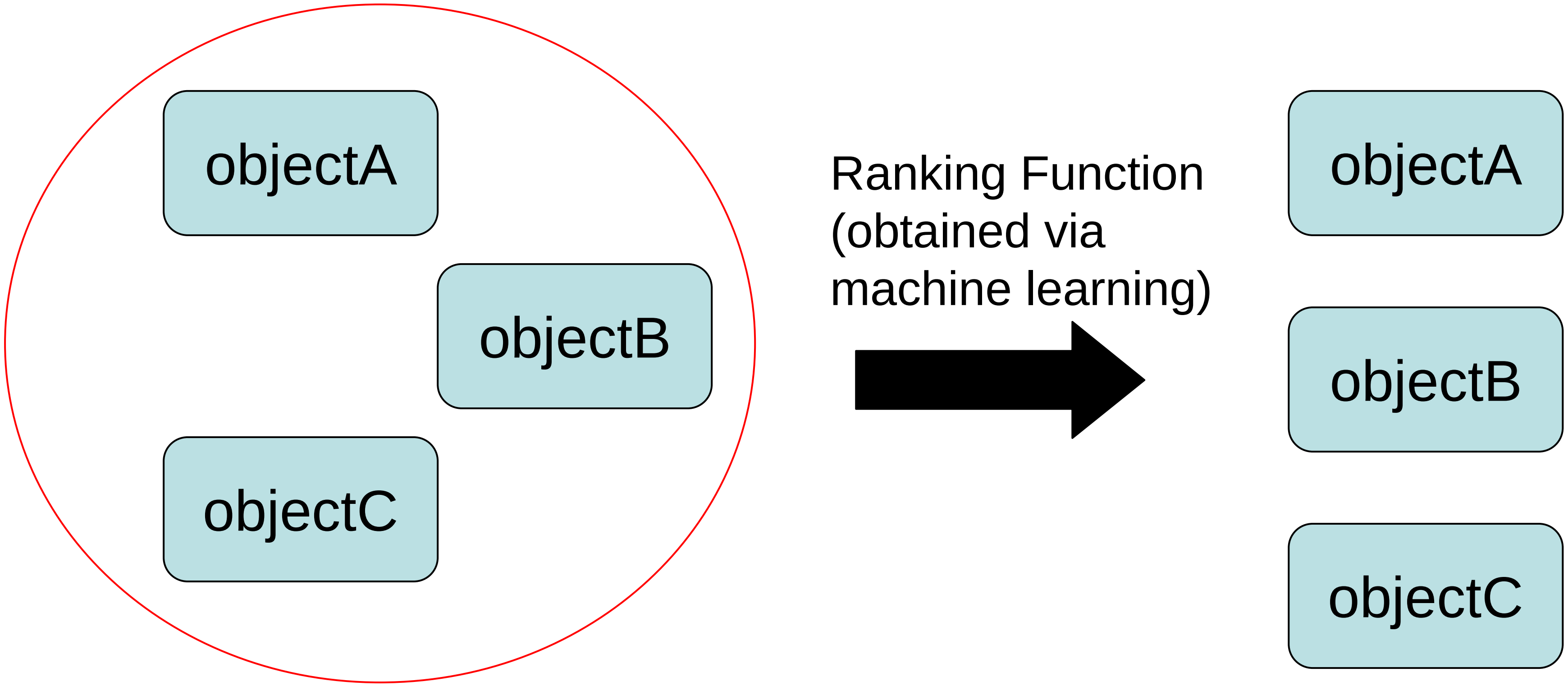
Learning to Rank with Partially-Labeled Data

Kevin Duh

University of Washington

The Ranking Problem

- Definition: Given a set of objects, sort them by preference.



Application: Web Search

You enter “uw” into the searchbox…

All webpages containing the term “uw”:

University of Wyoming - New Thinking
Official web site of the **University of Wyoming**, located in Laramie, Wyoming. Colleges, libraries, directories, faculty, student information and news.
www.**uw**yo.edu/ - 15k - Cached - Similar pages - Note this

UW Athletics - Official Site
Badgers news, team links, tickets, and facilities information.
www.**uw**badgers.com/ - 14k - Cached - Similar pages - Note this

University of Wisconsin-Madison
Skip to menu for main topics about the **University of Wisconsin**; Skip to search; Skip to news ... 2008 Board of Regents of the **University of Wisconsin** System.
www.wisc.edu/ - 14k - Cached - Similar pages - Note this

refresh.**uw**.hu ::
refresh.**uw**.hu - Gitáros Fórum. Gy.IK Gy.IK Keresés Keresés Taglista Taglista Csoportok Csoportok Regisztráció Regisztráció Profil ...
refresh.**uw**.hu/viewtopic.php?p=120127 - 10k - Cached - Similar pages - Note this

University of Washington
Offers information and news for prospective and current students, faculty, and staff. Highlights academic departments and athletics, serves as directory for ...
www.washington.edu/ - 15k - Cached - Similar pages - Note this

Results presented to user, after ranking:

1st **University of Washington**
Offers information and news for prospective and current students, faculty, and staff. Highlights academic departments and athletics, serves as directory for ...
www.washington.edu/ - 15k - Cached - Similar pages - Note this

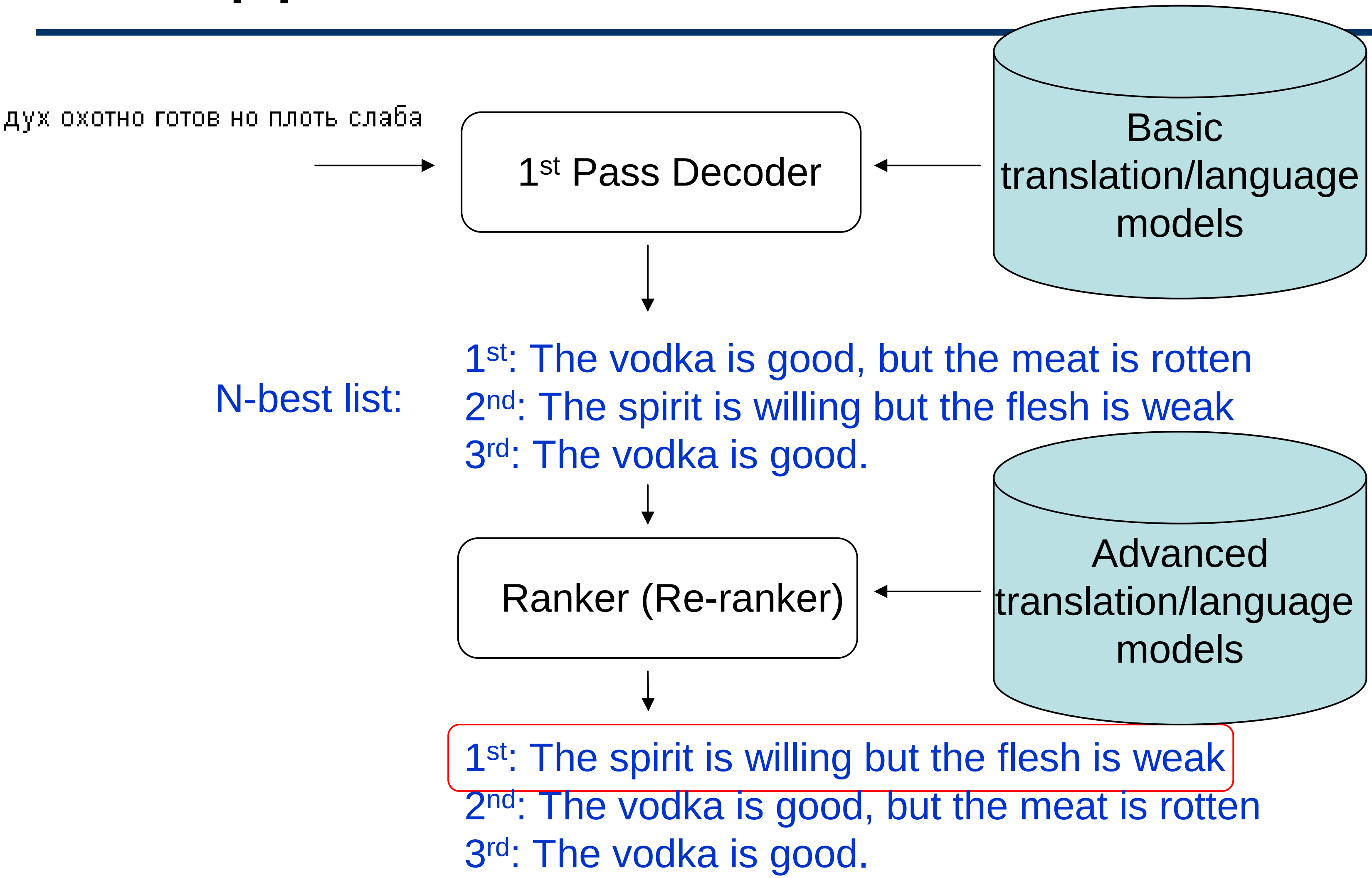
2nd **University of Wisconsin**-Madison
Skip to menu for main topics about the **University of Wisconsin**; Skip to search; Skip to news ... 2008 Board of Regents of the **University of Wisconsin** System.
www.wisc.edu/ - 14k - Cached - Similar pages - Note this

