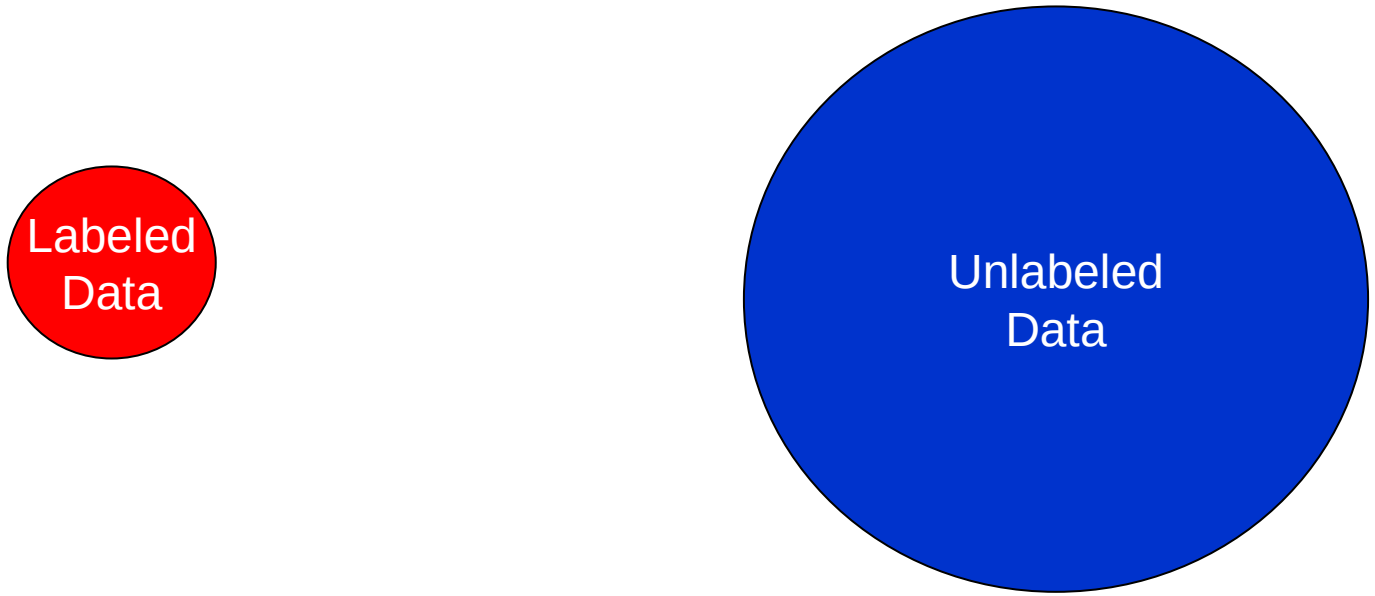
Learning to Rank with Partially-Labeled Data

Kevin Duh
University of Washington

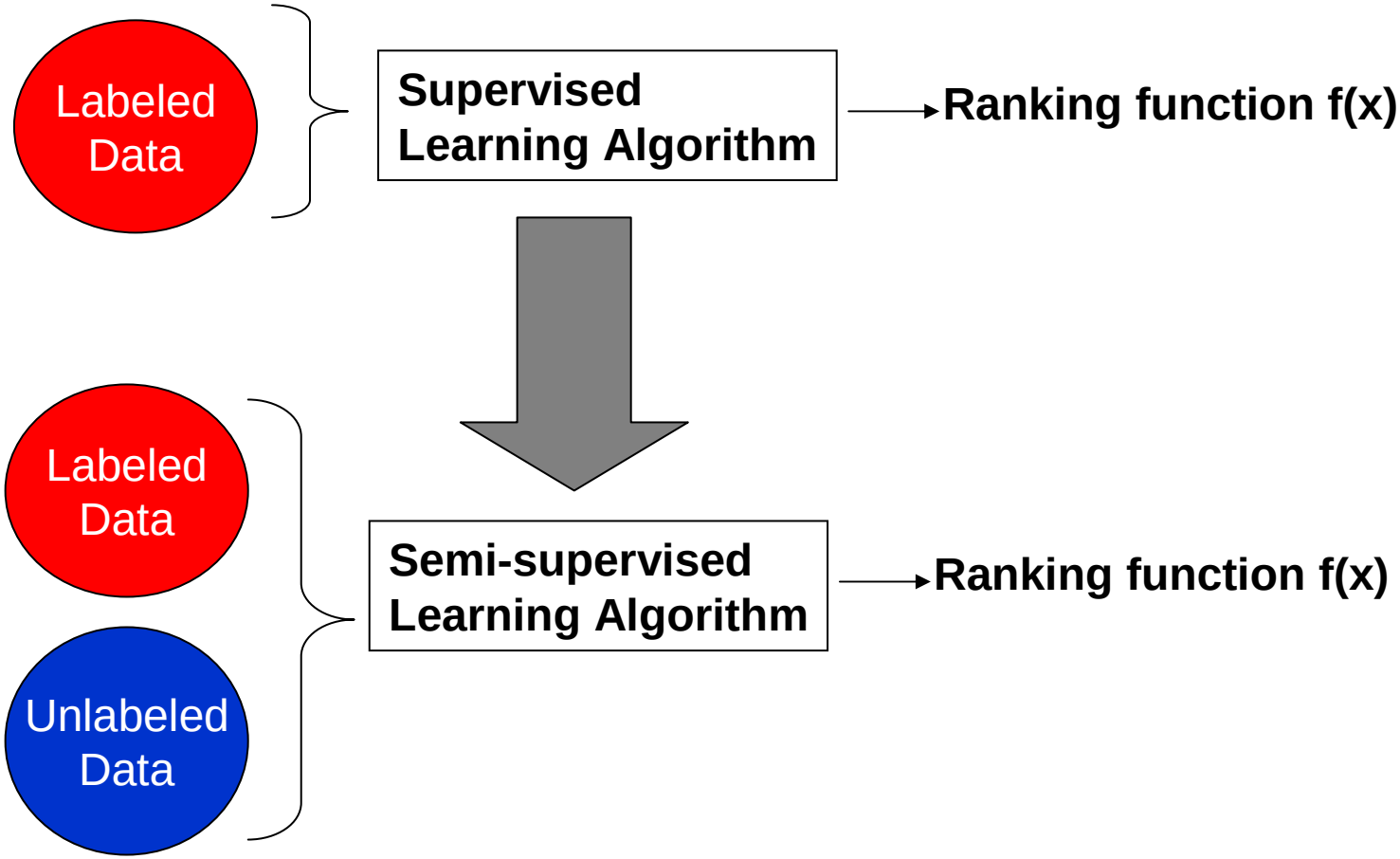
(Joint work with Katrin Kirchhoff)

Motivation

- Machine learning can be an effective solution for ranking problems in IR
  - But success depends on **quality** and **size** of training data



Problem Statement



Can we build a better ranker by adding cheap, unlabeled data?

