from the document collection through indexing, processing, and query formulation to the searcher.
Introduction

- Maximum margin learning-to-rank for document retrieval models are well known for their generalizeability.
- Probabilistic topic models have shown to give good performance on ad-hoc document retrieval [Wei and Croft, 2006].
- Why not incorporate both frameworks in a model?

A Common Approach

1. Train an existing topic model such as Latent Dirichlet Allocation (LDA).
2. Compute a topic-based similarity between the query and a document.
3. Use the similarity score as one of the features.
4. Train the learning-to-rank models.

A Problem

Such techniques are inherently sub-optimal due to error propagation.
Three approaches to learning to rank:

1. Pointwise - Ranking as a regression problem.

Query-document pairs are represented by a set of features.

These features are computed prior to conducting the learning.
UNSUPERVISED LATENT DIRICHLET ALLOCATION (LDA)

- LDA [Blei et al., 2003] posits that documents correspond to mixture of topics.
- Topic is a probability distribution over words.
- Each word $w$ has a latent topic assignment $z$.
- Topic weights for each document are stored in $\theta_d$.
- $\phi$ is a matrix of word-topic distributions.
- $\alpha$ and $\beta$ are prior distributions.
Unsupervised Latent Dirichlet Allocation (LDA)
Unsupervised Latent Dirichlet Allocation (LDA)

**UNSUPERVISED LATENT DIRICHLET ALLOCATION MODEL**

- \( \theta_d \) (Dirichlet distribution)
- \( \beta \) (Dirichlet distribution)
- \( \phi \) (Multinomial distribution)
- \( K \) (Number of topics)
- \( \alpha \) (Hyperparameter)
- \( M \) (Number of documents)
- \( N_d \) (Number of words in document)

Shoaib Jameel et al., Cardiff University

Learning-to-rank with Topic Models

October 12, 2015
LDA AS A MATRIX FACTORIZATION MODEL

Unsupervised Latent Dirichlet Allocation (LDA)

\[
\begin{align*}
\text{Documents} & \quad (d_1, d_2, \ldots, d_D) \\
\text{Terms} & \quad (v_1, v_2, \ldots, v_W) \\
\text{Topics} & \quad (v_1, v_2, v_W) \times \begin{pmatrix} k_1 & k_2 & k_3 \\ 0.50 & 0.11 & 0.10 \\ 0.44 & 0.27 & 0.34 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 0.00 & 0.00 & 0.47 \\ 0.19 & 0.05 & \cdot & \cdot & 0.10 \\ 0.01 & 0.43 & \cdot & \cdot & 0.52 \\ 0.03 & 0.15 & \cdot & \cdot & 0.24 \end{pmatrix}
\end{align*}
\]
Supervised Latent Dirichlet Allocation (sLDA)

- sLDA [Mcauliffe and Blei, 2008] posits that documents correspond to mixture of topics.
- Topic is a probability distribution over words.
- Each word $w$ has a latent topic assignment $z$.
- Topic weights for each document are stored in $\theta_d$.
- $\phi$ is a matrix of word-topic distributions.
- $\alpha$ and $\beta$ are prior distributions.
- Incorporates the document meta-data $y_d$ to generate topics, for example, classification labels, sentiment labels, etc.
Maximum Margin Models

Maximum Margin Models

- Optimal Hyperplane
- Maximum Margin

$\mathbf{x}_1, \mathbf{x}_2$
**LINEAR MAXIMUM MARGIN MODELS**

### Linear Maximum Margin Classifier

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| \eta \|^2 + C \sum \xi_{ik} \\
\text{subject to} & \quad \xi_{ik} \geq 0 \\
& \quad y_{ik}^u \eta^T (\psi_i^u - \psi_k^u) \geq 1 - \xi_{ik}, \forall u, i, k,
\end{align*}
\]