3rd **University of Wyoming** - New Thinking
Official web site of the **University of Wyoming**, located in Laramie, Wyoming. Colleges, libraries, directories, faculty, student information and news.
www.**uw**yo.edu/ - 15k - Cached - Similar pages - Note this

4th **UW** Athletics - Official Site
Badgers news, team links, tickets, and facilities information.
www.**uw**badgers.com/ - 14k - Cached - Similar pages - Note this

5th refresh.**uw**.hu ::
refresh.**uw**.hu - Gitáros Fórum. Gy.IK Gy.IK Keresés Keresés Taglista Taglista Csoportok Csoportok Regisztráció Regisztráció Profil ...
refresh.**uw**.hu/viewtopic.php?p=120127 - 10k - Cached - Similar pages - Note this

Application: Machine Translation

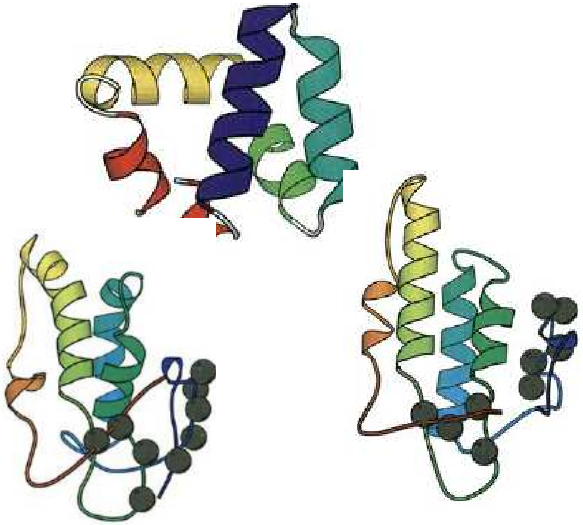


UNIVERSITY OF WASHINGTON

Application: Protein Structure Prediction

Amino Acid Sequence:
MMKLKSNQTRTYDGDGYKKRAACLCFSE

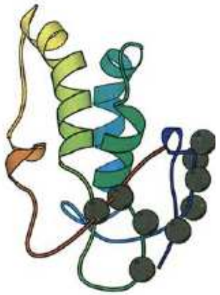
various protein folding simulations



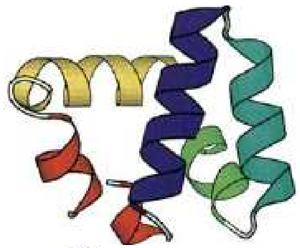
Candidate 3-D Structures

Ranker

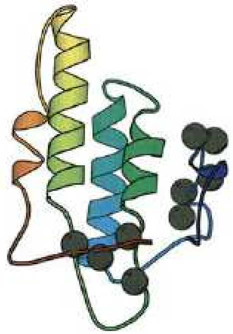
1st



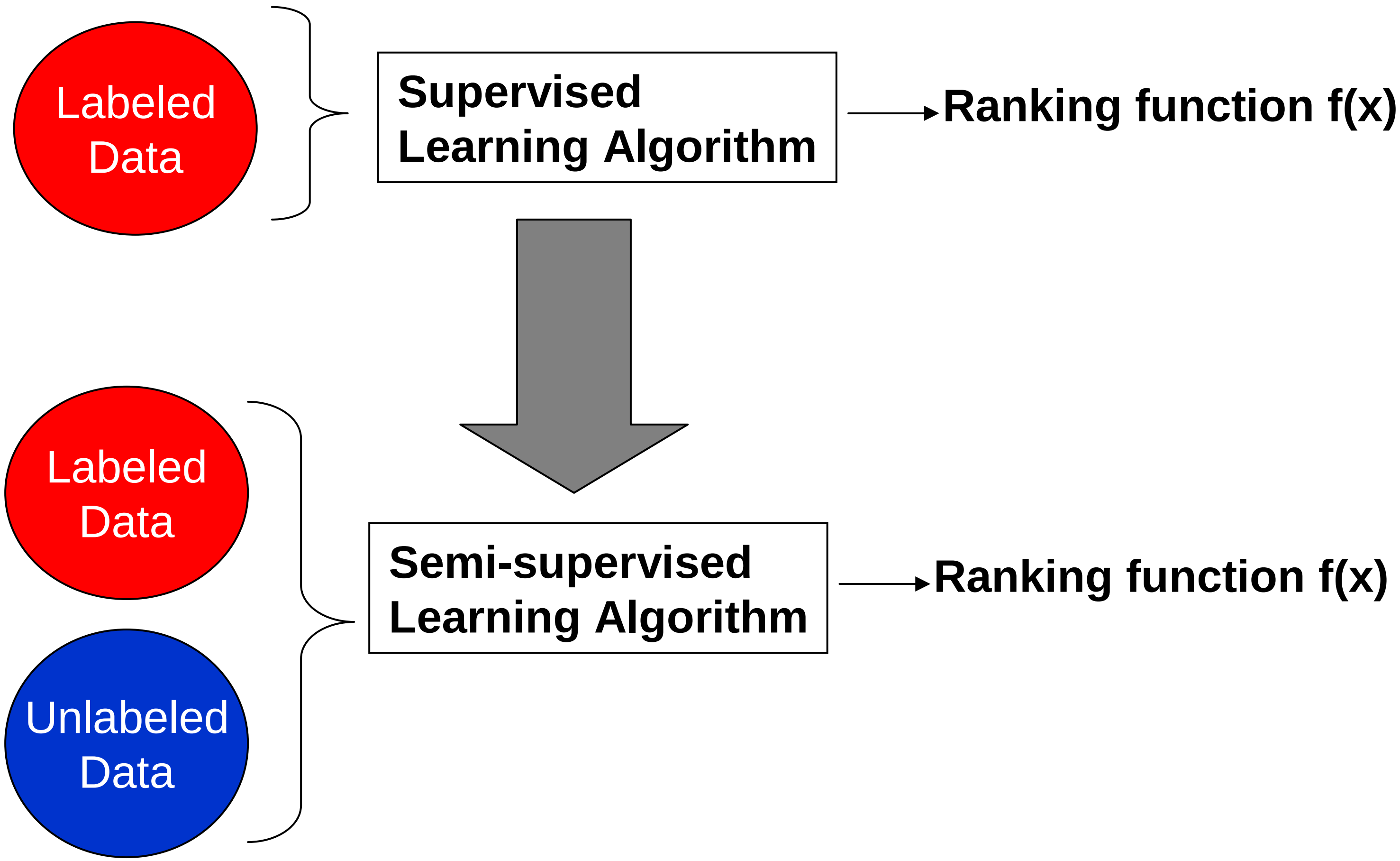
2nd



3rd



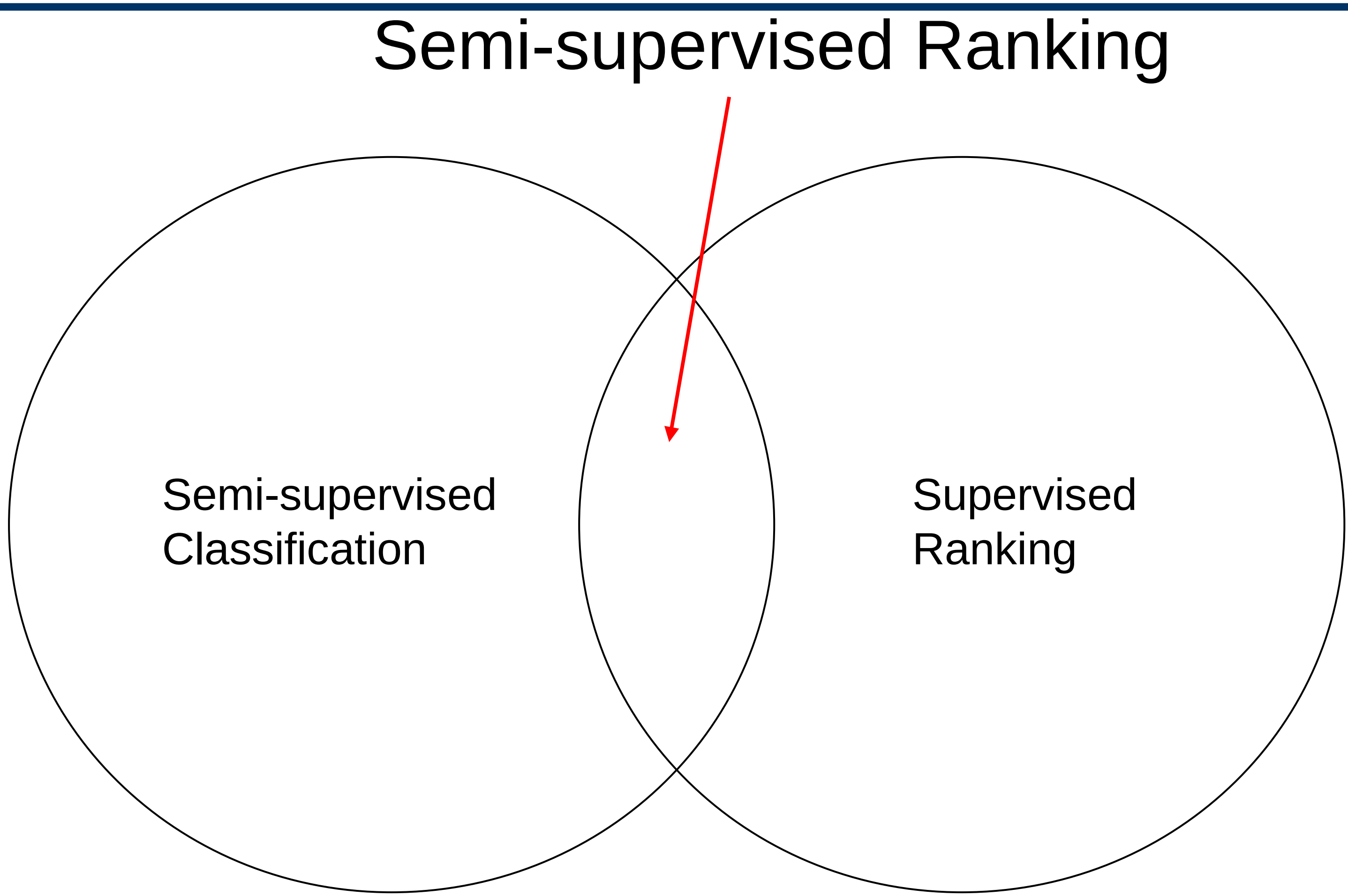
Goal of this thesis



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

UNIVERSITY OF WASHINGTON

Emerging field



Outline

1. Problem Setup
   1. Background in Ranking
   2. Two types of partially-labeled data
   3. Methodology
2. Manifold Assumption
3. Local/Transductive Meta-Algorithm
4. Summary

Ranking as Supervised Learning Problem

Query: UW

Labels

[University of Washington](#)
Offers information and news for prospective and current students, faculty, and staff. Highlights academic departments and athletics, serves as directory for ...
www.washington.edu/ - 15k - Cached - Similar pages - Note this

3 $x_1^{(i)} = [tfidf, pagerank, \ldots]$

[University of Wyoming - New Thinking](#)
Official web site of the **University of Wyoming**, located in Laramie, Wyoming. Colleges, libraries, directories, faculty, student information and news.
www.**uwyo**.edu/ - 15k - Cached - Similar pages - Note this

1 $x_2^{(i)} = [tfidf, pagerank, \ldots]$

[University of Wisconsin-Madison](#)
Skip to menu for main topics about the **University of Wisconsin**; Skip to search; Skip to news ... 2008 Board of Regents of the **University of Wisconsin** System.
www.wisc.edu/ - 14k - Cached - Similar pages - Note this

2 $x_3^{(i)} = [tfidf, pagerank, \ldots]$

Query: Seattle Traffic

[WSDOT Seattle Area Traffic - Traffic Conditions and Travel Alerts](#)
A map of current freeway **traffic** conditions for **Seattle** and surrounding areas; includes links to **traffic** cams, incident reports, mountain pass reports, ...
www.wsdot.wa.gov/**Traffic/seattle**/ - 38k - Cached - Similar pages - Note this

2 $x_1^{(j)} = [tfidf, pagerank, \ldots]$

[Seattle Praised for Traffic Efficiency : NPR](#)
Seattle and Tacoma's program to ease **traffic** flows is cited as the nation's most effective by the Texas Transportation Institute.
www.npr.org/templates/story/story.php?storyId=3905008 - Similar pages - Note this

1 $x_2^{(j)} = [tfidf, pagerank, \ldots]$

Ranking as Supervised Learning Problem

Query: UW

3 $x_1^{(i)} = [tfidf, pagerank, \ldots]$

1 $x_2^{(i)} = [tfidf, pagerank, \ldots]$

2 $x_3^{(i)} = [tfidf, pagerank, \ldots]$

Query: Seattle Traffic

2 $x_1^{(j)} = [tfidf, pagerank, \ldots]$

1 $x_2^{(j)} = [tfidf, pagerank, \ldots]$

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: MSR

MSR | MX Gear | One Brand Fits All...
Motocross gear, off road gear, and hard parts.
www.**msr**acing.com/ - 2k - Cached - Similar pages - Note this **?**

Microsoft Research Home
Corporate research division. Includes projects and publications, news and history, and job opportunities.
research.microsoft.com/ - 25k - Cached - Similar pages - Note this **?**

MSR Mountain Safety Research
