TFANT'07
Tokyo Univ.
March 12, 2007

Learning to Rank: A New Technology for Text Processing

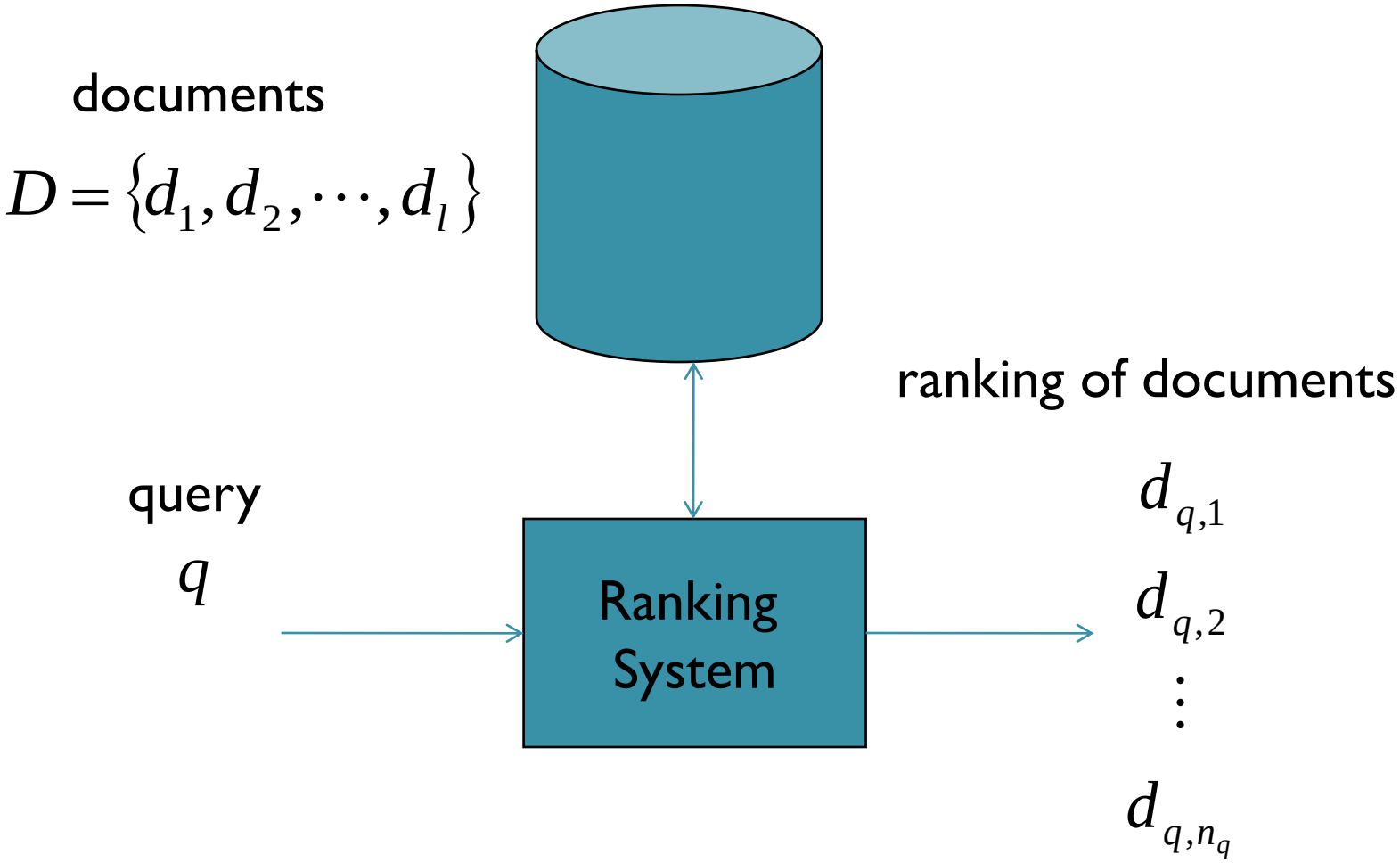
Hang Li
Microsoft Research Asia

Talk Outline

- What is 'Learning to Rank'?
- Ranking SVM
- Definition Search
- Ranking SVM for IR
- Summary

WHAT IS LEARNING TO RANK?