### Pairwise Maximum Margin Classifier

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| \eta \|^2 + C \sum \xi_{ik} \\
\text{subject to} & \quad \xi_{ik} \geq 0 \\
& \quad y_{ik}^u \eta^T (\psi_i^u - \psi_k^u) \geq 1 - \xi_{ik}, \forall u, i, k,
\end{align*}
\]

### Notations

- \( \psi_i^u \) - Feature vector.
- \( C \) - Regularization parameter.
- \( \eta \) - Weight vector.
- \( y_{ik}^u \) - Class label.
- \( \xi_{ik} \) - Non-negative slack variables.
Maximum Margin Supervised Topic Models (MedLDA)

- Topic model for document classification [Zhu et al., 2012].
- Regularizes the topic space with maximum margin constraints.
- Topic model parameter learning and maximum margin parameter learning are tightly integrated into a single model.
- Uses a latent linear discriminant function with a random weight vector.
- Encapsulates the topic feature weights.
- Learning is conducted iteratively:
  - Maximum margin solver.
  - Parameters of the topic model.
Study the **effectiveness** of the LDA model in ad-hoc retrieval task.

They proposed LDA-based document model within the language modeling framework.

**Basic Language Modeling for Ad-hoc IR**

\[
P(w|d) = \frac{N_d}{N_d + \mu} P_{ML}(w|d) + \left(1 - \frac{N_d}{N_d + \mu}\right) P_{ML}(w|collection)
\]

**LDA-based Language Modeling for Ad-hoc IR**

\[
P(w|d) = \lambda \left(\frac{N_d}{N_d + \mu} P_{ML}(w|d) + \left(1 - \frac{N_d}{N_d + \mu}\right) P_{ML}(w|collection)\right) + (1 - \lambda)(P_{LDA}(w|d))
\]

\[
P_{LDA}(w|d, \theta, \phi) = \sum_{z=1}^{K} P(w|z, \phi) P(z|\theta, d)
\]
**Our Model - Pairwise Regularized Bayesian Learning-to-Rank Model**

---

**Our Model** = Generative Learning + Discriminative Learning

- We propose an optimization algorithm to conduct learning-to-rank using topic models.
- We integrate the mechanism behind a basic topic model, LDA, with a maximum-margin pairwise learning-to-rank classifier.
- As a result, we obtain a unified learning-to-rank model, which conducts regularized Bayesian inference.
- In our model, the posterior distribution of the LDA model is regularized with a pairwise maximum margin classifier.
- This allows us to directly control the posterior distribution.
Posterior Regularization

**POSTERIOR REGULARIZATION - HIGH-LEVEL IDEA**

Bayes’ theorem

\[ P(M | x) = \frac{P(x | M)P(M)}{\int P(x | M)P(M)dM} \]

Likelihood

Prior

Posterior

Evidence of M given x
Let $M \in \mathcal{M}$ be the model space. The objective is to find the posterior distribution which we intend to infer. The posterior distribution of the Bayes’ Theorem is the solution to the following optimization problem:

**Bayes’ Optimization Model**

$$\text{minimize}_{P(M) \in \mathcal{M}} \quad \text{KL} \quad [P(M|x_1, x_2, \ldots, x_N)||P(M)] - \sum_{n=1}^{N} \int \log P(x_n|M)P(M) dM$$

subject to $P(M) \in \mathcal{P}$

Constraints on the posterior distribution.
Bayesian Posterior Regularization

**Bayesian Regularization**

Bayes’ Optimization Model

\[
\begin{align*}
\text{minimize} & \quad KL \quad [P(M|x_1, x_2, \ldots, x_N)||P(M)] - \sum_{n=1}^{N} \int \log P(x_n|M)P(M)\text{d}M \\
\text{subject to} & \quad P(M) \in \mathbb{P}
\end{align*}
\]