Outline

1. Problem Definition
   1. Ranking as a Supervised Learning Problem
   2. Two kinds of Partially-labeled Data
2. Proposed Method
3. Results and Analysis

Ranking as Supervised Learning Problem

Query: SIGIR

Labels

ACM SIGIR Special Interest Group on Information Retrieval Home Page - [翻譯此頁]
"Addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, storage, retrieval, and distribution ...
www.sigir.org/ - 10k - 頁庫存檔 - 類似網頁

3 $x_1^{(1)} = [tfidf, pagerank, ...]$

SIGIR 2004 - [翻譯此頁]
The 27th Annual International ACM SIGIR Conference will be held at The University of Sheffield, UK, from July 25 to July 29, 2004.
www.sigir.org/sigir2004/ - 9k - 頁庫存檔 - 類似網頁

1 $x_2^{(1)} = [tfidf, pagerank, ...]$

Special Inspector General for Iraq Reconstruction :SIGIR Homepage - [翻譯此頁]
Welcome to the Office of the Special Inspector General for Iraq Reconstruction (SIGIR), a temporary federal agency serving the American public as a watchdog ...
www.sigir.mil/ - 20k - 頁庫存檔 - 類似網頁

2 $x_3^{(1)} = [tfidf, pagerank, ...]$

Query: Hotels in Singapore

Singapore Hotels | All Hotels in Singapore Reservation Service ... - [翻譯此頁]
Singapore Hotels - Provides you with complete reservation services for hotels and resorts in Singapore. Sorted according to Price, Location, Class, Name.
hotels.online.com.sg/ - 31k - 頁庫存檔 - 類似網頁

2 $x_1^{(2)} = [tfidf, pagerank, ...]$

The Fullerton Hotel Singapore: Weekend Promotion
Get away for the weekend and bask in the luxury of The Fullerton Hotel Singapore. Relax in your elegant guest room or by the outdoor infinity pool, ...
www.fullertonhotel.com/en/promotions/WeekendSpecial.html - 18k - 頁庫存檔 - 類似網頁

1 $x_2^{(2)} = [tfidf, pagerank, ...]$

Ranking as Supervised Learning Problem

Query: SIGIR

3 $x_1^{(1)} = [tfidf, pagerank, ...]$

1 $x_2^{(1)} = [tfidf, pagerank, ...]$

2 $x_3^{(1)} = [tfidf, pagerank, ...]$

Train $f(x)$ such that:

$$f(x_1^{(1)}) > f(x_3^{(1)}) > f(x_2^{(1)})$$

$$f(x_1^{(2)}) > f(x_2^{(2)})$$

Test Query: Singapore Airport

Query: Hotels in Singapore

2 $x_1^{(2)} = [tfidf, pagerank, ...]$

1 $x_2^{(2)} = [tfidf, pagerank, ...]$

Welcome to Changi Airport - [翻譯此頁] ?
With more than 300 retail outlets and F&B outlets in Changi Airport, indulge yourself ... 2006 Civil Aviation Authority of Singapore, All rights reserved. ...
www.changiairport.com/changi/en/index.html?__locale=en - 44k - 頁庫存檔 - 類似網頁

Singapore Changi Airport - Wikipedia, the free encyclopedia - [翻譯此頁] ?
Growth in the global aviation transport was felt in Singapore, where Singapore International Airport at Paya Lebar, Singapore's third main civilian airport ...
en.wikipedia.org/wiki/Singapore_Changi_Airport - 292k - 頁庫存檔 - 類似網頁

Up to 70% off Singapore Airport Hotels at Wotif.com - [翻譯此頁] ?
Don't waste money on a taxi – instant confirmation on Sinagpore Airport Hotels from $165/night. Online bookings. Fast & secure site, and backed by a 24/7 ...
www.wotif.com/hotels/singapore-singapore-airport-east-coast-hotels.html - 14k -

Two kinds of Partially-Labeled Data

1. Lack of labels for some documents (depth)

Query1	Query2	Query3
Doc1 Label	Doc1 Label	Doc1 Label
Doc2 Label	Doc2 Label	Doc2 Label
Doc3 ?	Doc3 ?	Doc3 ?

Some references:
Amini+, SIGIR'08
Agarwal, ICML'06
Wang+, MSRA TechRep'05
Zhou+, NIPS'04
He+, ACM Multimedia '04

2. Lack of labels for some queries (breadth)

Query1	Query2	Query3
Doc1 Label	Doc1 Label	Doc1 ?
Doc2 Label	Doc2 Label	Doc2 ?
Doc3 Label	Doc3 Label	Doc3 ?

This paper
Truong+, ICMIST'06

Focus of this work: Transductive Learning

- Unlabeled data = Test data

  → Transductive Learning

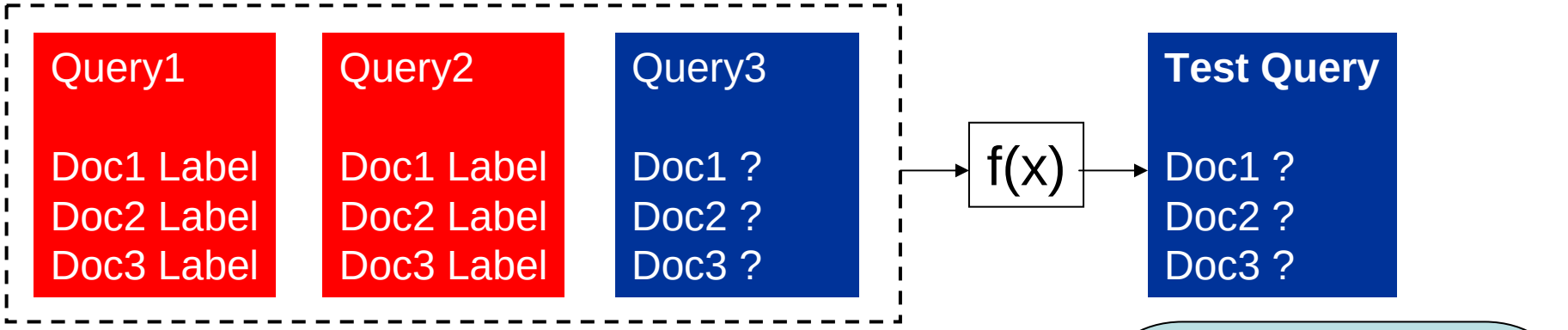
Query1	Query2	**Test Query**
Doc1 Label	Doc1 Label	Doc1 ?
Doc2 Label	Doc2 Label	Doc2 ?
Doc3 Label	Doc3 Label	Doc3 ?