Ranking Problem: Example = Document Retrieval



Ranking in Information Retrieval

- Document Retrieval
- Collaborative Filtering
- Key Term Extraction
- Expert Finding
- ……

Means for Information Access

Information Extraction

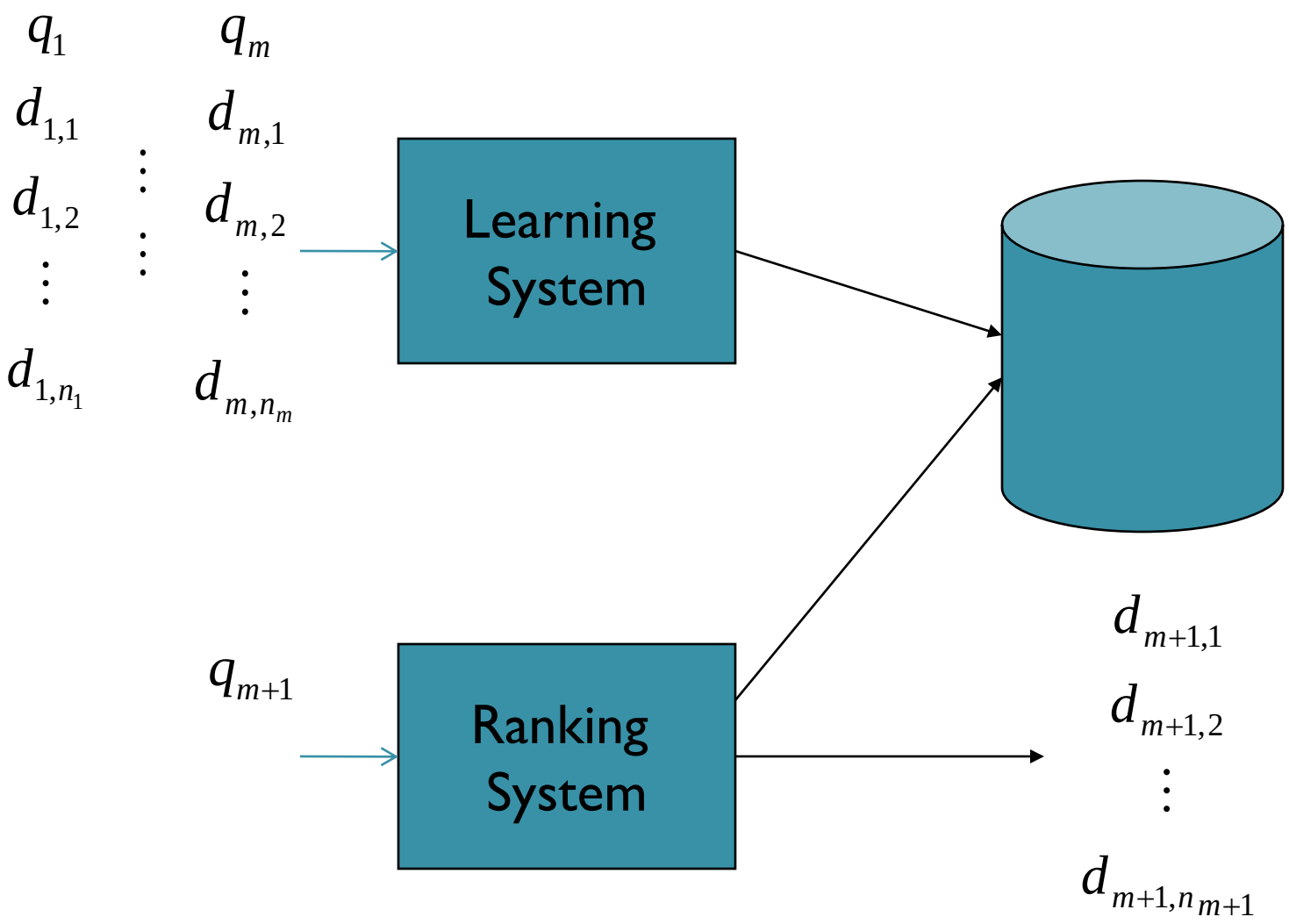
Ranking

Multi-document Summarization

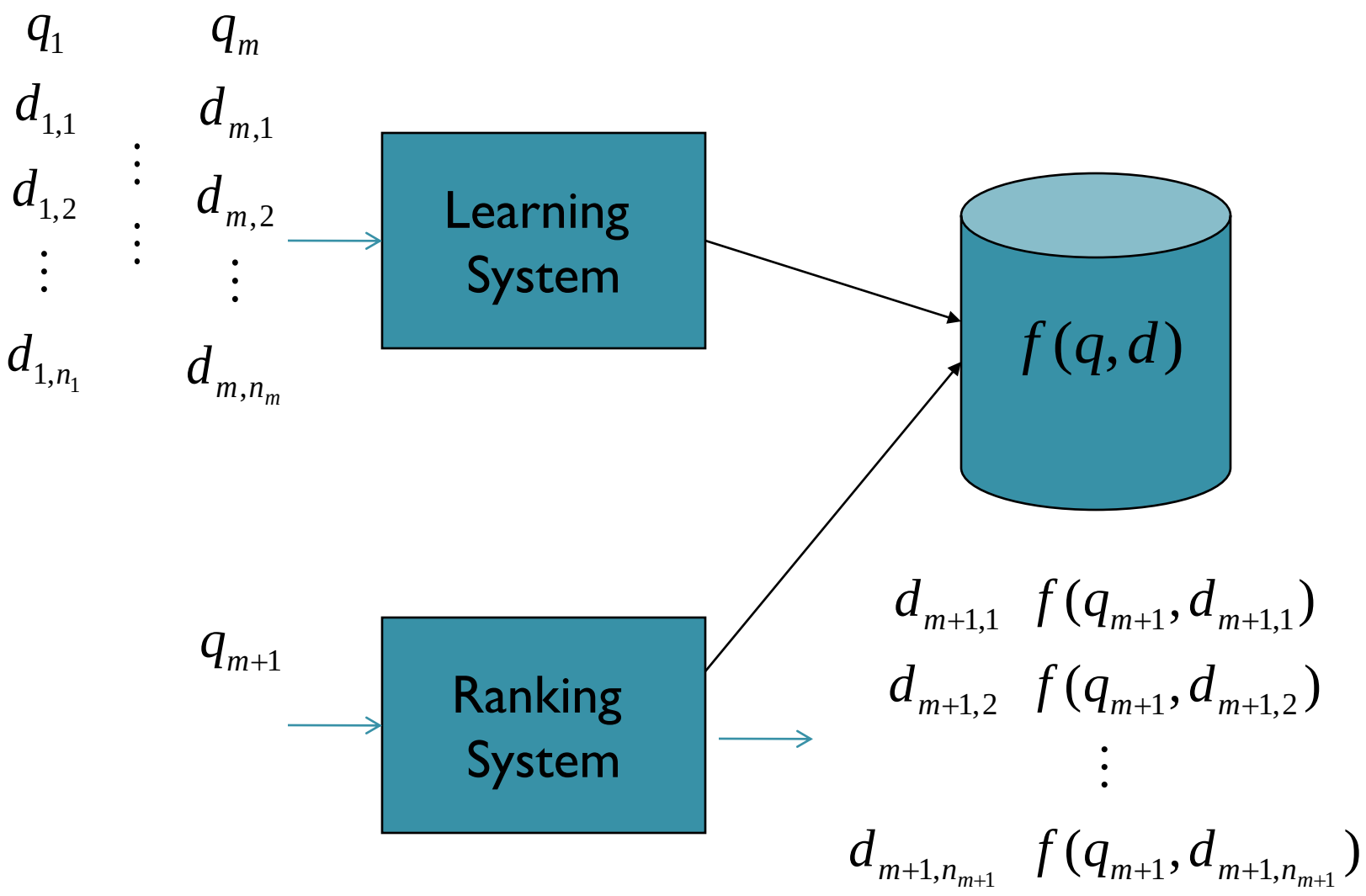
NLP and TM Problems Can Be Formalized as Ranking

