Microsoft Research

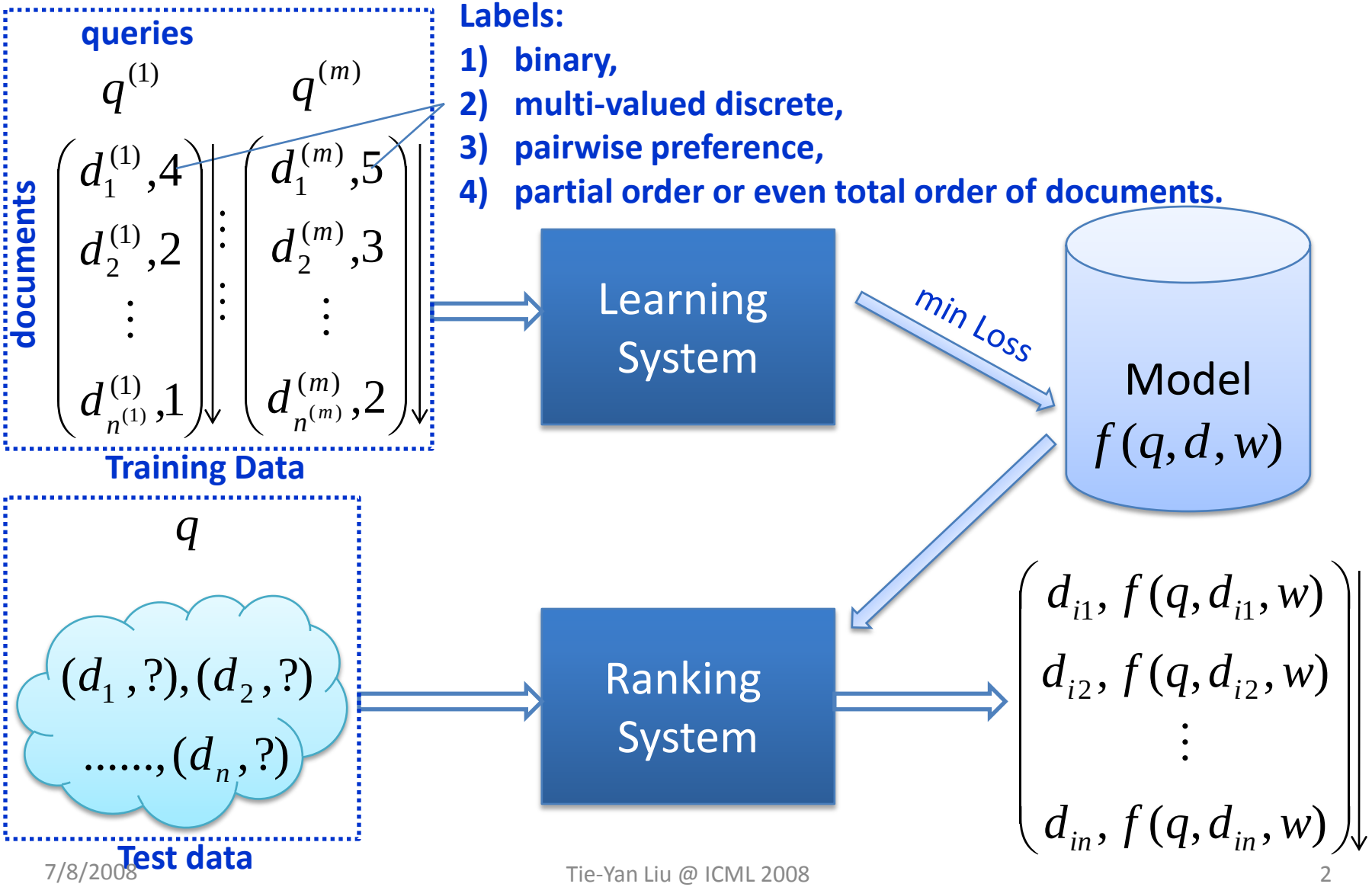
Query-Level Stability and Generalization in Learning to Rank

Yanyan Lan*, **Tie-Yan Liu**, Tao Qin, Zhiming Ma, Hang Li

Microsoft Research Asia

Chinese Academy of Science

Learning to Rank for Information Retrieval



State-of-the-art Approaches

- Pointwise
  - (Ordinal) regression / classification
  - Pranking, MCRank, etc.
- Pairwise
  - Preference learning
  - Ranking SVM, IR-SVM, RankNet, etc.
- Listwise
  - Direct optimization of IR measure
    - AdaRank, SVM-MAP, SoftRank, LambdaRank, etc.
  - Listwise loss minimization
    - RankCosine, ListNet, ListMLE, etc.