This is the home page for **Mountain Safety Research** ®, manufacturers of the most reliable and functional backcountry gear in the world.
www.**msr**gear.com/ - 11k - Cached - Similar pages - Note this **?**

Semi-supervised Data: Some labels are missing

Labels

Query: UW

[University of Washington](#)
Offers information and news for prospective and current students, faculty, and staff. Highlights academic departments and athletics, serves as directory for ...
www.washington.edu/ - 15k - Cached - Similar pages - Note this

3 $x_1^{(i)} = [tfidf, pagerank, \ldots]$

University of Wyoming - New Thinking
Official web site of the **University of Wyoming**, located in Laramie, Wyoming. Colleges, libraries, directories, faculty, student information and news.
www.**uwyo**.edu/ - 15k - Cached - Similar pages - Note this

1 $x_2^{(i)} = [tfidf, pagerank, \ldots]$

University of Wisconsin-Madison
Skip to menu for main topics about the **University of Wisconsin**; Skip to search; Skip to news ... 2008 Board of Regents of the **University of Wisconsin** System.
www.wisc.edu/ - 14k - Cached - Similar pages - Note this

~~2~~ X $x_3^{(i)} = [tfidf, pagerank, \ldots]$

Query: Seattle Traffic

WSDOT **Seattle** Area **Traffic** - **Traffic** Conditions and Travel Alerts
A map of current freeway **traffic** conditions for **Seattle** and surrounding areas; includes links to **traffic** cams, incident reports, mountain pass reports, ...
www.wsdot.wa.gov/**Traffic/seattle**/ - 38k - Cached - Similar pages - Note this

~~2~~ X $x_1^{(j)} = [tfidf, pagerank, \ldots]$

Seattle Praised for **Traffic** Efficiency : NPR
Seattle and Tacoma's program to ease **traffic** flows is cited as the nation's most effective by the Texas Transportation Institute.
www.npr.org/templates/story/story.php?storyId=3905008 - Similar pages - Note this

~~1~~ X $x_2^{(j)} = [tfidf, pagerank, \ldots]$

Two kinds of Semi-supervised 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 thesis
Duh&Kirchhoff, SIGIR'08
Truong+, ICMIST'06

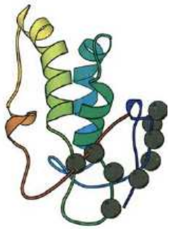
Why "Breadth" Scenario

- Information Retrieval: Long tail of search queries

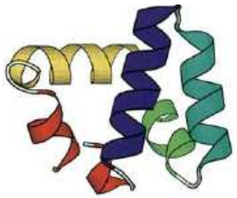
"20-25% of the queries we will see today, we have never seen before"

– Udi Manber (Google VP), May 2007

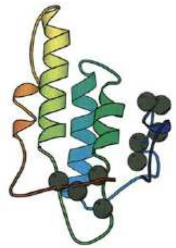
- Machine Translation and Protein Prediction:
  - Given references (costly), computing labels is trivial



reference



candidate 1
similarity=0.3



candidate 2
similarity=0.9

Methodology of this thesis

1. ***Make an assumption*** about how can unlabeled lists be useful
   - Borrow ideas from semi-supervised classification

2. ***Design a method*** to implement it
   - 4 unlabeled data assumptions & 4 methods

3. ***Test on various datasets***
   - Analyze when a method works and doesn't work

Datasets

<u>Information Retrieval datasets</u>
- from LETOR distribution [Liu'07]
- TREC: Web search / OHSUMED: Medical search
- Evaluation: MAP (measures how high relevant documents are on list)

	TREC 2003	TREC 2004	OHSUMED	Arabic translation	Italian translation	Protein prediction
# lists	50	75	100	500	500	100
label type	2 level	2 level	3 levels	conti-nuous	conti-nuous	conti-nuous
avg # objects per list	1000	1000	150	260	360	120
# features	44	44	25	9	10	25

Datasets

<u>Machine Translation datasets</u>
- from IWSLT 2007 competition, UW system [Kirchhoff'07]
- translation in the travel domain
- Evaluation: BLEU (measures word match to reference)

	TREC 2003	TREC 2004	OHSUMED	Arabic translation	Italian translation	Protein prediction
# lists	50	75	100	500	500	100
label type	2 level	2 level	3 levels	conti-nuous	conti-nuous	conti-nuous
avg # objects per list	1000	1000	150	260	360	120
# features	44	44	25	9	10	25

Datasets

Protein Prediction dataset
- from CASP competition [Qiu/Noble'07]
- Evaluation: GDT-TS (measures closeness to true 3-D structure)