Bayesian Regularization

\[
\begin{align*}
\text{minimize} & \quad KL \quad [P(M|x_1, x_2, \ldots, x_N)||P(M)] - \sum_{n=1}^{N} \int \log P(x_n|M)P(M)\text{d}M + \\
& \quad f(\chi) \\
\text{subject to} & \quad P(M) \in \mathbb{P}(\chi)
\end{align*}
\]

\[\text{Can encode rich information or knowledge.}\]
**LDA as a Regularized Bayesian Model**

**Posterior of LDA**

\[
P(\Theta, Z, \Phi | W, \alpha, \beta) = \frac{P_0(\Theta, Z, \Phi | \alpha, \beta)P(W | \Theta, Z, \Phi)}{P(W | \alpha, \beta)},
\]

where,

- \( P_0(\Theta, Z, \Phi | \alpha, \beta) \) - Prior probability.
- \( w^d = \{w^d_n\}_{n=1}^{N^d} \)
- \( W = \{w^d\}_{d=1}^D \)
- \( z^d = \{z^d_n\}_{n=1}^{N^d} \)
- \( Z = \{z^d\}_{d=1}^D \)
- \( \Theta = \{\theta^d\}_{d=1}^D \)
- \( \Phi = \{\phi_1, \phi_2, \ldots, \phi_K\} = V \times K \) - Matrix of topic distribution parameters
- \( \alpha \) - Parameter of the Dirichlet prior on the per-document topic distributions
- \( \beta \) - Parameter of the Dirichlet prior on the per-topic word distributions
Regularized Bayesian Learning to Rank

Design Details

- Combine ideas from pairwise maximum margin learning to rank with topic models.
- Convex function can be borrowed from the maximum margin pairwise learning to rank model.
- We choose the basic topic model - LDA
- Formed an aggregated document set consisting of queries and documents.
- Incorporate the latent topic information in the form of a latent topic similarity between the document and the query during learning.

\[
\begin{align*}
\text{minimize} & \quad KL[P(\Theta, Z, \Phi | W, \alpha, \beta) || P_0(\Theta, Z, \Phi | \alpha, \beta)] - \mathbb{E}_P[\log P(W | \Theta, Z, \Phi)] \\
& + M(\xi) \\
\text{subject to} & \quad P(\Theta, Z, \Phi | \alpha, \beta) \in \mathbb{P}_{\text{new}}(\xi), \\
& \quad \xi \geq 0
\end{align*}
\]
Full Regularized Bayesian Learning to Rank Model

\[
\text{minimize} \quad \left\{ \begin{array}{l}
\text{KL}[P(\Theta, Z, \Phi | W, \alpha, \beta) || P_0(\Theta, Z, \Phi | \alpha, \beta)] - \mathbb{E}_P[\log P(W | \Theta, Z, \Phi)] \\
+ \frac{1}{2} (||\eta||^2 + \eta_t^2) + \frac{C}{X} \sum_{u=1}^{X} \frac{1}{|B_u|} \sum_{(i,k) \in B_u} \xi_{(i,k)}^u \\
\end{array} \right. \\
\text{subject to} \quad \xi_{(i,k)}^u \geq |h_i^u - h_k^u| - y_{ik}^u \left[ \eta^T (\psi_i^u - \psi_k^u) + \eta_t (\Upsilon_i^u - \Upsilon_k^u) \right], \forall u, i, k \\
P(\Theta, Z, \Phi | \alpha, \beta) \in \mathbb{P}(\xi_{(i,k)}^u) \\
\xi_{(i,k)}^u \geq 0, \forall u, i, k.
\]

Convex Function - Pairwise Maximum Margin

Dynamic Topic Similarity
SOLVING THE OPTIMIZATION MODEL

Iterative Procedure

1. Monte Carlo method - Collapsed Gibbs sampling.
   2. Iteratively compute:
      1. Maximum margin separation of the points, i.e. we determine $\eta$ and $\eta_i$ given $P(\Theta, Z, \Phi | W, \alpha, \beta)$.
      2. $P(\Theta, Z, \Phi | W, \alpha, \beta)$ given $(\eta, \eta_i)$.
3. Both steps are repeated for a given number of iterations or until the sampler converges to a steady state.