Question

- How about the generalization ability of these learning to rank algorithms?
- Since pointwise and pairwise approaches can be based on conventional machine learning technologies, corresponding theories can be applied to analyze their generalization ability.
  - Pointwise: classification / regression
  - Pairwise : pairwise classification / U-statistics
- However, are these expected risks really what one cares about in real applications?

Information Retrieval as Example

- In information retrieval, widely used evaluation measures are MAP, NDCG, etc.
- These measures are first calculated for a ranking model w.r.t. a particular query. Then the measures are averaged on all queries in the test collection.

$$NDCG(n) = \frac{1}{|Q|}\sum_{q \in Q} Z_n^q \sum_{j=1}^{n} (2^{r_q(j)} - 1)/\log(1+j))$$

Evaluation is conducted at query-level in information retrieval, and one does not care the errors at document or document pair level !!

Generalization in Learning to Rank for IR

- The generalization ability should also be analyzed at the query level.
- Existing work can only give the generalization ability at document level or document pair level, which is not consistent with the evaluation in information retrieval.
- New theory needs to be developed.

Our Work

- A two-layer probabilistic framework for ranking
- The definition of query-level losses and risks in learning to rank.
- The proposal of query-level stability, and a generalization theory for learning to rank based on it.
- The experimental verification of the correctness of the proposed theory.

Two-Layer Probabilistic Framework for Ranking

- Query – Associate
  - *q* stands for query, which is viewed as a random variable sampled from query space *Q* according to an unknown probability distribution P_Q.
  - $\left(w^{(q)}, g(w^{(q)})\right)$ stands for the associate of the query and its ground-truth, which is viewed as a random variable sampled from space $\Omega \times \mathcal{G}$ according to an unknown probability distribution D_q.

Associates in Different Approaches

Approach	$w^{(q)}$	$g(w^{(q)})$
Pointwise	Document	Relevance score
Pairwise	Document pair	Order
Listwise	Document set	Permutation

Training Data

$$\{(q_1, S_1), \cdots, (q_r, S_r)\}$$

$q_1, \cdots, q_r$ *i.i.d* $\sim P_Q$

$$S_i = \left(\left(w_1^{(i)}, g(w_1^{(i)})\right), \cdots, \left(w_{n_i}^{(i)}, g(w_{n_i}^{(i)})\right)\right)$$

$$\left(w_1^{(i)}, g(w_1^{(i)})\right), \cdots, \left(w_{n_i}^{(i)}, g(w_{n_i}^{(i)})\right) \text{ i.i.d } \sim D_{q_i}$$

Query-level Loss and Risk

- Expected Query-Level Loss

$$L(f;q) = \int_{\Omega \times \mathcal{G}} l(f;\omega^{(q)}, g(\omega^{(q)}))\, D_q(d\omega^{(q)}, dg(\omega^{(q)}))$$

- Empirical Query Level Loss

$$\hat{L}(f;q) = \frac{1}{n_q} \sum_{j=1}^{n_q} l(f;\omega_j^{(q)}, g(\omega_j^{(q)}))$$

- Expected Query-Level Risk

$$R_l(f) = E_{\mathcal{Q}} L(f;q) = \int_{\mathcal{Q}} L(f;q)\, P_{\mathcal{Q}}(dq)$$