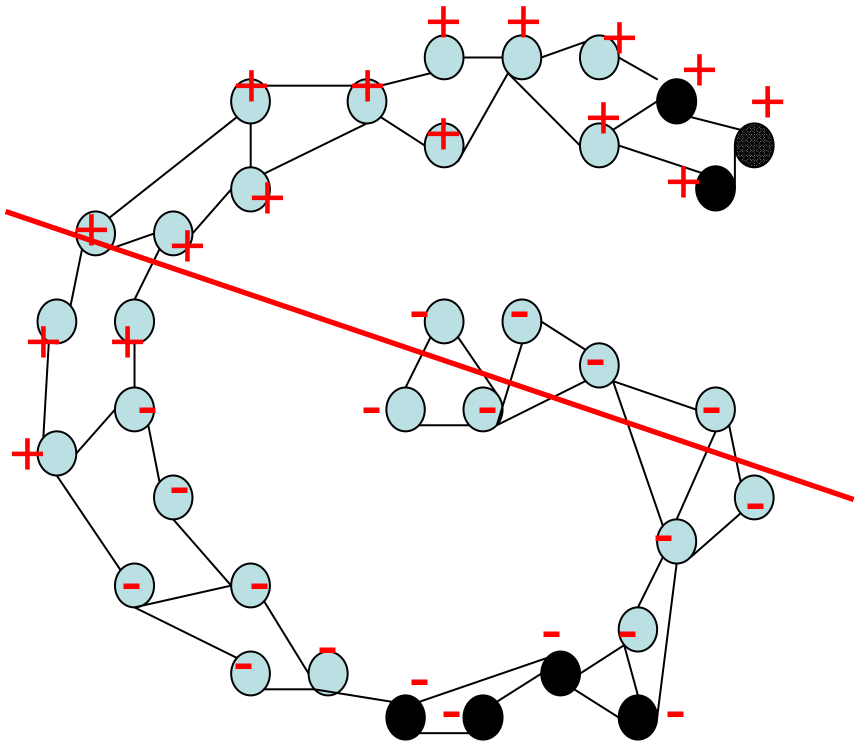
	TREC 2003	TREC 2004	OHSUMED	Arabic translation	Italian translation	Protein prediction
# lists	50	75	100	500	500	100
label type	2 level	2 level	3 levels	conti-nuous	conti-nuous	conti-nuous
avg # objects per list	1000	1000	150	260	360	120
# features	44	44	25	9	10	25

Outline

1. Problem Setup
2. Manifold Assumption
    - Definition
    - Ranker Propagation Method
    - List Kernel similarity
3. Local/Transductive Meta-Algorithm
4. Summary

Manifold Assumption in Classification

-Unlabeled data can help discover underlying data manifold
-Labels vary smoothly over this manifold



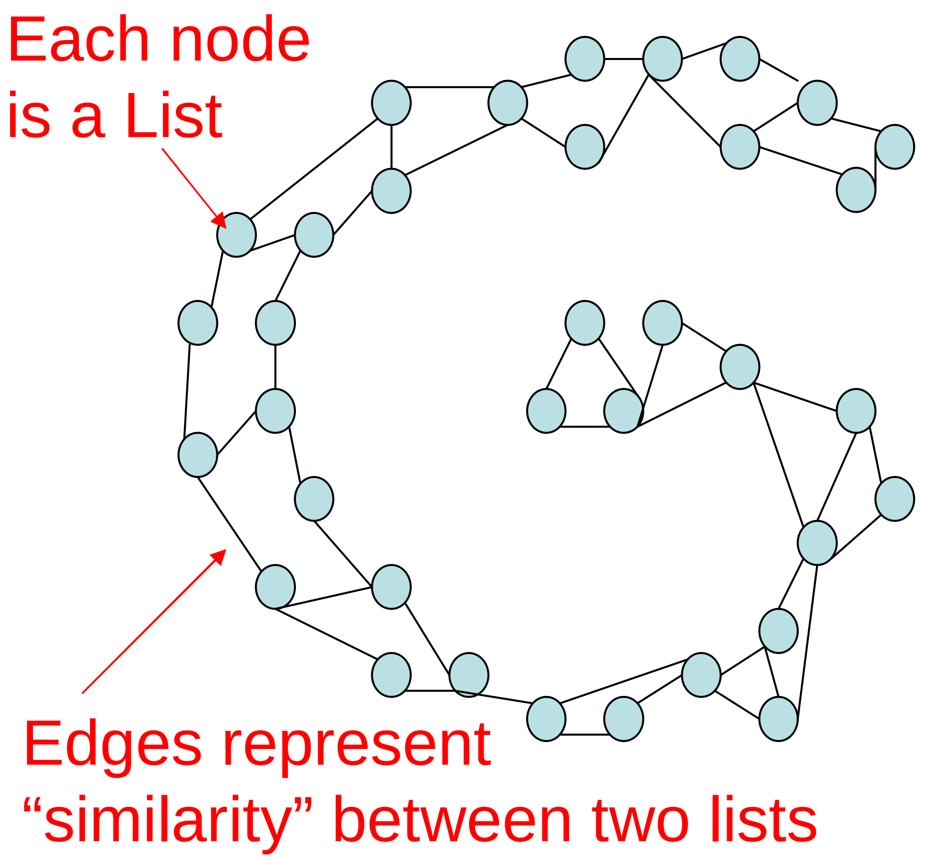
Prior work:

1. How to give labels to test samples?
   - Mincut [Blum01]
   - Label Propagation [Zhu03]
   - Regularizer+Optimization [Belkin03]

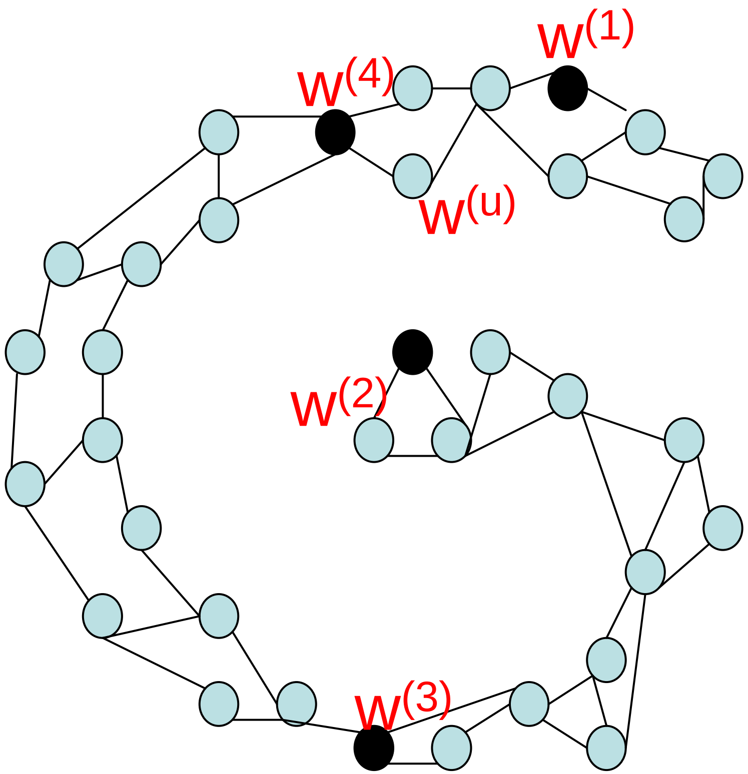
2. How to construct graph?
   - k-nearest neighbors, eps-ball
   - data-driven methods [Argyriou05,Alexandrescu07]

Manifold Assumption in Ranking

Ranking functions vary smoothly over the manifold



Ranker Propagation



Algorithm:

1. For each train list, fit a ranker

$$F(x) = w^T x \qquad w \in R^d, x \in R^d$$

2. Minimize objective:

$$\sum_{ij \in edges} K^{(ij)} \| w^{(i)} - w^{(j)} \|^2$$

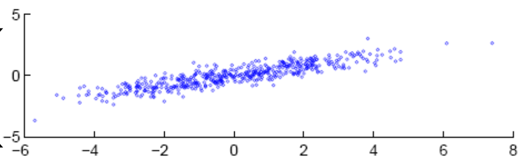
Ranker for list i

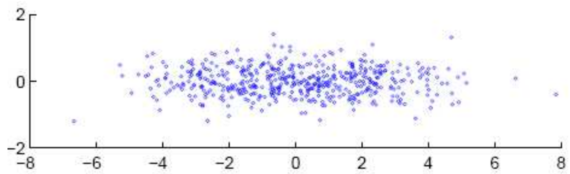
Similarity between list i,j

$$W^{(u)} = -inv(L^{(uu)}) L^{(ul)} W^{(l)}$$

Similarity between lists: Desirable properties

- Maps two lists of feature vectors to scalar

K(

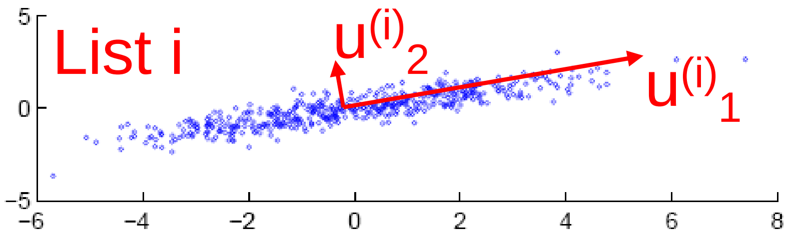
, 

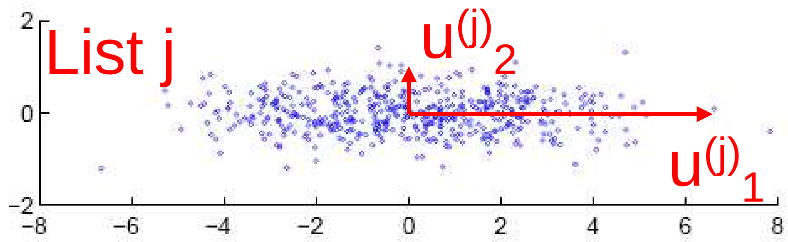
) =0.7

- Work on variable length lists (different N in N-best)
- Satisfy symmetric, positive semi-definite properties
- Measure rotation/shape differences