- Machine Translation
- Paraphrasing
- Sentiment Analysis
- ……

Learning to Rank



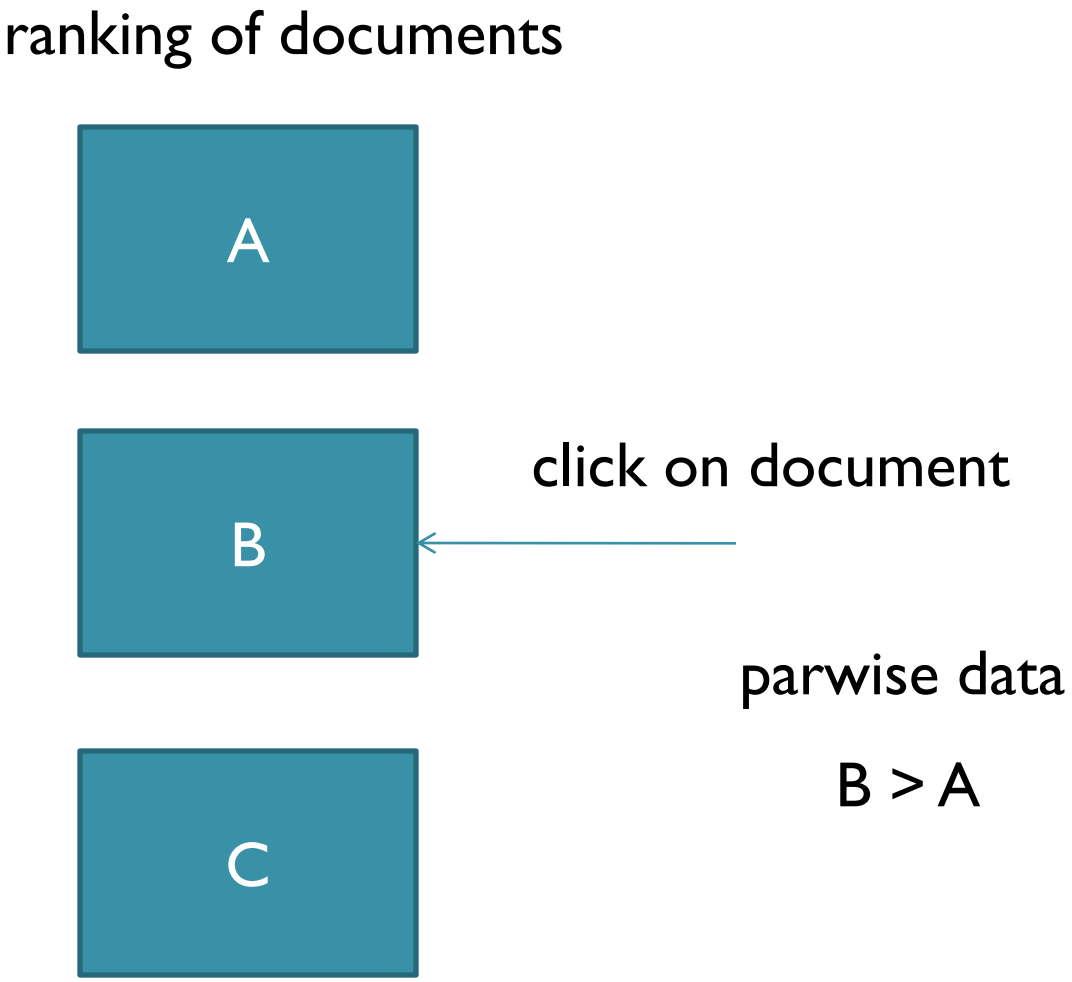
Score-based Ranking



Data Labeling Methods

- Listwise
- Pointwise
  - Score
  - Rank (e.g., relevant, partially relevant, irrelevant)
- Pairwise

Pairwise Data Labeling Method Joachims (2002)



Evaluation Measures

- MRR (Mean Reciprocal Rank)
- MAP (Mean Average Precision)
- NDCG (Normalized Discounted Cumulative Gain)
- Kendall's Tau

Kendall's Tau

$$\tau = \frac{2P}{\frac{1}{2}n(n-1)} - 1$$

- Number of pairs $\frac{1}{2}n(n-1)$
- Number of concordant pairs P
- Example

  A B C

  C A B

$$\tau = \frac{2 \times 1}{3} - 1 = -\frac{1}{3}$$

NDCG

- query: q_i
- DCG at position *m*: $N_i = n_i \sum_{j=1}^{m} (2^{r(j)} - 1) / \log(1 + j)$
- NDCG at position *m*: average over queries
- Example
  - (3, 3, 2, 2, 1, 1, 1) rank *r*
  - (7, 7, 3, 3, 1, 1, 1) gain $2^{r(j)} - 1$
  - (1, 0.63, 0.5, 0.43, 0.39, 0.36, 0.33) discount $1/\log(1 + j)$
  - (7, 11.41, 12.91, 14.2, 14.59, 14.95, 15.28)

$$\sum_{j=1}^{m} (2^{r(j)} - 1) / \log(1 + j)$$

RANKING SVM

Ranking SVM
Herbrich et al (2000)