- Empirical Query-Level Risk

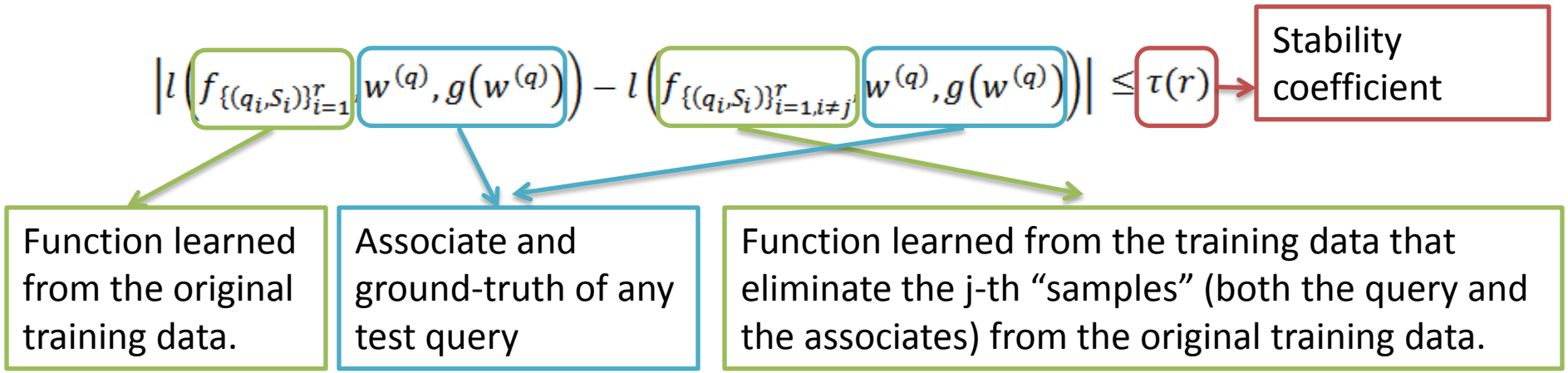
$$\widehat{R_l}(f) = \frac{1}{r} \sum_{i=1}^{r} \hat{L}(f;q_i)$$

Generalization in Learning to Rank for IR

- The goal of Learning to Rank is to minimize the expected query-level risk $R_l\left(f\right)$.
- However, as the distribution is unknown, one minimizes the empirical query-level risk $\widehat{R_l}\left(f\right)$.
- The generalization in learning to rank for IR is concerned with the bound of the difference between the expected and empirical query-level risks $R_l\left(f\right) - \widehat{R_l}\left(f\right)$.

Query-level Stability Theory

- Query-Level Stability
  - Query-level stability represents the degree of change in the loss of prediction when randomly removing a query and its associates from the training data.



Query-level Stability Theory

- Generalization Bound based on Query-Level Stability

$$R_l\left(f_{\{(q_i,S_i)\}_{i=1}^r}\right) \le \widehat{R_l}\left(f_{\{(q_i,S_i)\}_{i=1}^r}\right) + 2\tau(r) + (4r\tau(r) + B)\sqrt{\frac{\ln\frac{1}{\delta}}{2r}}$$

  - The bound is related to the number of training queries *r*, and the stability coefficient $\tau(r)$.
  - If $\tau(r) \to 0$ as $r \to \infty$, then as *r* tends to infinity, the bound tends to zero, i.e. the generalization ability is good.

Case Study

- Apply the query-level stability theory to analyze the generalization ability of the pairwise approach.
- We take support vector based algorithms as examples.
  - Ranking SVM
    (*Joachims, T. KDD 2002; Herbrich, R., Obermayer, K., et al. ICANN 1999*)
  - IRSVM
    (*Cao, Y., Xu, J., et al. SIGIR 2006; Qin, T., Liu, T., et al. IP&M 2007*)

The Algorithms under Investigation

- Ranking SVM

$$\min \frac{1}{\sum_{i=1}^{r} n_i} \sum_{i=1}^{r} \sum_{j=1}^{n_i} l_h(f; z_j^{(i)}, y_j^{(i)}) + \lambda \|f\|_K^2$$

Hinge Loss

Document pair and order

- IRSVM

$$\min \frac{1}{r} \sum_{i=1}^{r} \frac{1}{n_i} \sum_{j=1}^{n_i} l_h(f; z_j^{(i)}, y_j^{(i)}) + \lambda \|f\|_K^2$$

IRSVM introduces query-level normalization to the loss function of Ranking SVM.

Stability of the Algorithms

- Ranking SVM

$$\tau(r) = \frac{4\kappa^2}{\lambda r} \frac{1+\frac{\sigma}{\mu\sqrt{\frac{\delta}{r}}}}{1-\frac{\varepsilon}{\mu}} \qquad O(\frac{1}{\sqrt{r}})$$

- IRSVM

$$\tau(r) = \frac{4\kappa^2}{\lambda r} \qquad O(\frac{1}{r})$$

Detailed deductions can be found in the paper.