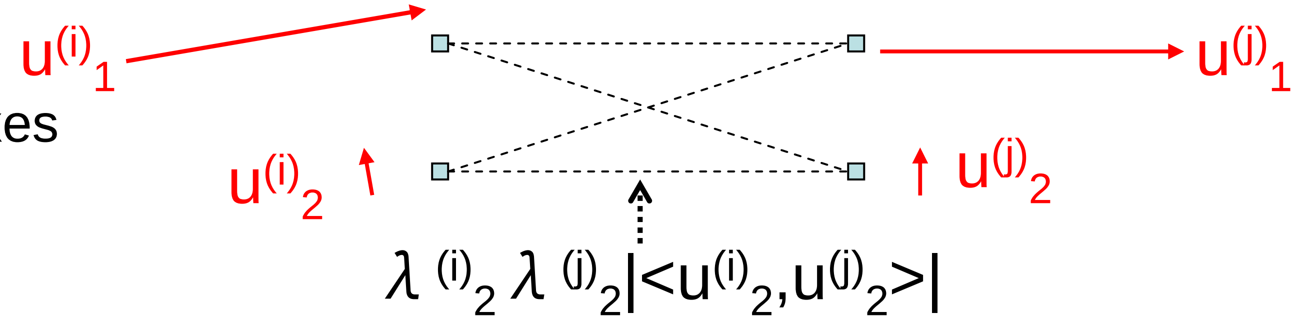
List Kernel

Step 1: PCA





Step 2: Compute similarity between axes

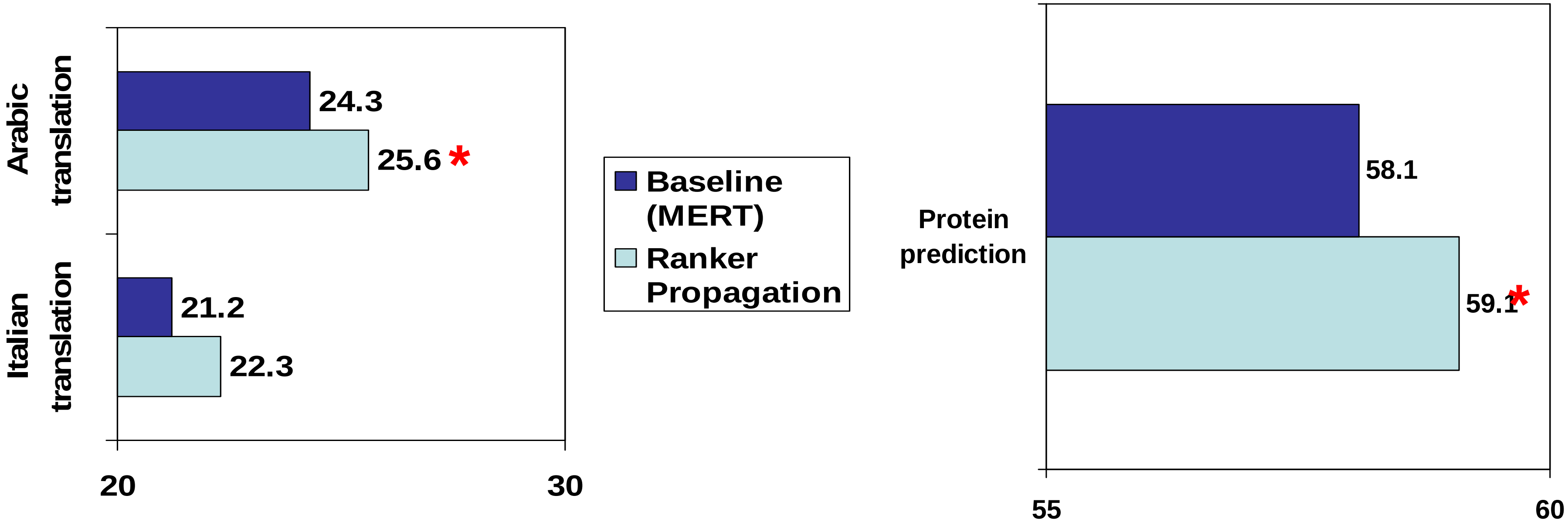


Step 3: Maximum Bipartite Matching

$$K^{(ij)} = \sum_{m=1}^{M} \lambda_m^{(i)} \lambda_{a(m)}^{(j)} \left| \langle u_m^{(i)}, u_{a(m)}^{(j)} \rangle \right| / \| \lambda^{(i)} \| \cdot \| \lambda^{(j)} \|$$

Evaluation in Machine Translation & Protein Prediction

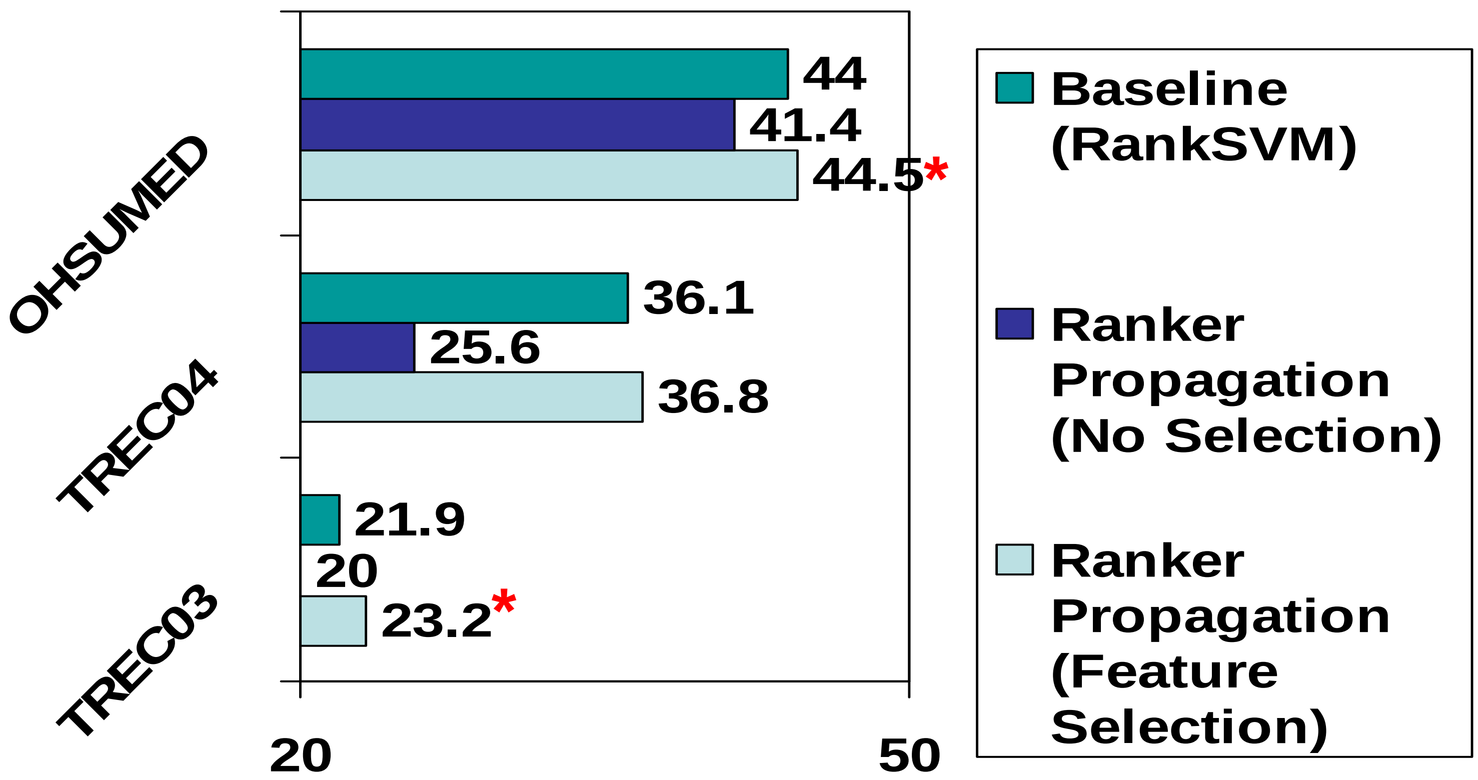
Ranker Propagation (with List Kernel) outperforms Supervised Baseline (MERT linear ranker)



