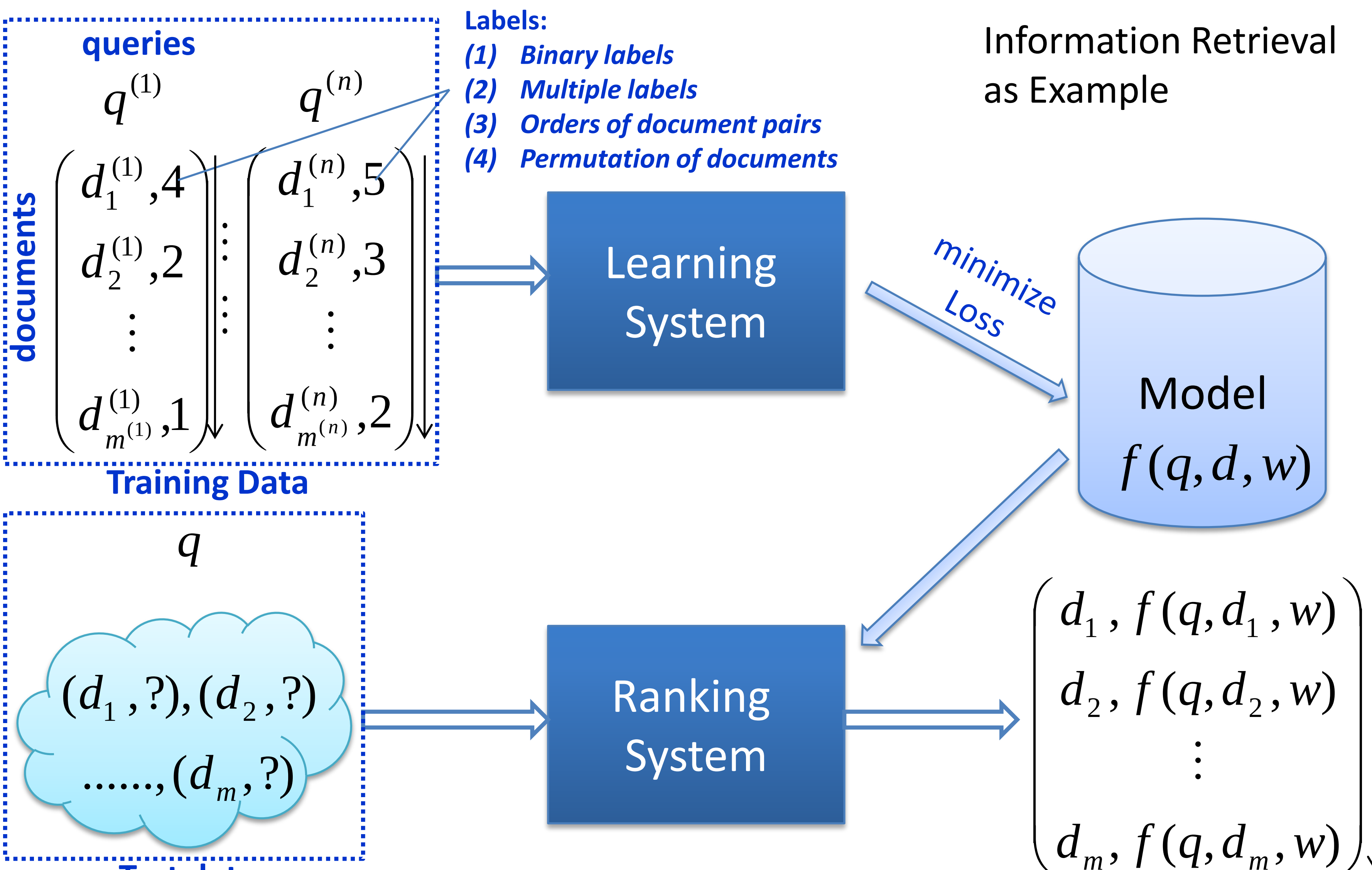
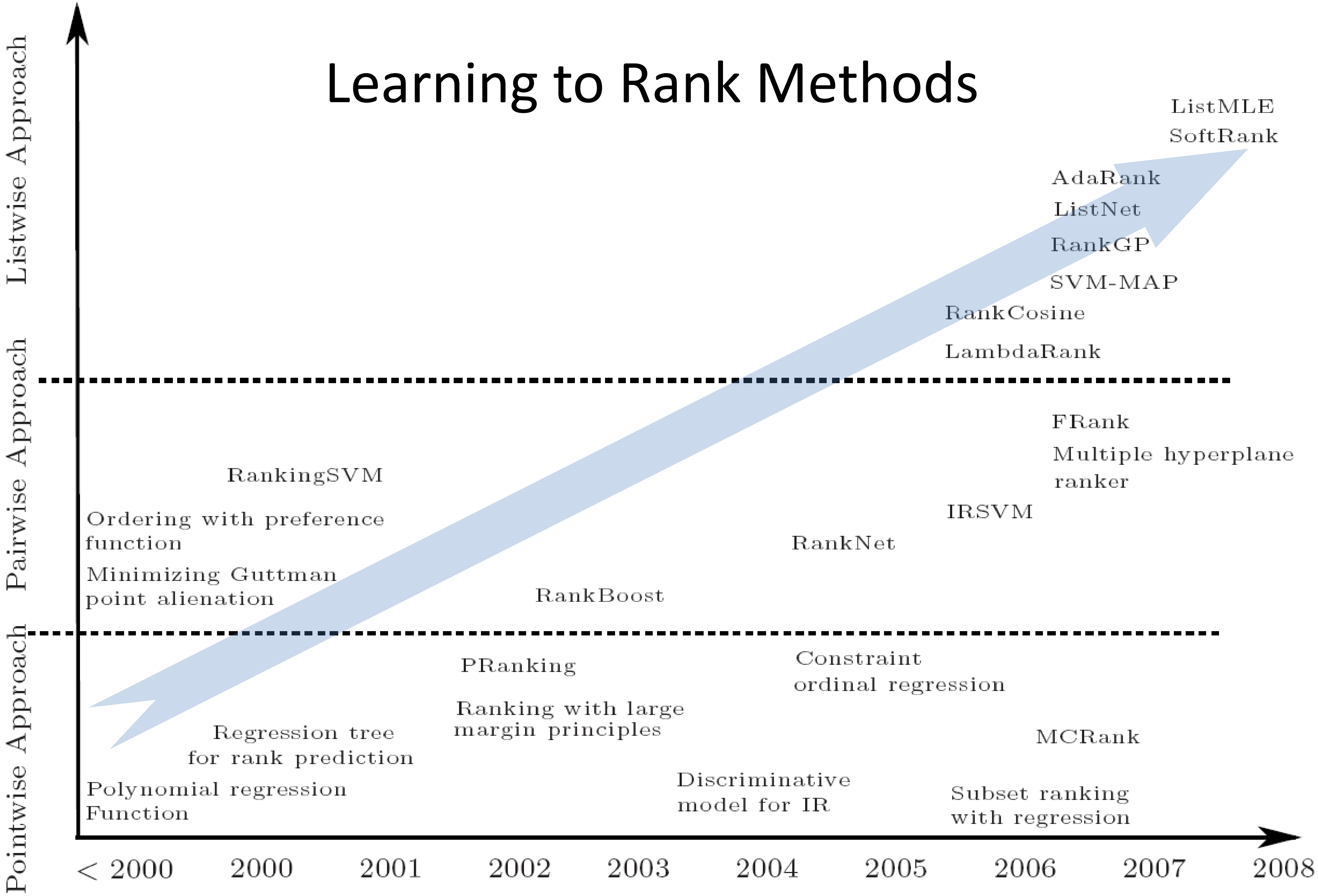
Generalization Analysis of Listwise Learning-to-Rank Algorithms