- Input space: X
- Ranking function $f : X \to R$
- Ranking: $x_i \succ x_j \Leftrightarrow f(x_i; w) > f(x_j; w)$
- Linear ranking function: $f(x; w) = \langle w, x \rangle$

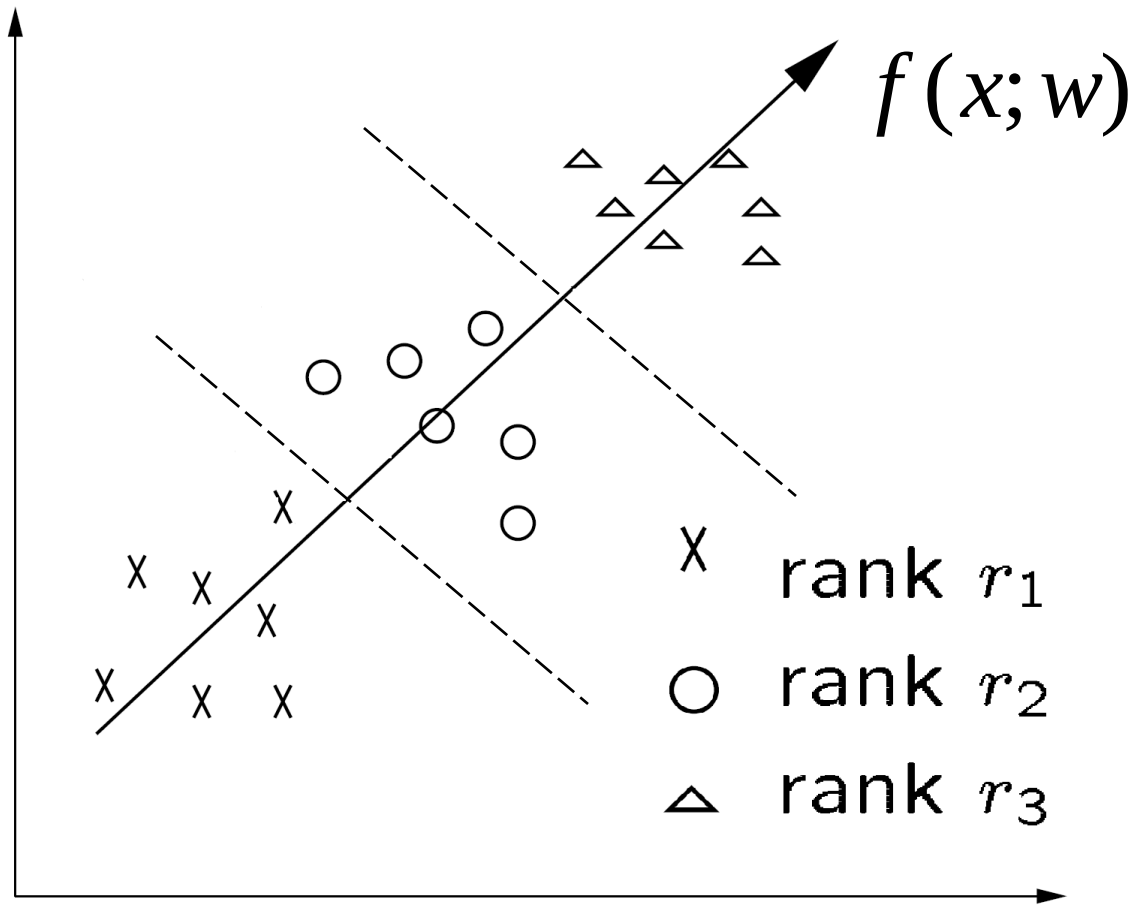
$$\langle w, x^{(1)} - x^{(2)} \rangle > 0 \quad \Leftrightarrow \quad f(x^{(1)}; w) > f(x^{(2)}; w)$$

- Transforming to binary classification:

$$(\vec{x}^{(1)} - \vec{x}^{(2)}, z), \; z = \begin{cases} +1 & x^{(1)} \succ x^{(2)} \\ -1 & x^{(2)} \succ x^{(1)} \end{cases}$$

Ranking SVM (cont')