\[
P(Z | \alpha) \propto \frac{P(W, Z | \alpha, \beta)}{\Omega} \cdot \left( e^{\frac{1}{2} \Delta} - \Psi \right).
\]

$\Omega$ is the normalization constant.
**Sampler Formulations**

\[
P(Z|\alpha) \propto \prod_{z=1}^{K} \left( \frac{\Gamma\left(\sum_{s=1}^{\mid \beta \mid} \beta_{s}\right)}{\prod_{s=1}^{\mid \beta \mid} \Gamma(\beta_{s})} \int_{w=1}^{V} \phi_{zw}^{n_{zw}+\beta_{w}-1} d\phi_{z} \right) .
\]

Expanding using Dirichlet - Word-Topic statistics

\[
\prod_{d=1}^{D} \left( \frac{\Gamma\left(\sum_{s=1}^{\mid \alpha \mid} \alpha_{s}\right)}{\prod_{s=1}^{\mid \alpha \mid} \Gamma(\alpha_{s})} \int_{z=1}^{K} \theta_{dz}^{\rho_{dz}+\alpha_{z}-1} d\theta_{d} \cdot e^{\frac{1}{2} \Delta - \Psi} \right)
\]

Expanding using Dirichlet - Document-Topic statistics

**Pairwise Maximum Margin Regularizer**

\[
\Delta = \left[ \sum_{u=1}^{X} \sum_{(d,k) \in B_u} \sum_{\hat{u}=1}^{X} \sum_{(\hat{d}\hat{k}) \in B_{\hat{u}}} \lambda_{uk} \lambda_{d\hat{k}} \cdot \left[ (\psi_{d}^{u} - \psi_{k}^{u}) + (\gamma_{d}^{u} - \gamma_{k}^{u}) \right] \right] \cdot \left[ (\psi_{\hat{d}}^{\hat{u}} - \psi_{\hat{k}}^{\hat{u}}) + (\gamma_{\hat{d}}^{\hat{u}} - \gamma_{\hat{k}}^{\hat{u}}) \right]
\]

\[
\Psi = \sum_{u=1}^{X} \sum_{(d,k) \in B_u} \lambda_{dk} (h_{d}^{u} - h_{k}^{u})
\]
**Transition Equations**

\[
P(z_n^d = t | z_{-n}, w = v, w_{-n}, \alpha, \beta) \propto \left( \frac{n_{tv} + \beta_{w_n^d} - 1}{\left[ \sum_{v=1}^{V} n_{tv} + \beta_v \right] - 1} \right) \left( p_{dt} + \alpha_t - 1 \right) e^{\frac{1}{2} \Delta - \Psi}
\]

Word-topic statistics

Document-topic statistics

Regularizer

Posterior distributions are computed as:

**Building Word-Topic Matrix**

\[
\phi_{zv} = \frac{n_{zv} + \beta_v}{\left( \sum_{v=1}^{V} n_{zv} + \beta_v \right)}
\]

**Building Document-Topic Matrix**

\[
\theta_{dz} = \frac{p_{dz} + \alpha_z}{\left( \sum_{z=1}^{K} p_{dz} + \alpha_z \right)} e^{\frac{1}{2} \Delta - \Psi}
\]
**EXPERIMENTS AND RESULTS**

**Datasets**
- We need **original documents** to build a topic model.
- Many standard learning to rank test collections do not come with original documents, except OHSUMED.
- Used **Terrier Information Retrieval** platform to compute query-document features.
- Experimented on standard **TREC collections**:
  1. OHSUMED - 45 normalized features
  2. AQUAINT - 39 normalized features
  3. WT2G - 39 normalized features
  4. ClueWeb-2009 - 91 normalized features