- Main question: How can knowledge of the test list help our learning algorithm?

Why transductive learning?

Inductive (semi-supervised) learning:

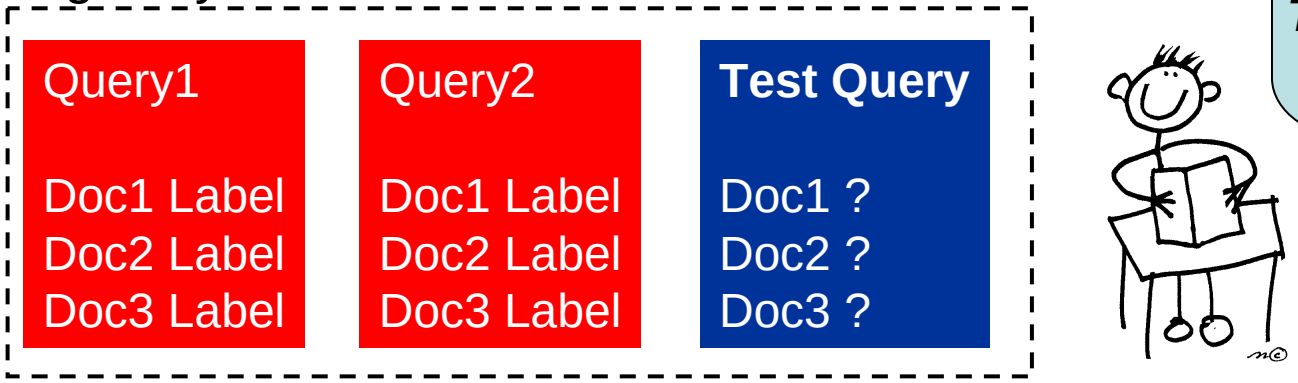
Need to generalize to new data

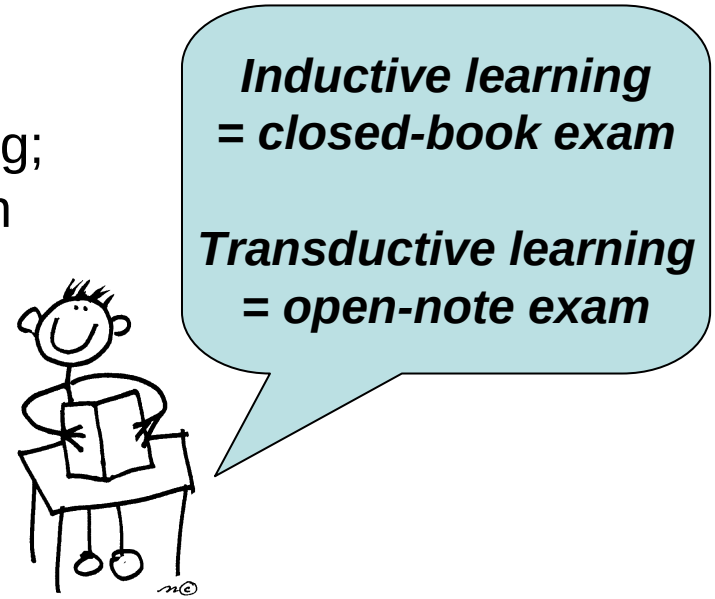


Transductive learning:

Test data is fixed and observed during learning;