- Ranking Model: $f(x;w)$

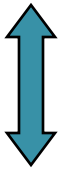


Ranking SVM (cont')

$$\min_{w,\xi} \frac{1}{2} \| w \|^2 + C\xi_i$$

$$z_i \left\langle w, x_i^{(1)} - x_i^{(2)} \right\rangle \geq 1 - \xi_i$$

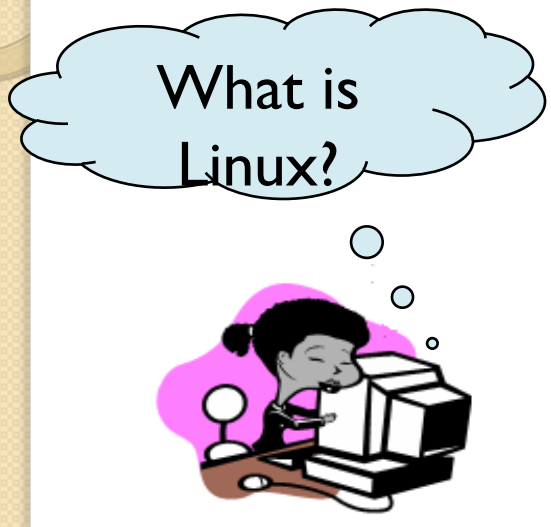
$$\xi_i \geq 0$$



$$\min_{w} \sum_{i=1}^{l} \left[1 - z_i \left\langle w, x_i^{(1)} - x_i^{(2)} \right\rangle \right]_+ + \lambda \| w \|^2$$

DEFINITION SEARCH

Definition Search (Xu 2005)



Linux is an excellent product.

Linux is a Unicode platform.

Linux is a free Unix-type operating system originally created by Linus Torvalds with the assistance of developers around the world.

Linux is an open source operating system that was derived from Unix in 1991.

Linux is the platform for the communication application for the leader network

Definitions Can Be Ranked

- Good definition
  - Contain general notion and important properties of term
  - E.g., "Linux is a Unix-based operating system that was developed in 1991 by Linus Torvalds, then a student in Finland."
- Indifferent definition
  - Between 'good' and 'bad'
  - E.g., "Linux is the best-known product distributed under the GPL."
- Bad definition
  - Option, impression, or feeling of people about term
  - E.g., "Linux is an excellent product."

Extracting and Ranking Definition Candidates

- Identifying term of each paragraph
  - First Base NP of first sentence
  - Two Base NPs separated by 'of' or 'for'
- Extracting definition candidates
  - <term> is a|an|the *
  - <term>, *, a|an|the *
  - <term> is one of *
- Ranking definition candidates using Ranking SVM

Features in Ranking SVM

1. <*term*> occurs at the beginning of the paragraph.
2. <*term*> begins with ‘the’, ‘a’, or ‘an’.
3. All the words in <*term*> begin with uppercase letters.
4. Paragraph contains predefined negative words, e.g. ‘he’, ‘she’, ‘said’
5. <*term*> contains pronouns.
6. <*term*> contains ‘of’, ‘for’, ‘and’, ‘or‘ or ‘,’.
7. <*term*> re-occurs in the paragraph.
8. <*term*> is followed by ‘is a ’, ‘is an’ or ‘is the ’.
9. Number of sentences in the paragraph.
10. Number of words in the paragraph.
11. Number of the adjectives in the paragraph.
12. Bag of words: words frequently occurring within a window after <*term*>

Experimental Results on Ranking Definitional Paragraphs

	Error Rate	R-precision	Top 1 Precision	Top 3 Precision
BM25	0.513	0.289	0.221	0.642
Random Ranking	0.436	0.322	0.347	0.632
Ranking SVM	**0.271**	**0.518**	**0.550**	**0.887**

RANKING SVM FOR IR

Direct Application of Ranking SVM to Document Retrieval

- Query and document → feature vector
- Combining instance pairs from all queries