**Train-Validate-Test Splits**
- **Training set** - approximately 60% of query-document pairs
- **Testing set** - approximately 20% of query-document pairs
- **Validation set** - approximately 20% of query-document pairs
### Comparative Models and Evaluation Metric

#### Comparative Models

- **Listwise** - Listnet, AdaRank-MAP, AdaRank-NDCG, Coordinate Ascent, LambdaMART, SVMMAP, and MART
- **Pairwise** - RankNet, RankBoost, RankSVM-Struct, and LambdaRank
- **Pointwise** - Linear Regression - L2 Norm, and Random Forests
- **DirectRank** - recently proposed [Tan et al., 2013].

#### Evaluation Metric

**Normalized Cumulative Discounted Gain (NDCG):**

\[
W(q_s) = \frac{1}{T_n} \sum_{i=1}^{n} \frac{2^{r(i)} - 1}{\log(1 + i)}
\]

- \(T_n\) - Normalization constant.
- \(r(i)\) - Rank label.
- \(n\) - Length of ranked list.
## Results - NDCG@10

<table>
<thead>
<tr>
<th>Models</th>
<th>AQUAINT</th>
<th>WT2G</th>
<th>ClueWeb09</th>
<th>OHSUMED</th>
</tr>
</thead>
<tbody>
<tr>
<td>ListNet</td>
<td>0.217</td>
<td>0.188</td>
<td>0.331</td>
<td>0.427</td>
</tr>
<tr>
<td>AdaRank-NDCG</td>
<td>0.292</td>
<td>0.489</td>
<td>0.211</td>
<td>0.431</td>
</tr>
<tr>
<td>AdaRank-MAP</td>
<td>0.253</td>
<td>0.486</td>
<td>0.212</td>
<td>0.426</td>
</tr>
<tr>
<td>SVMMAP</td>
<td>0.274</td>
<td>0.469</td>
<td>0.333</td>
<td>0.408</td>
</tr>
<tr>
<td>Coordinate Ascent</td>
<td>0.276</td>
<td>0.512</td>
<td>0.275</td>
<td>0.448</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>0.299</td>
<td>0.463</td>
<td>0.211</td>
<td>0.408</td>
</tr>
<tr>
<td>LambdaRank</td>
<td>0.061</td>
<td>0.266</td>
<td>0.224</td>
<td>0.238</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.213</td>
<td>0.511</td>
<td>0.214</td>
<td>0.430</td>
</tr>
<tr>
<td>RankSVM</td>
<td>0.288</td>
<td>0.465</td>
<td>0.345</td>
<td>0.309</td>
</tr>
<tr>
<td>RankBoost</td>
<td>0.262</td>
<td>0.478</td>
<td>0.292</td>
<td>0.421</td>
</tr>
<tr>
<td>DirectRank</td>
<td>0.295</td>
<td>0.496</td>
<td>0.381</td>
<td>0.432</td>
</tr>
<tr>
<td>MART</td>
<td>0.256</td>
<td>0.479</td>
<td>0.235</td>
<td>0.425</td>
</tr>
<tr>
<td>RankNet</td>
<td>0.262</td>
<td>0.438</td>
<td>0.363</td>
<td>0.439</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.281</td>
<td>0.471</td>
<td>0.266</td>
<td>0.433</td>
</tr>
<tr>
<td>Semantic Search</td>
<td>0.255</td>
<td>0.481</td>
<td>0.367</td>
<td>0.423</td>
</tr>
<tr>
<td><strong>Our Model</strong></td>
<td><strong>0.301</strong></td>
<td><strong>0.531</strong></td>
<td><strong>0.394</strong></td>
<td><strong>0.454</strong></td>
</tr>
</tbody>
</table>
## Results - NDCG@10