Arguably, transduction is easier than induction

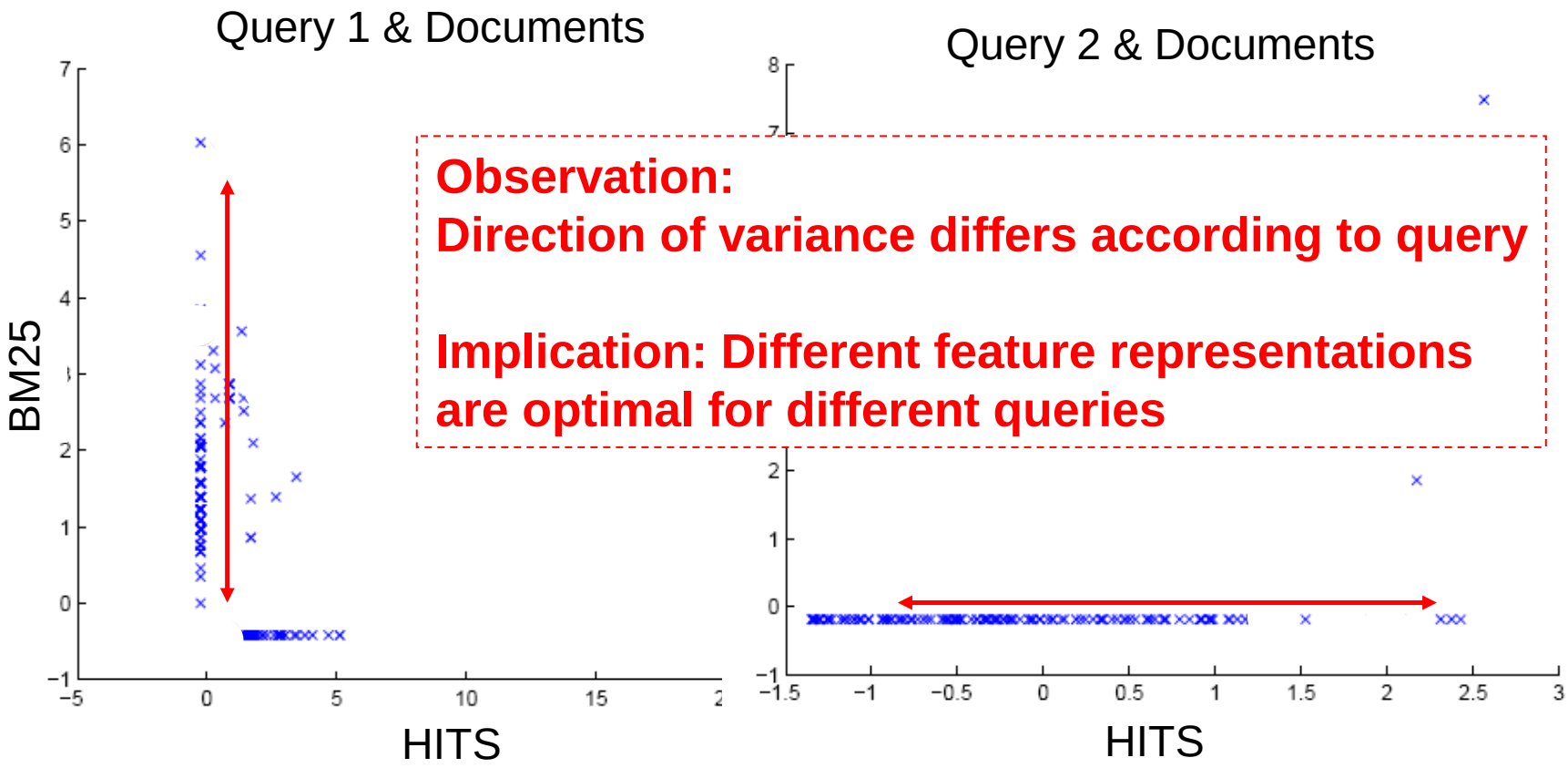




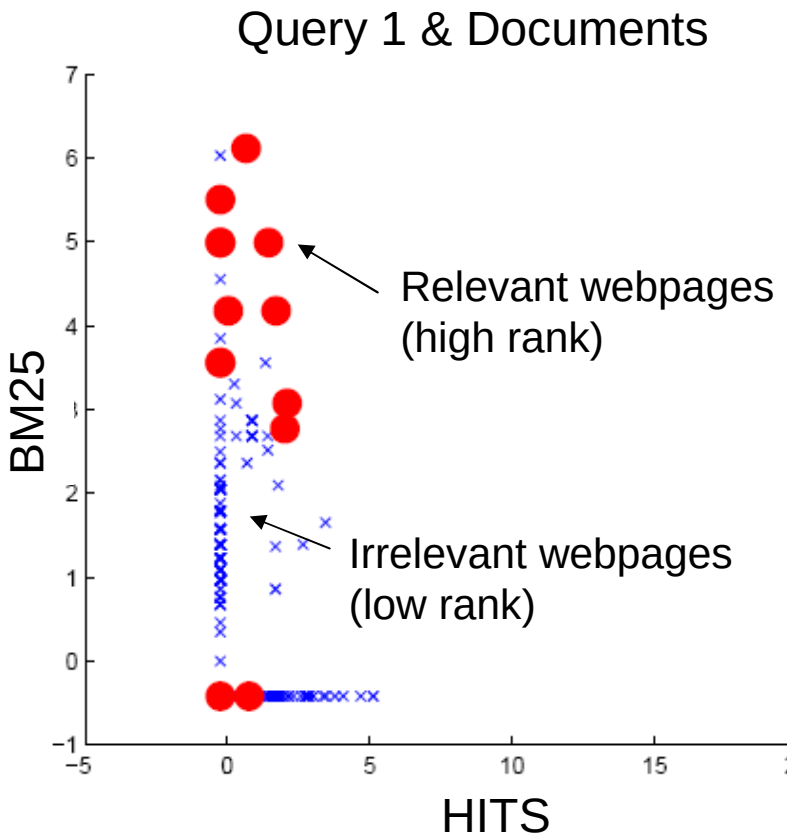
Outline

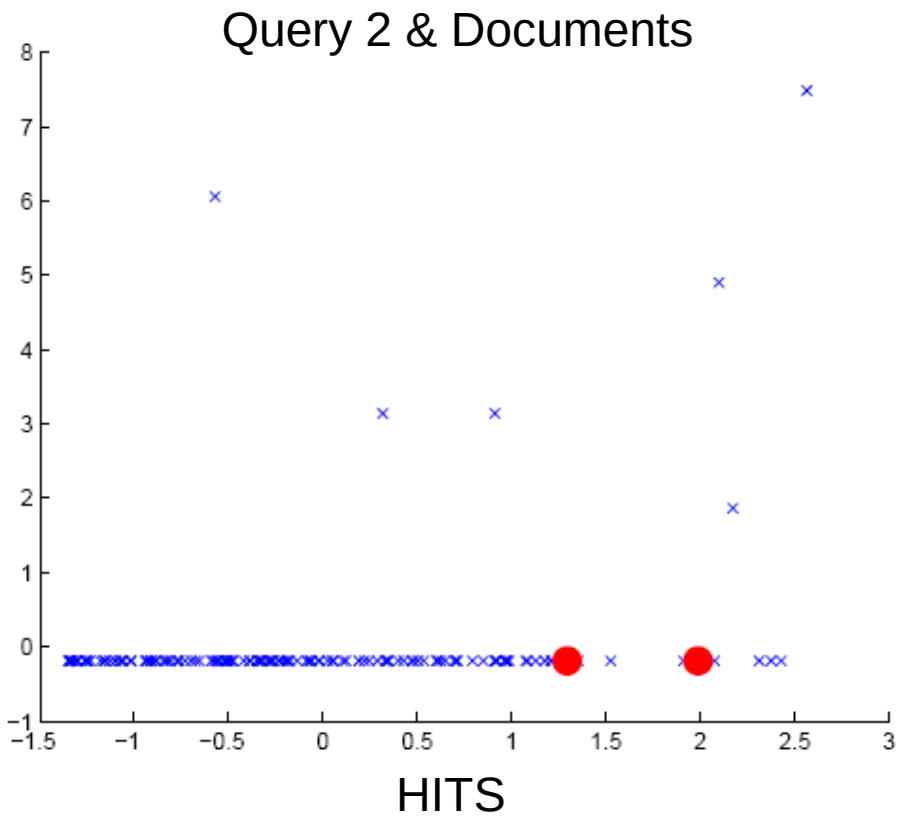
1. Problem Definition
2. Proposed Method
    1. Intuition
    2. Details of proposed algorithm
3. Results and Analysis

Thought Experiment: What information does unlabeled data provide?