Indicates statistically significant improvement (p<0.05) over baseline

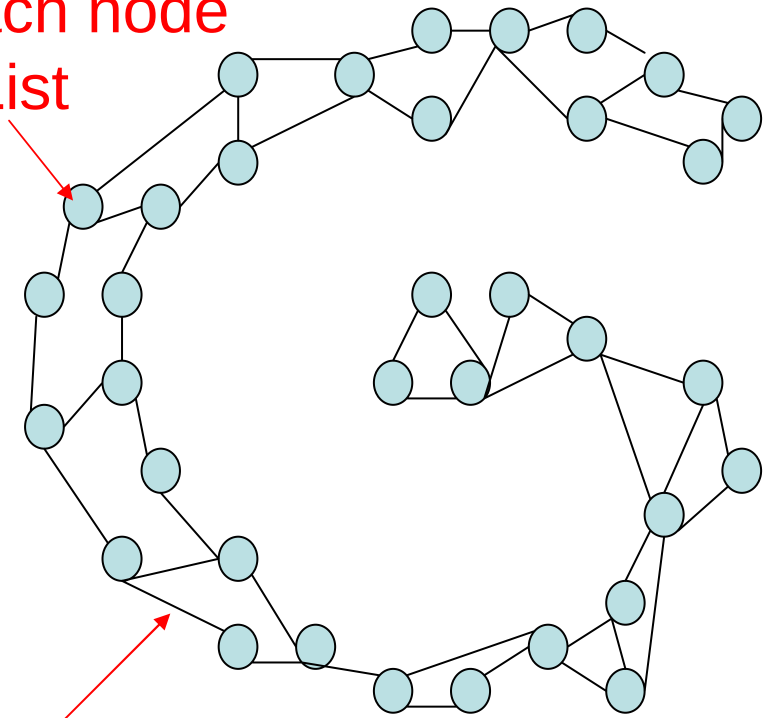
Evaluation in Information Retrieval

1. List Kernel did not give good similarity
2. Feature selection is needed



Summary

1. Each node is a List



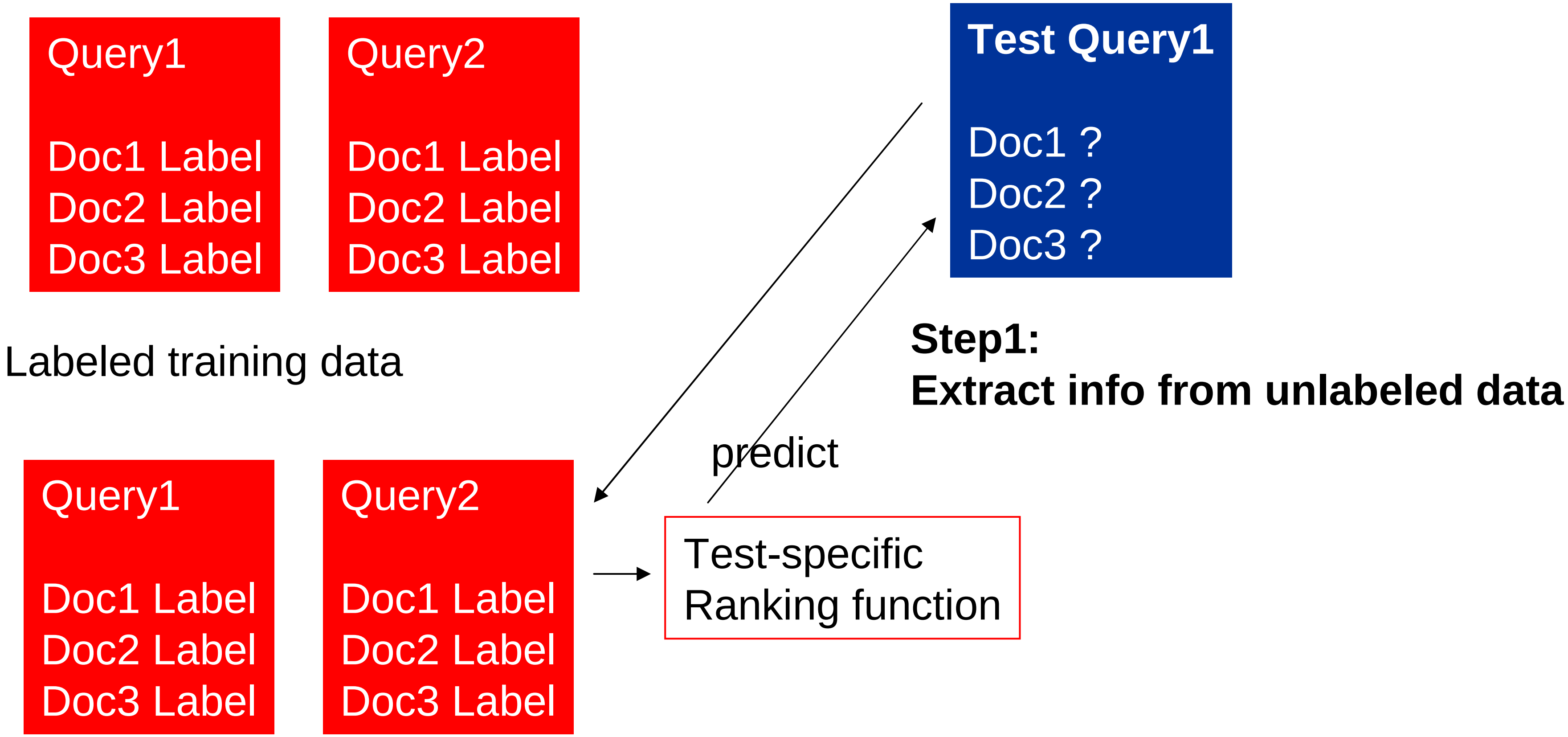
3. Ranker Propagation computes rankers that are smooth over manifold

2. Edge similarity = List Kernel

Outline

1. Problem Setup
2. Manifold Assumption
3. Local/Transductive Meta-Algorithm
   1. Change of Representation Assumption
   2. Covariate Shift Assumption
   3. Low Density Separation Assumption
4. Summary

Local/Transductive Meta-Algorithm



Step2:
Train with extracted unlabel info as bias

Local/Transductive Meta-Algorithm

- Rationale: Focus only on one unlabeled (test) list each time
  - Ensure that the information extracted from unlabeled data is directly applicable

- The name:
  - Local = ranker is targeted at a single test list
  - Transductive = training doesn't start until test data is seen

- Modularity:
  - We will plug-in 3 different unlabeled data assumptions

RankBoost [Freund03]

Query: UW

3 $x_1^{(i)} = [tfidf, pagerank, \ldots]$

2 $x_2^{(i)} = [tfidf, pagerank, \ldots]$

1 $x_3^{(i)} = [tfidf, pagerank, \ldots]$

Objective: maximize pairwise accuracy

$$F(x_1^{(i)}) > F(x_2^{(i)})$$

$$F(x_2^{(i)}) > F(x_3^{(i)})$$

$$F(x_1^{(i)}) > F(x_3^{(i)})$$

Initialize distribution over pairs $D_0(p,q)$ $\forall x_p$ ranked-above x_q

For t=1..T

Train weak ranker h_t to maximize $D_t(p,q) \cdot \mathrm{I}_{\{F(x_p) > F(x_q)\}}$