Generalization Bounds of the Algorithms

- Ranking SVM

$$R_l\left(f_{\{(q_i,S_i)\}_{i=1}^r}\right) \le \widehat{R_l}\left(f_{\{(q_i,S_i)\}_{i=1}^r}\right) + \frac{8\kappa^2}{\lambda r}\frac{1+\frac{\sigma}{\mu\sqrt{\frac{\delta}{r}}}}{1-\frac{\varepsilon}{\mu}} + \left(\frac{16\kappa^2\frac{1+\frac{\sigma}{\mu\sqrt{\frac{\delta}{r}}}}{1-\frac{\varepsilon}{\mu}} + \lambda(1+2C\kappa)}{\lambda}\right)\sqrt{\frac{\ln\frac{1}{\delta}}{2r}}$$

much looser

- IRSVM

$$R_l\left(f_{\{(q_i,S_i)\}_{i=1}^r}\right) \le \widehat{R_{l_h}}\left(f_{\{(q_i,S_i)\}_{i=1}^r}\right) + \frac{8\kappa^2}{\lambda r} + \frac{16\kappa^2+\lambda(1+2C\kappa)}{\lambda}\sqrt{\frac{\ln\frac{1}{\delta}}{2r}}$$

much tighter

IV

Discussions

- When *r* tends to infinity, the upper bound of IRSVM will tends to zero and the convergent rate is $O\left(\frac{1}{\sqrt{r}}\right)$
- For Ranking SVM, as *r* tends to infinity, the upper bound may be a constant: $r\tau(r)\sqrt{\frac{1}{r}} = O(1)$
- In other words, for Ranking SVM, with infinite number of training queries, there is still a gap between the query-level empirical and query-level expected risks.

Experiment (1)

- Query-Level Stability
  - 1200 queries, 200 for training, 500 for validation and 500 for testing, labeled as “relevant” and “irrelevant”.
  - First train two models using Ranking SVM and IRSVM respectively, denoted as f_0 and f_0'.
  - Randomly remove a query from the training set, and train two new models, denoted as f_i and f_i'.

$$\Delta_i = \max_{q \in \mathcal{T}} \max_{z \in S_q} |l_h(f_0, z^{(q)}, y^{(q)}) - l_h(f_i, z^{(q)}, y^{(q)})|,$$

$$\Delta_i' = \max_{q \in \mathcal{T}} \max_{z \in S_q} |l_h(f_0', z^{(q)}, y^{(q)}) - l_h(f_i', z^{(q)}, y^{(q)})|.$$

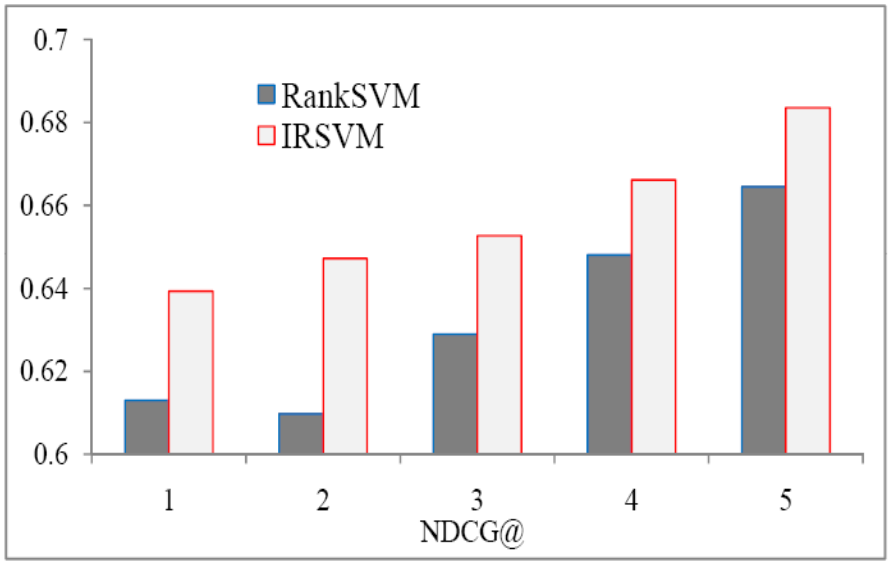
Experiment (1)