Yanyan Lan, Tie-Yan Liu, Zhiming Ma, and Hang Li
Microsoft Research Asia
Chinese Academy of Sciences

Learning to Rank



Learning to Rank Methods



Problem Studied in This Work

- Generalization ability of listwise algorithms

Related Work

- A two-layer learning framework was proposed and query-level generalization ability of pairwise algorithms was discussed

Our Contributions

- Proposal of listwise learning framework
- Analysis on generalization ability of listwise algorithms using Rademacher Average
- Analysis on generalization bounds of ListMLE, ListNet, RankCosine

Outline of Talk

- Introduction
- Listwise Algorithms
- Listwise Learning Framework
- Generalization Analysis of Listwise Algorithms
- Conclusion and Future Work

Pointwise, Pairwise, and Listwise Approaches

Same Model
Different Loss Functions

$$q\begin{pmatrix} x_1, y_1 \\ \vdots \\ x_m, y_m \end{pmatrix}$$

Pointwise Approach	Pairwise Approach	Listwise Approach
		Permutation
$S_q = \{(x_i, y_i)\}_{i=1}^{m}$	$S_q = \{(x_i, x_j) : y_i > y_j\}$	$S_q = \{(x_1, \cdots, x_m, y)\}$
$\sum_{i=1}^{m} l(f; x_i, y_i)$	$\sum_{(x_i, x_j) \in S_q} l(f; x_i, x_j, y_i, y_j)$	$l(f; x_1, \cdots x_m, y)$

Listwise Loss Functions

ϕ Transformation Function

z	$f(z)$	$g_y(z)$	$\phi(f(z))$	$\phi(g_y(z))$
x_1	$f(x_1)$	$g_y(x_1)$	$\phi(f(x_1))$	$\phi(g_y(x_1))$
x_2	$f(x_2)$	$g_y(x_2)$	$\phi(f(x_2))$	$\phi(g_y(x_2))$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	$f(x_m)$	$g_y(x_m)$	$\phi(f(x_m))$	$\phi(g_y(x_m))$

Lose Functions in Listwise Algorithms

- ListMLE $$l(f;z,y) = -\log P(y|z;f)$$

$$P(y|z;f) = \prod_{i=1}^{m} \frac{\phi(f(x_{y(i)}))}{\sum_{j=i}^{m} \phi(f(x_{y(j)}))}$$

- ListNet $$l(f;z,y) = -\sum_{\forall \pi \in \mathcal{Y}} P(\pi|z;g_y) \log P(\pi|z;f)$$

$$P(\pi|z;g_y) = \prod_{i=1}^{m} \frac{\phi(g_y(x_{\pi(i)}))}{\sum_{j=i}^{m} \phi(g_y(x_{\pi(j)}))} \quad P(\pi|z;f) = \prod_{i=1}^{m} \frac{\phi(f(x_{\pi(i)}))}{\sum_{j=i}^{m} \phi(f(x_{\pi(j)}))}$$

- RankCosine $$l(f;z,y) = \frac{1}{2}\left(1 - \frac{\phi(g_y(z))^T \phi(f(z))}{\|\phi(g_y(z))\| \|\phi(f(z))\|}\right)$$

Outline of Talk

- Introduction
- Listwise Algorithms
- Listwise Learning Framework
- Generalization Analysis of Listwise Algorithms
- Conclusion and Future Work

Listwise Learning Framework

- Data is represented as (z, y) , where z is feature vector set $z = (x_1, \cdots, x_m)$ and y is ground-truth permutation
- (z, y) are random variables according to distribution $P(\cdot, \cdot)$
- Training Data: $(z_1, y_1), \cdots, (z_n, y_n)$
- Expected Risk:

$$R_l(f) = \int_{\mathcal{Z} \times \mathcal{Y}} l(f; z, y) \, P(dz, dy)$$

- Empirical Risk:

$$\widehat{R}_l(f; S) = \frac{1}{n} \sum_{i=1}^{n} l(f; z_i, y_i)$$

Generalization Analysis

- Goal of learning = to minimize expected risk $R_l(f)$
- Distribution is unknown we instead minimize empirical risk $\widehat{R}_l(f;S)$
- Generalization analysis is concerned with upper bound of difference between expected and empirical risks $\sup_{f\in\mathcal{F}}(R_l(f) - \widehat{R}_l(f;S))$

Outline of Talk

- Introduction
- Listwise Algorithms
- Listwise Learning Framework
- Generalization Analysis of Listwise Algorithms
- Conclusion and Future Work

Our Analysis Technique

- Using Rademacher Average
- Generalization Bound based on Rademacher Average of Compound Function
- Further Deriving Bounds of the Rademacher Average for Different Algorithms