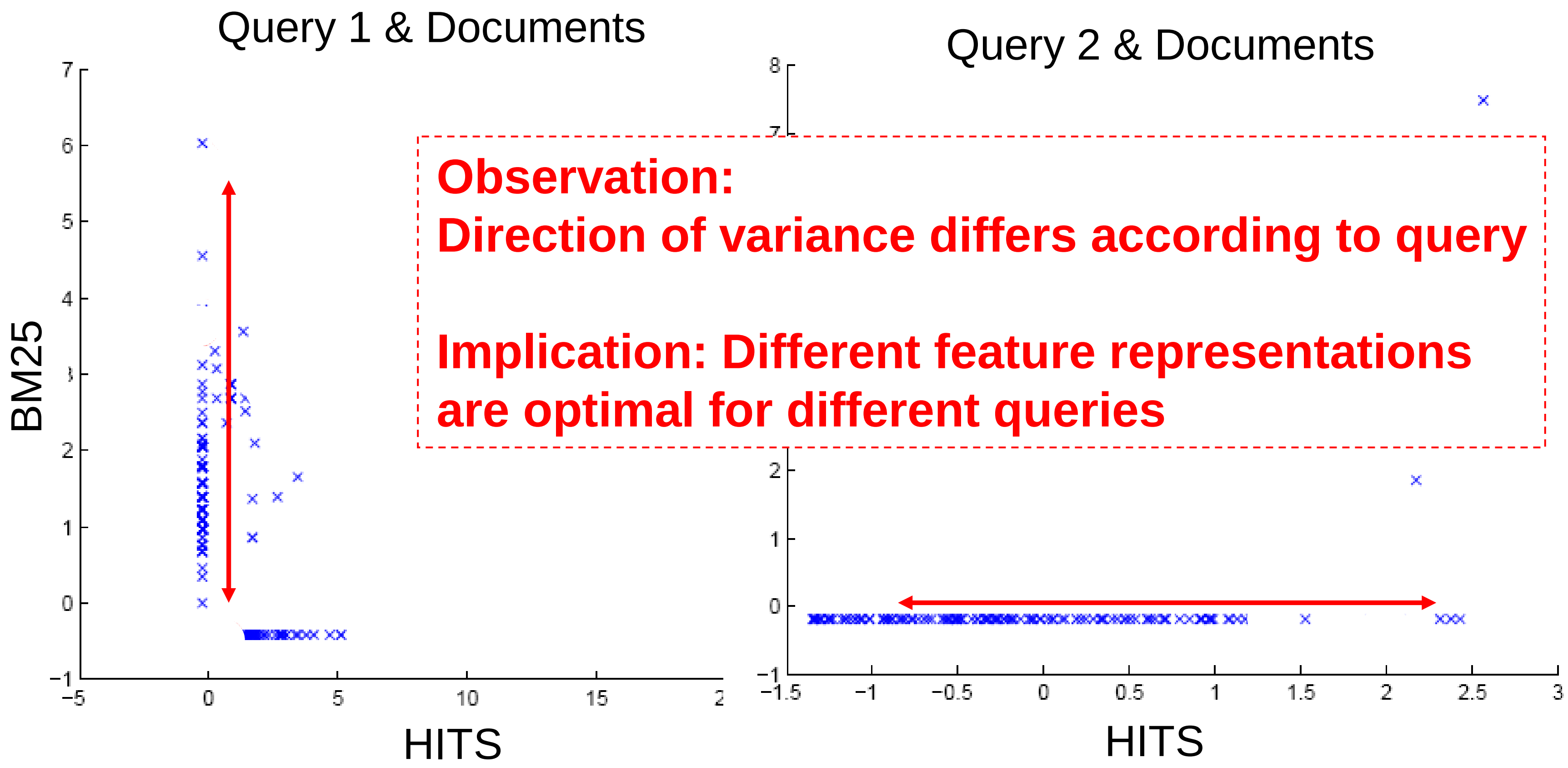
Update distribution $D_{t+1}(p,q) = D_t(p,q) \exp\{\alpha_t (h_t(x_p) - h_t(x_q))\}$

Final ranker

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

Change of Representation Assumption

"Unlabeled data can help discover better feature representation"



Feature Generation Method

Query1

Doc1 Label
Doc2 Label
Doc3 Label

Query2

Doc1 Label
Doc2 Label
Doc3 Label

x: initial feature representation

Test Query1

Doc1 ?
Doc2 ?
Doc3 ?

Kernel Principal Component Analysis outputs projection matrix **A**

predict

Query1

Doc1 Label
Doc2 Label
Doc3 Label

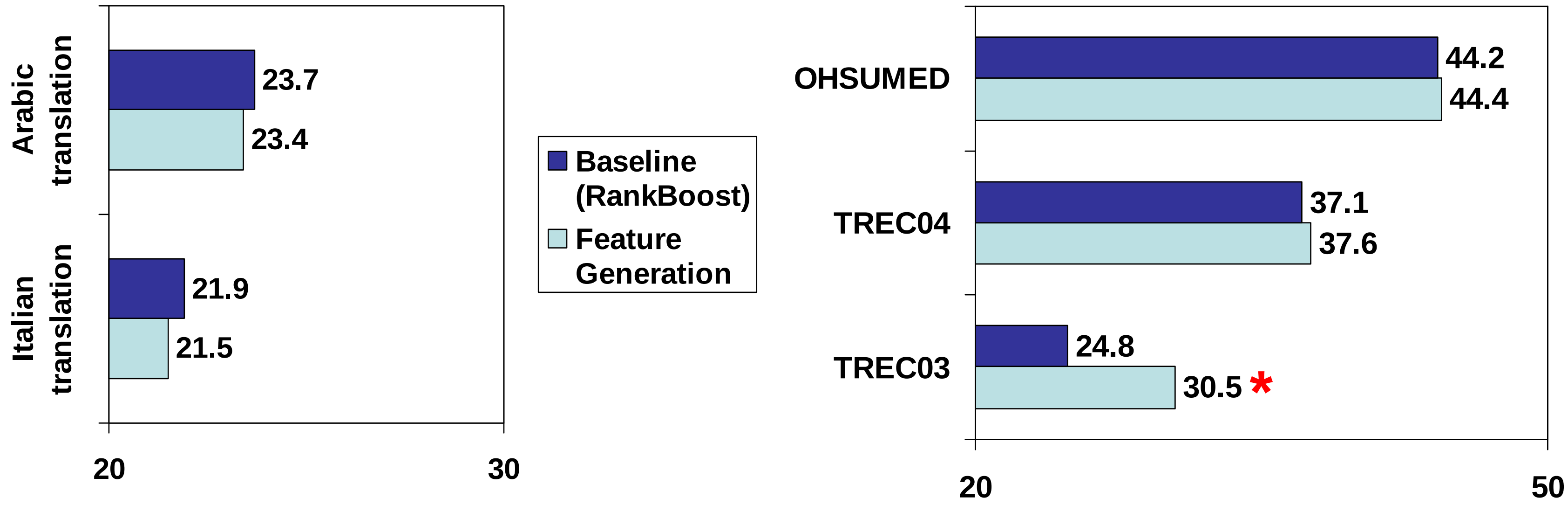
Query2

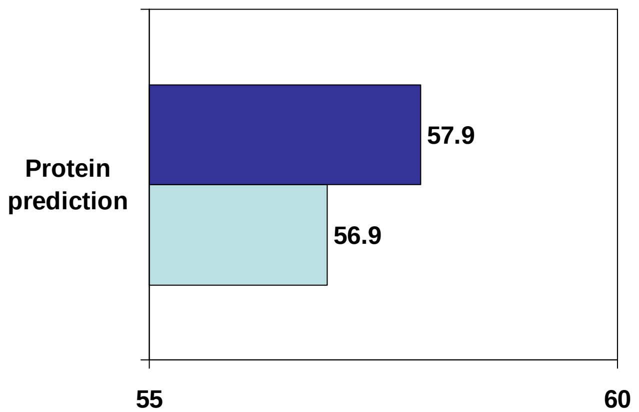
Doc1 Label
Doc2 Label
Doc3 Label

Ranker trained by Supervised RankBoost

z=A'x: new feature representation

Evaluation (Feature Generation)





- Feature Generation works for Information Retrieval

-But degrades for other datasets

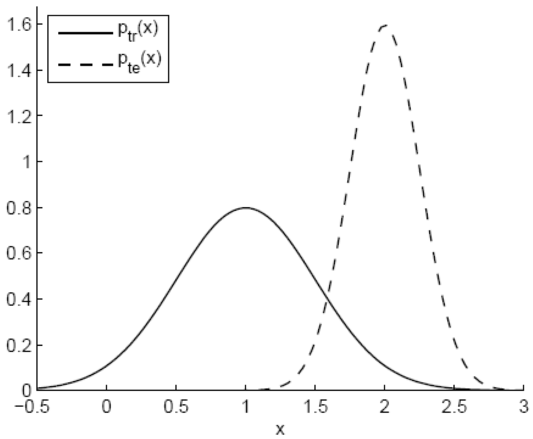
Analysis: Why didn't it work for Machine Translation?

- 40% of weights are for Kernel PCA features
- Pairwise Training accuracy actually improves:
  - 82% (baseline) → 85% (Feature Generation)

- We're increasing the model space *and* optimizing on the wrong loss function
- Feature Generation more appropriate if pairwise accuracy correlates with evaluation metric

Covariate Shift Assumption in Classification (Domain Adaptation)

If training & test distributions differ in marginals p(x),
optimize on weighted data to reduce bias



$$F_{ERM} = \arg\min_F \frac{1}{n} \sum_{i=1}^{n} Loss(F, x_i, y_i)$$

$$F_{IW} = \arg\min_F \frac{1}{n} \sum_{i=1}^{n} \frac{p_{test}(x_i)}{p_{train}(x_i)} Loss(F, x_i, y_i)$$

KLIEP method [Sugiyama08] for generating importance weights *r*

$$\min_r KL(p_{test}(x) \,||\, r(x) p_{train}(x))$$

Covariate Shift Assumption in Ranking

- Each test list is a "different domain"
- Optimize ***weighted*** pairwise accuracy