Problems with Direct Application

- Top sensitiveness

  d: *definitely relevant*, p: *partially relevant*, n: *not relevant*

  ranking 1: p d p n n n n

  ranking 2: d p n p n n n

- Query normalization

  q1: d p p n n n n

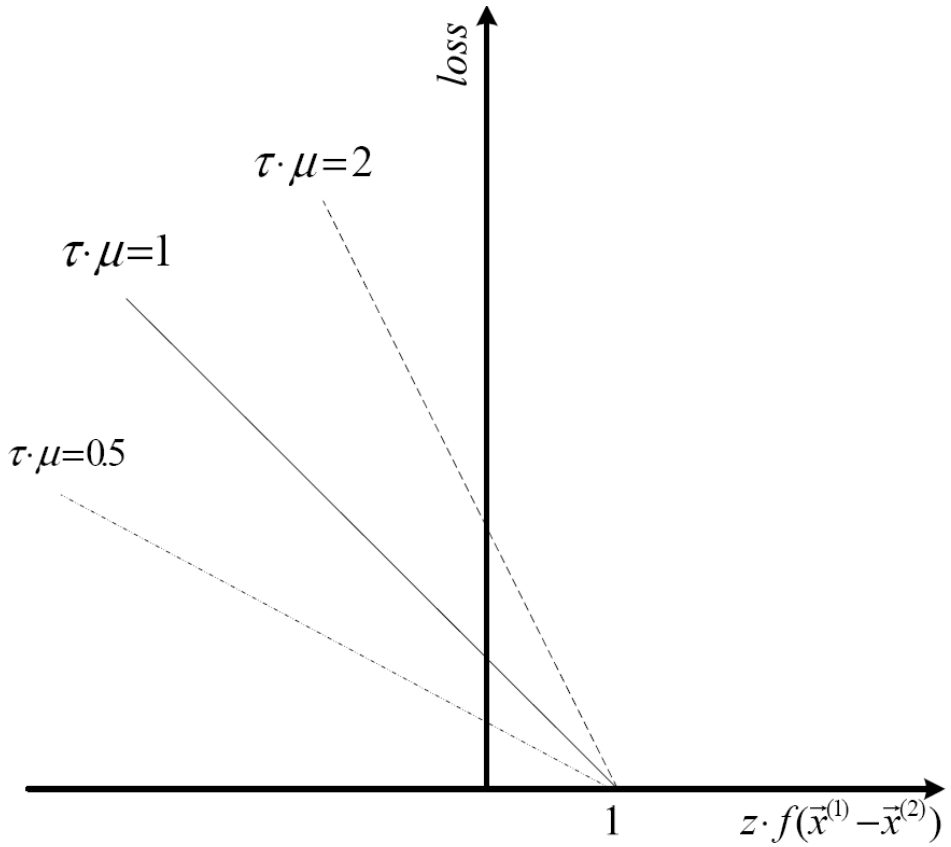
  q2: d d p p p n n n n n

  q1 pairs: 2*(d, p) + 4*(d, n) + 8*(p, n) = 14

  q2 pairs: 6*(d, p) + 10*(d, n) + 15*(p, n) = 31

New Loss function

$$\min_{\vec{w}} L(\vec{w}) = \sum_{i=1}^{l} \tau_{k(i)} \mu_{q(i)} \left[1 - z_i \left\langle \vec{w}, \vec{x}_i^{(1)} - \vec{x}_i^{(2)} \right\rangle \right]_+ + \lambda \left\| \vec{w} \right\|^2$$



Ranking SVM for IR (Cao et al 2006)