Rademacher Average

- For a function class $\mathcal{G}$, empirical Rademacher Average is defined as:

$$\widehat{\mathcal{R}}(\mathcal{G}) = E_\sigma \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(X_i)$$

where $X_i, i = 1, \cdots, n$ are i.i.d. random variables, and $\sigma_i, i = 1, \cdots, n$ are i.i.d. random variables, with probability ½ to take 1 or -1, σ stands for $\{\sigma_1, \cdots, \sigma_n\}$

Generalization Bound based on Rademacher Average of Compound Function

- Generalization bound based on Rademacher Average :

$$\sup_{f\in\mathcal{F}}(R_{l_\mathcal{A}}(f) - \widehat{R}_{l_\mathcal{A}}(f;S)) \leq 2\widehat{\mathcal{R}}(l_\mathcal{A}\circ\mathcal{F}) + \sqrt{\frac{2\ln\frac{2}{\delta}}{n}}$$

- $\widehat{\mathcal{R}}(l_\mathcal{A}\circ\mathcal{F})$ Rademacher Average of the compound function class, whose outer function is listwise loss function and inner function is ranking function.

$$\widehat{\mathcal{R}}(l_\mathcal{A}\circ\mathcal{F}) = E_\sigma \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i l_\mathcal{A}(f;z_i,y_i)$$

Upper Bounds of Rademacher Average

- Upper bounds of $\widehat{\mathcal{R}}(l_{\mathcal{A}} \circ \mathcal{F})$ of ListMLE, ListNet, and RankCosine can be represented as

$$\widehat{\mathcal{R}}(l_{\mathcal{A}} \circ \mathcal{F}) \leq C_{\mathcal{A}}(\phi) N(\phi) \widehat{\mathcal{R}}(\mathcal{F})$$

$$N(\phi) = \sup_{x \in [-BM, BM]} \phi'(x)$$

$$C_{ListMLE}(\phi) = \frac{2}{\phi(-BM)(\log m + \log \frac{\phi(BM)}{\phi(-BM)})}$$

$$C_{ListNet}(\phi) = \frac{2m!}{\phi(-BM)(\log m + \log \frac{\phi(BM)}{\phi(-BM)})}$$

$$C_{RankCosine}(\phi) = \frac{\sqrt{m}}{2\phi(-BM)}$$

$\widehat{\mathcal{R}}(\mathcal{F})$ has been studied in previous work, e.g., for linear function class, $\widehat{\mathcal{R}}(\mathcal{F}) \leq \frac{2BM}{\sqrt{n}}$, where $\forall x \in \mathcal{X}, \forall f \in \mathcal{F}, |f(x)| \leq BM$

Generalization Bound

- With probability at least $1-\delta$

$$\sup_{f\in\mathcal{F}}(R_{l_{\mathcal{A}}}(f)-\widehat{R}_{l_{\mathcal{A}}}(f;S))\leq\frac{4BM}{\sqrt{n}}C_{\mathcal{A}}(\phi)N(\phi)+\sqrt{\frac{2\ln\frac{2}{\delta}}{n}}$$

- The bound is related to:
  - $C_{\mathcal{A}}(\phi)$, algorithm-dependent factor, determined by loss function and transformation function ϕ
  - $N(\phi)$, algorithm-independent factor, only determined by transformation function ϕ
  - Order $O(\frac{1}{\sqrt{n}})$

Discussions

- When number of training samples → ∞, the generalization bounds → 0 at rate of $O(\frac{1}{\sqrt{n}})$
- When length of list ≥ 6, bound of ListMLE is tightest among the three algorithms
- In most cases, the use of a linear transformation function will result in a tighter bound than sigmoid and exponential transformation functions

Outline of Talk

- Introduction
- Listwise Algorithms
- Listwise Learning Framework
- Generalization Analysis of Listwise Algorithms
- Conclusion and Future Work

Conclusions

- Proposal of framework, which enables theoretical analysis on the listwise approach
- Proof of theorem that gives a general generalization bound of listwise ranking algorithms on the basis of Rademacher Average
- Investigations on generalization bounds of three listwise algorithms

Future Work

- Investigate approximation error - the difference between surrogate loss and true loss of ranking
- Experimentally verify the correctness of our theoretical findings
- Apply the proof technique to other approaches

Thank you!