<table>
<thead>
<tr>
<th>Models</th>
<th>AQUAINT</th>
<th>WT2G</th>
<th>ClueWeb09</th>
<th>OHSUMED</th>
</tr>
</thead>
<tbody>
<tr>
<td>ListNet</td>
<td>0.217</td>
<td>0.194</td>
<td>0.344</td>
<td>0.431</td>
</tr>
<tr>
<td>AdaRank-NDCG</td>
<td>0.296</td>
<td>0.496</td>
<td>0.224</td>
<td>0.434</td>
</tr>
<tr>
<td>AdaRank-MAP</td>
<td>0.253</td>
<td>0.489</td>
<td>0.232</td>
<td>0.431</td>
</tr>
<tr>
<td>SVMMAP</td>
<td>0.278</td>
<td>0.472</td>
<td>0.348</td>
<td>0.415</td>
</tr>
<tr>
<td>Coordinate Ascent</td>
<td>0.276</td>
<td>0.515</td>
<td>0.275</td>
<td>0.449</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>0.299</td>
<td>0.465</td>
<td>0.245</td>
<td>0.415</td>
</tr>
<tr>
<td>LambdaRank</td>
<td>0.060</td>
<td>0.268</td>
<td>0.244</td>
<td>0.245</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.213</td>
<td>0.514</td>
<td>0.244</td>
<td>0.431</td>
</tr>
<tr>
<td>RankSVM</td>
<td>0.291</td>
<td>0.469</td>
<td>0.372</td>
<td>0.318</td>
</tr>
<tr>
<td>RankBoost</td>
<td>0.262</td>
<td>0.478</td>
<td>0.301</td>
<td>0.425</td>
</tr>
<tr>
<td>DirectRank</td>
<td>0.295</td>
<td>0.503</td>
<td>0.386</td>
<td>0.433</td>
</tr>
<tr>
<td>MART</td>
<td>0.295</td>
<td>0.481</td>
<td>0.266</td>
<td>0.426</td>
</tr>
<tr>
<td>RankNet</td>
<td>0.262</td>
<td>0.441</td>
<td>0.369</td>
<td>0.441</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.281</td>
<td>0.475</td>
<td>0.311</td>
<td>0.439</td>
</tr>
<tr>
<td>Semantic Search</td>
<td>0.255</td>
<td>0.489</td>
<td>0.367</td>
<td>0.423</td>
</tr>
<tr>
<td><strong>Our Model</strong></td>
<td><strong>0.301</strong></td>
<td><strong>0.531</strong></td>
<td><strong>0.394</strong></td>
<td><strong>0.454</strong></td>
</tr>
</tbody>
</table>
## Query-Level Performance

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>&gt;3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQUAINT</td>
<td>54.33</td>
<td>59.74</td>
<td>69.12</td>
<td>78.66</td>
</tr>
<tr>
<td>WT2G</td>
<td>59.64</td>
<td>67.35</td>
<td>73.92</td>
<td>73.92</td>
</tr>
<tr>
<td>ClueWeb</td>
<td>61.23</td>
<td>72.98</td>
<td>79.76</td>
<td>83.21</td>
</tr>
<tr>
<td>LETOR OHSUMED</td>
<td>57.34</td>
<td>63.65</td>
<td>69.44</td>
<td>75.61</td>
</tr>
</tbody>
</table>

**Table:** Query level performance in terms of the percentage of winning numbers for different query lengths.

**Winning Number is defined as follows**

\[
W_i = \sum_{j=1}^{n} \sum_{k=1}^{m} I(NDCG_i(j) > NDCG_k(j))
\]

**Notations**

- \( n \) - Number of datasets.
- \( k \) - Indexes of compared models.
CONCLUSIONS

- Novel LTR model that combines latent topic information with maximum margin learning in a unified way.
- Latent topic representation is directly regularized with a pairwise maximum margin constraint.
- Topic similarity is computed in the regularized latent topic space.
- As our experiments show, such an integrated approach outperforms the existing two-stage de-coupled approach to incorporating latent topics in LTR models.
REFERENCES