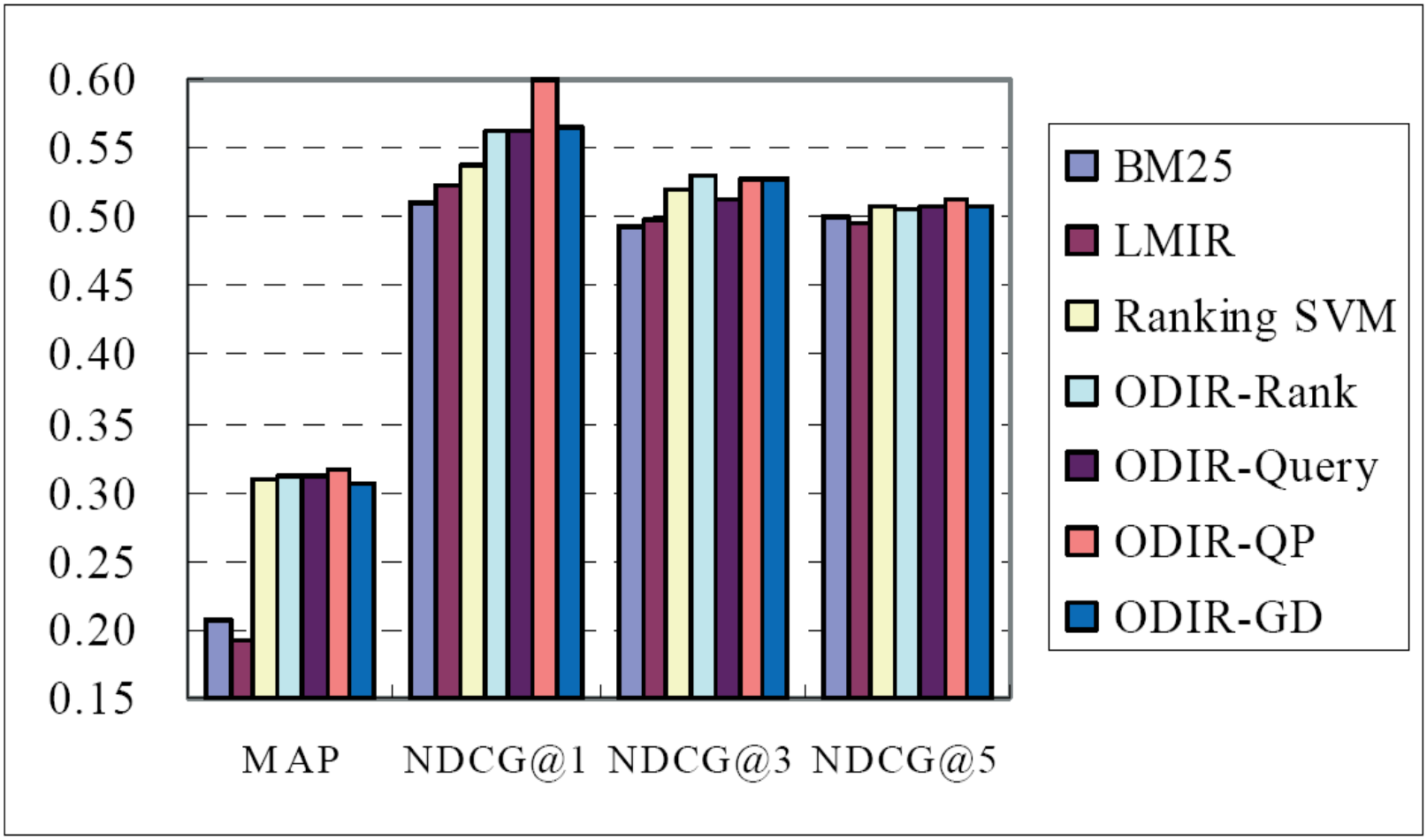
$$\min_{\vec{w}} L = \sum_{i=1}^{\ell} \tau_{k(i)} \mu_{q(i)} \left[1 - z_i \left\langle \vec{w}, \vec{x}_i^{(1)} - \vec{x}_i^{(2)} \right\rangle \right]_+ + \lambda \left\| \vec{w} \right\|^2 ,$$

$$\min_{\vec{w}} M(\vec{w}) = \frac{1}{2} \left\| \vec{w} \right\|^2 + \sum_{i=1}^{\ell} C_i \cdot \xi_i$$

$$\text{subject to } \xi_i \geq 0, \quad z_i \left\langle \vec{w}, \vec{x}_i^{(1)} - \vec{x}_i^{(2)} \right\rangle \geq 1 - \xi_i \quad i = 1, \cdots, \ell$$

$$\textit{where } C_i = \frac{\tau_{k(i)} \mu_{q(i)}}{2\lambda}$$

Experimental Results (OHSUMED)



SUMMARY

Summary

- What is ‘Learning to Rank’?
- Ranking SVM
- Definition Search
- Ranking SVM for IR
- Summary

Future Directions

- Inventing new algorithms
  - e.g., directly optimizing evaluation measure
- Developing new labeling methods
- Finding more appropriate evaluation measures
- Applying to new applications

References

- Herbrich, R., Graepel, T., & Obermayer, K. (2000). Large Margin Rank Boundaries for Ordinal Regression. . Advances in Large Margin Classifiers (pp. 115-132).
- Joachims T. (2002), Optimizing Search Engines Using Clickthrough Data, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining.
- Jun Xu, Yunbo Cao, Hang Li, and Min Zhao, Ranking Definitions with Supervised Learning Method, Proc. of WWW 2005 industry track, 811-819.
- Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon, Adapting Ranking SVM to Document Retrieval, Proc. of SIGIR 2006, 186-193.
- Yu-Ting Liu, Tie-Yan Liu, Tao Qin, Zhi-Ming Ma, Hang Li, Supervised Rank Aggregation, Proc. of WWW-2007, to appear, 2007.

THANK YOU