Table 1. Comparison of Query-level Stability

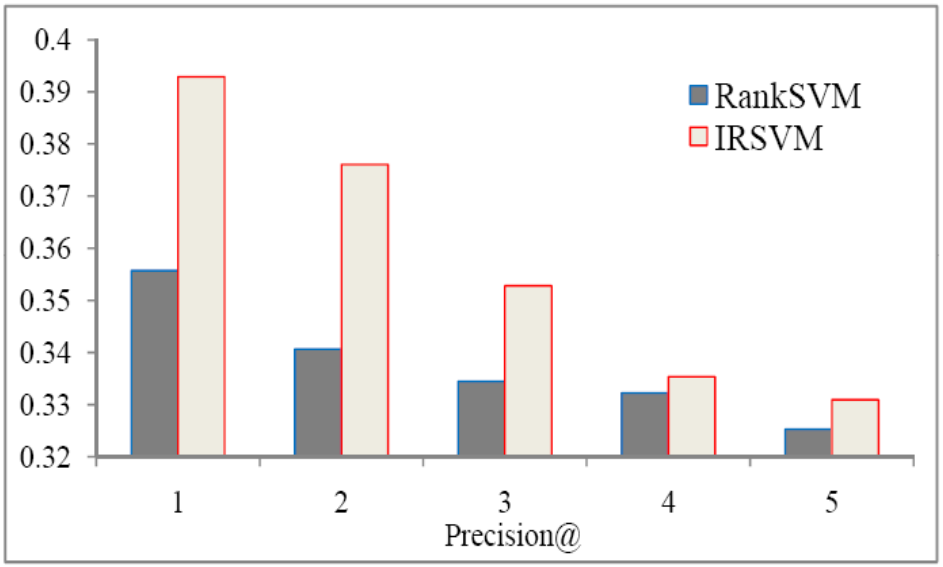
i	1	2	3	4	5	6
Δ_i	3.59	1.14	0.88	0.81	1.84	1.15
Δ_i'	0.07	0.07	0.06	0.06	0.05	0.24
	7	8	9	10	11	12
	0.89	1.30	0.90	1.42	1.38	1.39
	0.18	0.06	0.09	0.08	0.11	0.15
	13	14	15	16	17	18
	0.56	1.43	1.42	1.01	1.13	1.34
	0.11	0.13	0.14	0.11	0.06	0.11
	19	20	21	22	23	24
	1.04	0.86	0.43	0.51	0.64	0.92
	0.08	0.05	0.09	0.20	**0.27**	0.14
	25	26	27	28	29	30
	0.50	0.88	**4.53**	0.99	1.13	0.62
	0.18	0.08	0.12	0.09	0.21	0.14

Experiment (2)

- Ranking performance



(a) NDCG@1-5



(b) Precision@1-5

Conclusions

- A proposal on conducting generalization analysis on learning to rank algorithms at query level is made.
- A two-layer probabilistic formulation of learning to rank is proposed.
- A new methodology for analyzing generalization ability of learning to rank algorithms on the basis of query-level stability is proposed.
- The proposed theory is applied to learning to rank algorithms of Ranking SVM and IRSVM. The correctness of the theory has been verified by experiments.

Future Work

- We have taken SVM based ranking algorithms as examples. It is interesting to know whether we can obtain similar results for other algorithms, such as RankBoost, etc.
- The proposed formulation for ranking and the tool of query-level stability can also be used to analyze the generalization ability of other approaches.
- It is worth checking whether new learning to rank algorithms can be derived under the guide of the theoretical study.

Acknowledgement:
Liwei Wang (Peking Univ.)

Tutorial on Learning to Rank @ SIGIR 2008
Workshop on Learning to Rank @ SIGIR 2008

tyliu@microsoft.com
http://research.microsoft.com/users/tyliu/

Comparison with Previous Work

	Assumptions on i.i.d.	Labels compatible
Classification	Documents Do not distinguish different queries	Binary Multi-valued discrete
Pairwise classification	Document pairs Do not distinguish different queries	Binary Multi-valued discrete Pairwise preferences
U-statistics	Documents with the same label No assumption on the i.i.d. of associates. Do not distinguish different queries	Binary Multi-valued discrete
Two-Layer framework	Queries Associates conditioned on query	Binary Multi-valued discrete Pairwise preference Partial or total order