Good results can be achieved by:
Ranking Query 1 by BM25 only
Ranking Query 2 by HITS only





Proposed Method: Main Ideas

Main Assumptions:

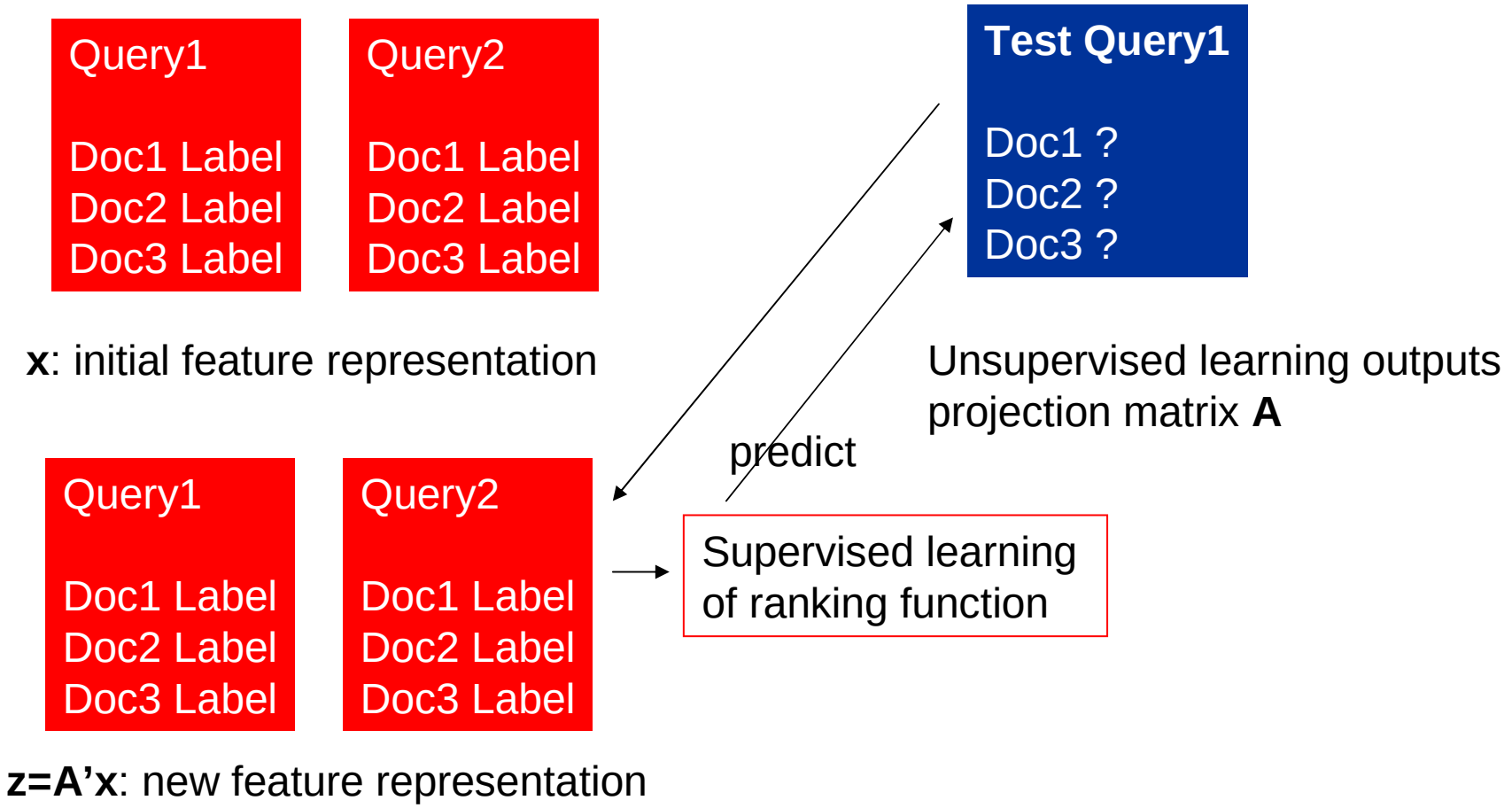
1. Different queries are best modeled by different features
2. Unlabeled data can help us discover this representation

Two-Step Algorithm:

Requires:
- DISCOVER(): unsupervised method for finding useful features
- LEARN(): supervised method for learning to rank

For each Test List:
- Run DISCOVER()
- Augment Feature Representation
- Run LEARN() and Predict

Proposed Method: Illustration

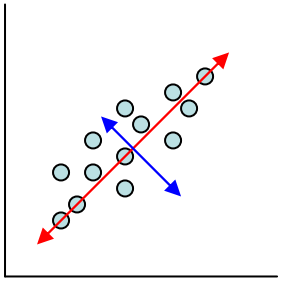


DISCOVER( ) Component

- Goal of DISCOVER( ):

  Find useful patterns on the test list

- Principal Components Analysis (PCA)
  - Discovers direction of maximum variance
  - View low variance directions as noise



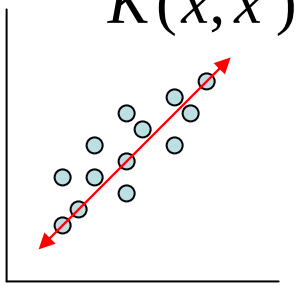
- Kernel PCA *[Scholkopf+, Neural Computation 98]*
  - Non-linear extension to PCA via the Kernel Trick

  1. Maps inputs non-linearly to high-dimensional space.
  2. Performs PCA in that space

Kernels for Kernel PCA

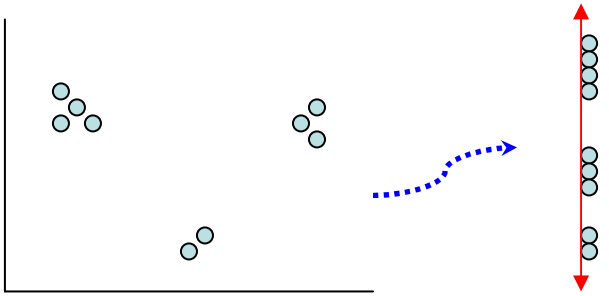
Linear

$$K(x, x') = < x, x' >$$



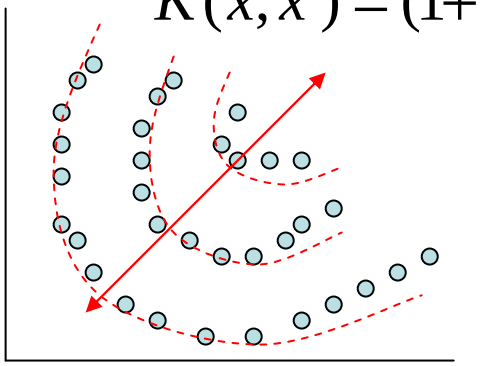
Gaussian

$$K(x, x') = \exp(-\beta \| x - x' \|)$$



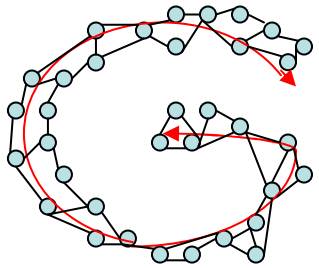
Polynomial

$$K(x, x') = (1 + < x, x' >)^d$$



Diffusion

$K(x, x') =$ Random walk between x, x' on graph



LEARN( ) Component

- Goal of LEARN( ):
  - Optimize some ranking metric on labeled data
- RankBoost *[Freund+, JMLR 2003]*
  - Inherent Feature Selection
  - Few parameters to tune
- Other supervised ranking methods are possible:
  - RankNet, Rank SVM, ListNet, FRank, SoftRank, etc.

Summary of Proposed Method

- Relies on unlabeled test data to learn good feature representation
- "Adapts" the supervised learning process to each test list

- Caveats:
  - DISCOVER() may not always find features that are helpful for LEARN()
  - Run LEARN() at query time → Computational speedup is needed in practical application

Outline

1. Problem Definition
2. Proposed Method
3. Results and Analysis
    1. Experimental Setup
    2. Main Results
    3. Deeper analysis into where things worked and failed

Experiment Setup (1/2)

- LETOR Dataset *[Liu+, LR4IR 2007]*:

	TREC03	TREC04	OHSUMED
# of queries	50	75	106
Average # of documents/query	1000	1000	150
# of original features	44	44	25

- Additional features generated by Kernel PCA:
  - 5 kernels: Linear, Polynomial, Gaussian, Diffusion 1, Diffusion 2
  - Extract 5 principal components for each

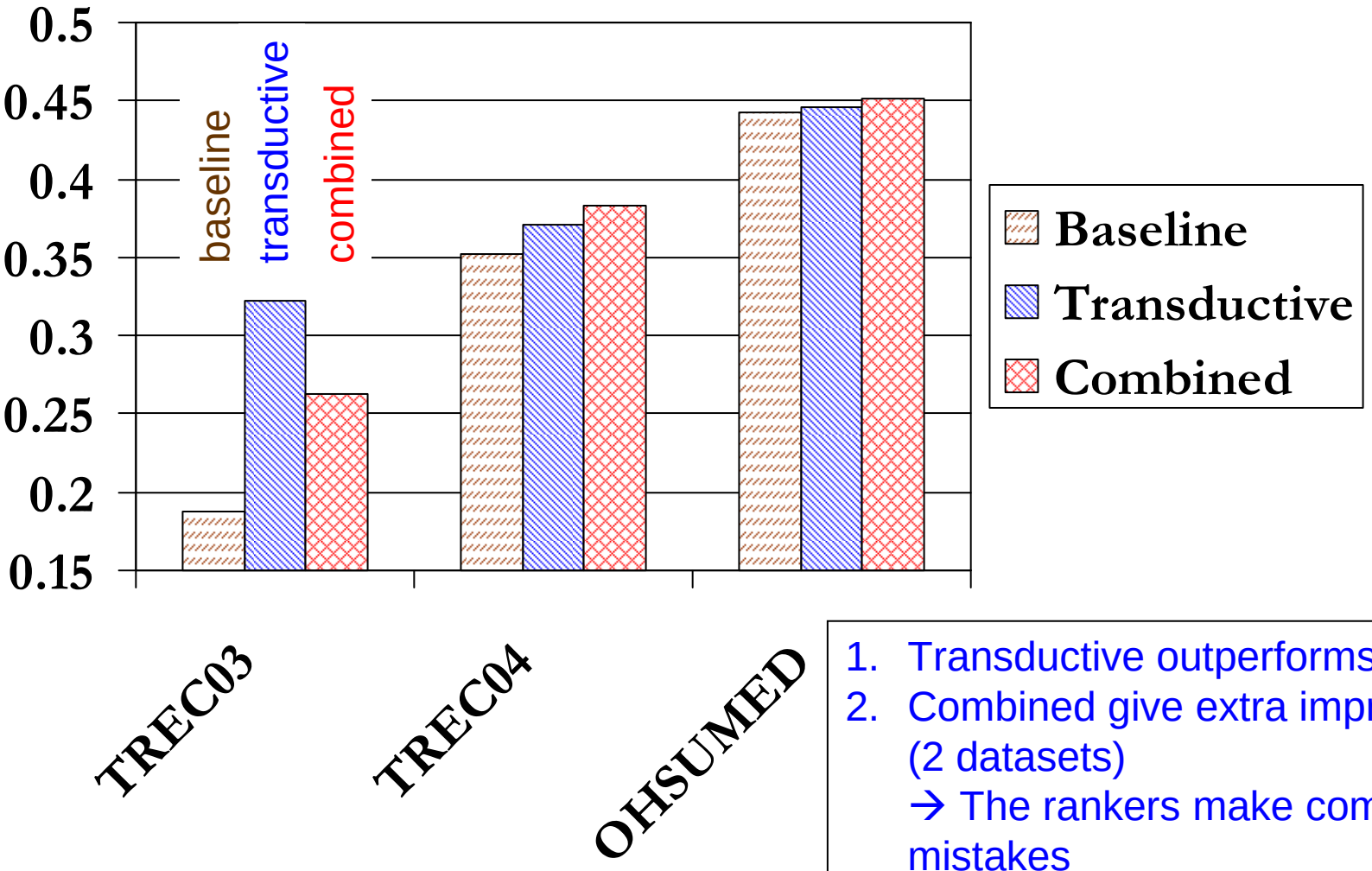
Experiment Setup (2/2)

- Comparison of 3 systems:
  - Baseline: Supervised RankBoost
  - Transductive: Proposed method:
    Kernel PCA + Supervised RankBoost
  - Combined: Average of Baseline, Transductive outputs

$$f(x^{(i)}) = sort\{ f_{baseline}(x_n^{(i)}) + f_{transductive}(x_n^{(i)})\}$$

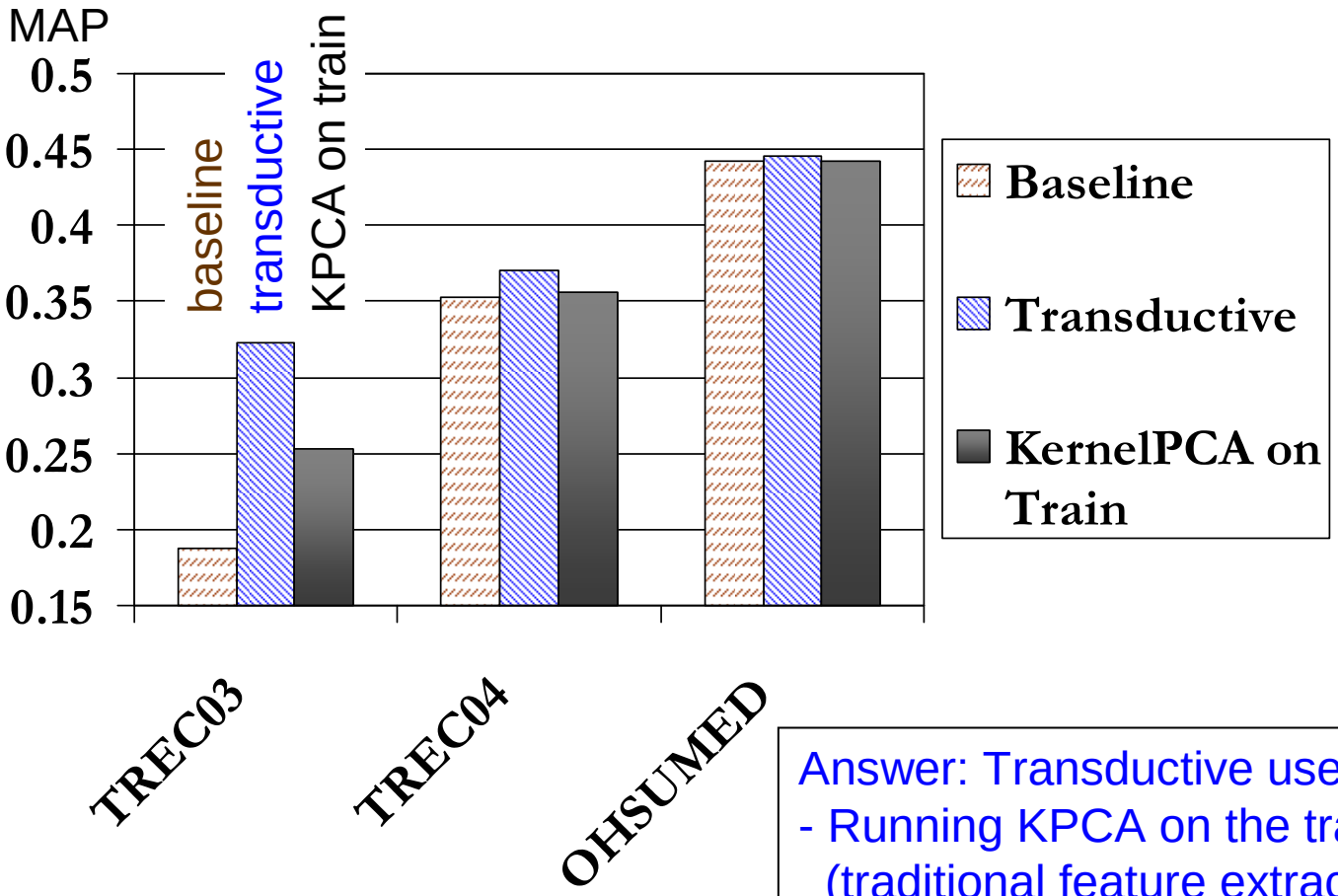
- Evaluation:
  - Mean Averaged Precision (MAP)
  - Normalized Discount Cumulative Gain (NDCG) ← see the paper

Overall Results (MAP)



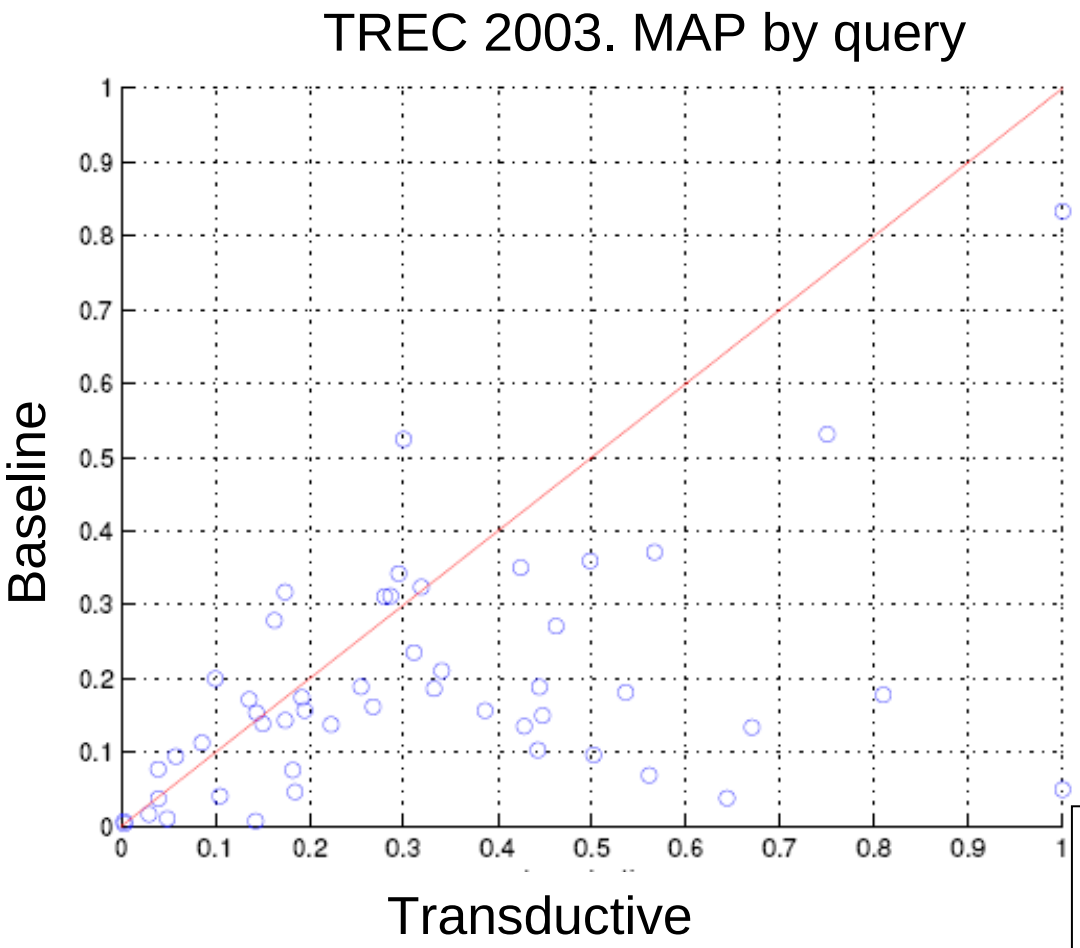
1. Transductive outperforms Baseline
2. Combined give extra improvements (2 datasets)
   → The rankers make complementary mistakes

Did improvements come from Kernel PCA per se, or its transductive use?



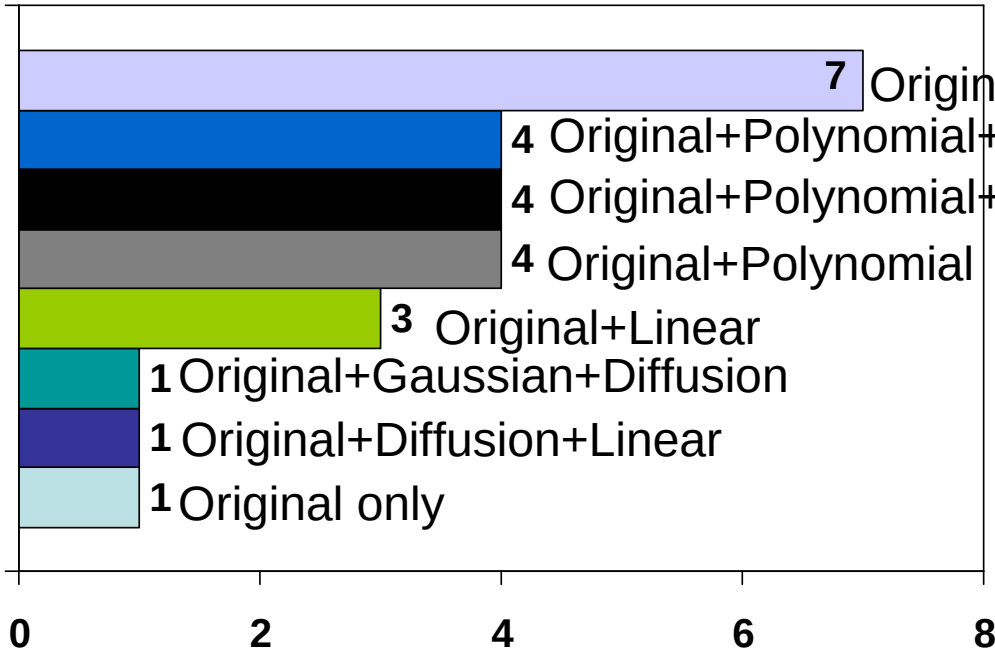
Answer: Transductive use
- Running KPCA on the training set (traditional feature extraction) gives little gains
- Gains are a result of test-specific rankers

Do results vary by query?



Answer:
- Yes. For some queries, it is better not to use the transductive method

What kernels are most useful?



1. Pick top 25 rankers where MAP improved by over 20% (TREC04)
2. Plot histogram of the most important five features

Answer: There is a diversity of kernels that lead to good performance.
Different test list have different structure

Conclusion

- Unlabeled data can be useful for ranking problems
- Two-step transductive algorithm:
  - Adapts the supervised component using a feature representation that better models the test list
- Overall results are positive
  - but results vary at the query-level
- Future work:
  - Computational speed-up
  - Different LEARN() and DISCOVER() components
  - Other ways to exploit unlabeled data

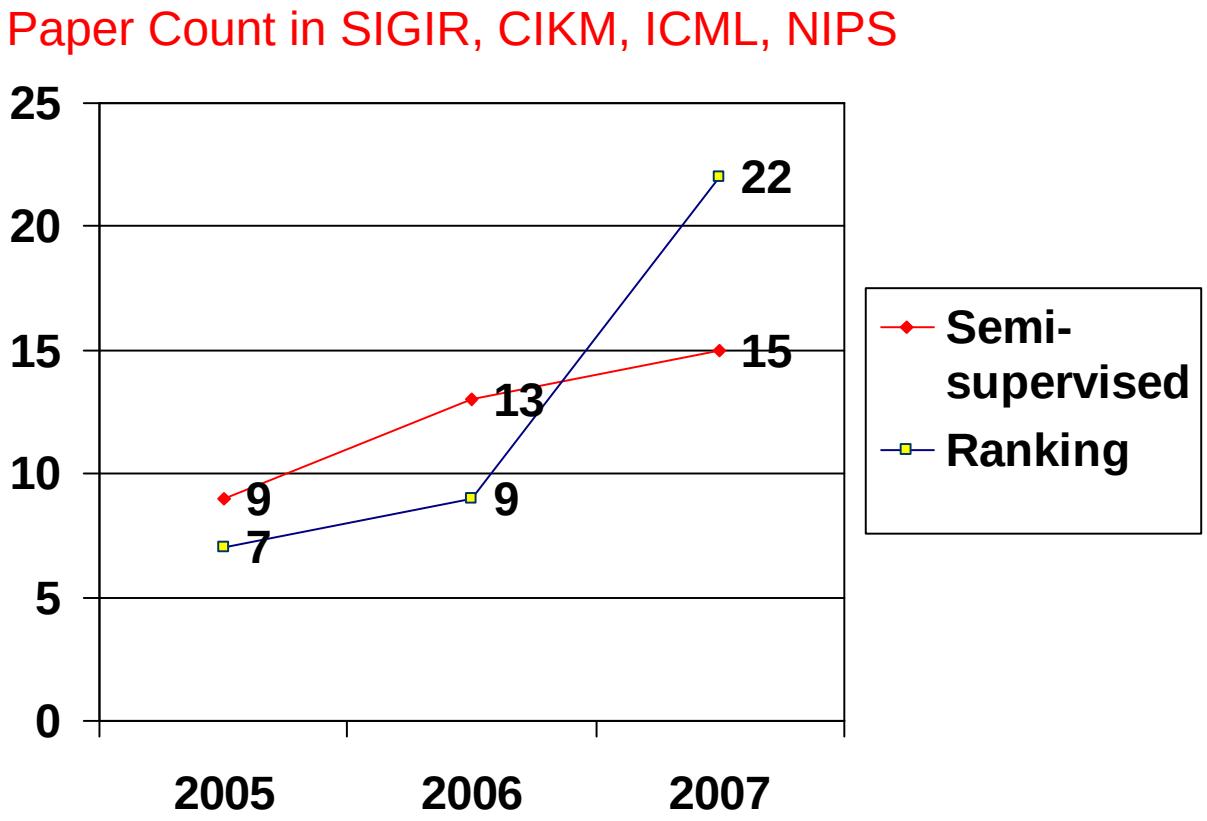
Thanks for your attention!

Acknowledgments:

- U.S. National Science Foundation Graduate Fellowship
- Travel Grant supported by:
  - SIGIR
  - Dr. Amit Singhal (made in honor of Donald B. Crouch)
  - Microsoft Research (in honor of Karen Spark Jones)

The time is ripe for Semi-supervised Ranking!

- Both Semi-supervised Classification and Learning to Rank have become well-established sub-fields with many techniques



Computation Time (OHSUMED)

- On Intel x86-32 (3GHz CPU)
  - Kernel PCA (Matlab/C-Mex): 4.3sec/query
  - Rankboost (C++): 0.7sec/iteration
  - Total time (Assuming 150 iterations): 109sec/query

    (233sec/query for TREC)

- Kernel PCA: O(n^3) for n documents
  - Sparse KPCA: O(n)