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

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

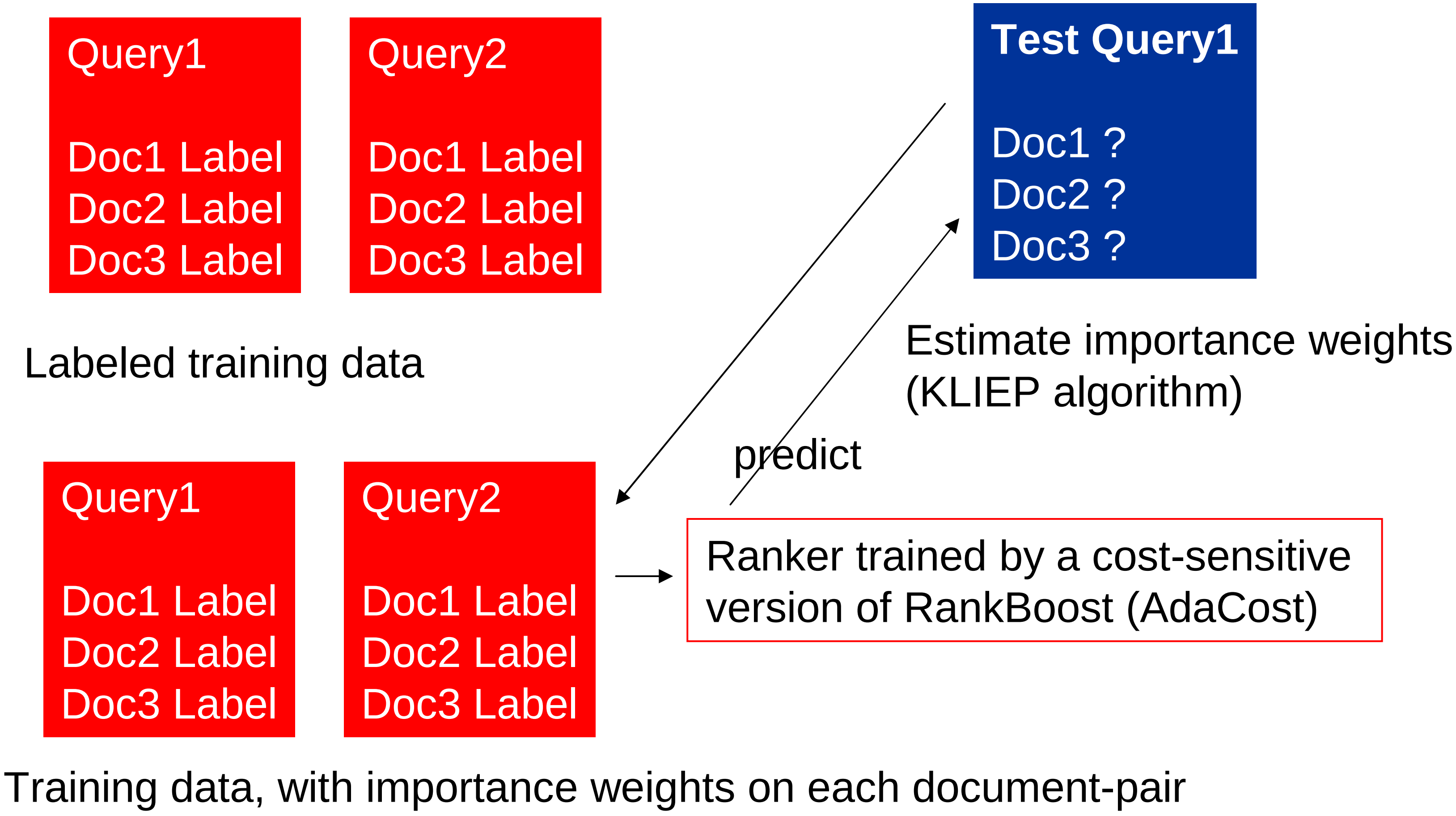
1 $x_3^{(i)} = [tfidf, pagerank, ...]$

$$F(x_1^{(i)}) > F(x_2^{(i)})$$
$$F(x_2^{(i)}) > F(x_3^{(i)})$$
$$F(x_1^{(i)}) > F(x_3^{(i)})$$

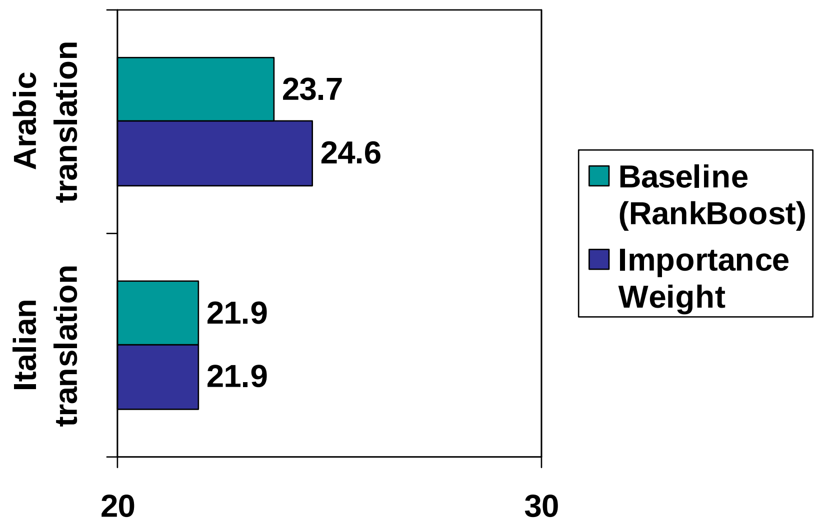
- Define density on pairs

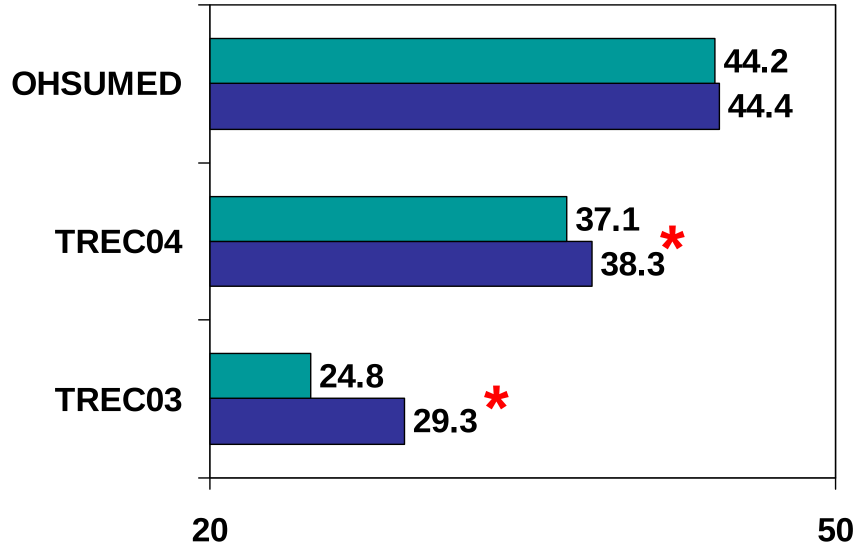
$$p_{train}(x) \rightarrow p_{train}(s) \qquad s = x_p^{(i)} - x_q^{(i)}$$

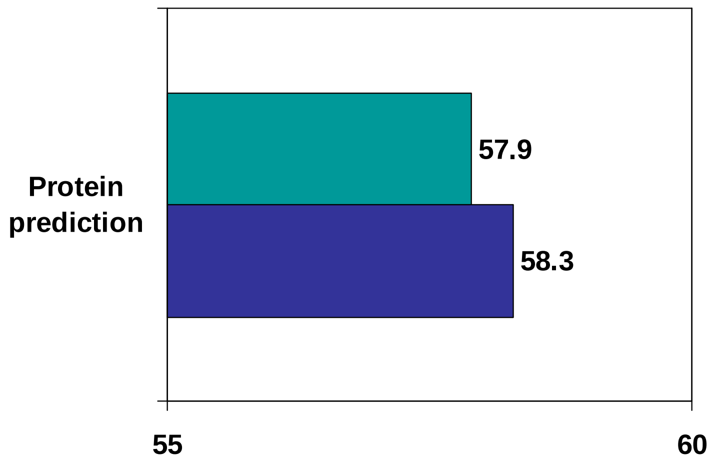
Importance Weighting Method



Evaluation (Importance Weighting)







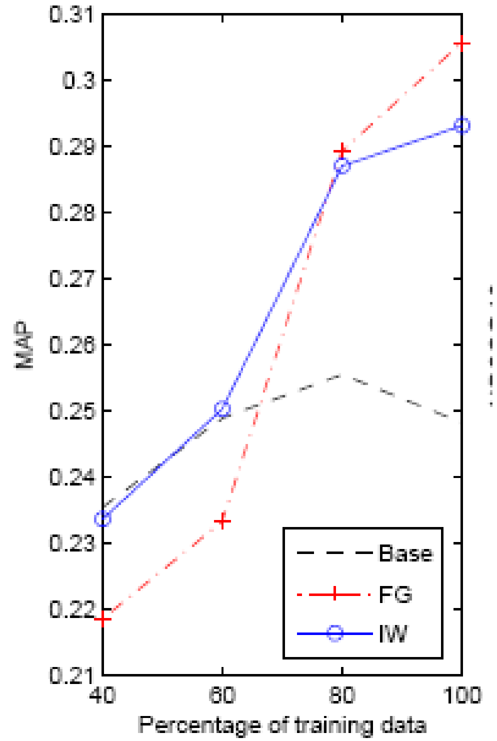
Importance Weighting is a stable method that improves or equals Baseline

Stability Analysis

How many lists are improved/degraded by the method?
Importance Weighting is most conservative
and rarely degrades in low data scenario

PROTEIN PREDICTION	% lists changed
Importance Weighting	32%
Feature Generation	45%
Pseudo Margin (next)	70%

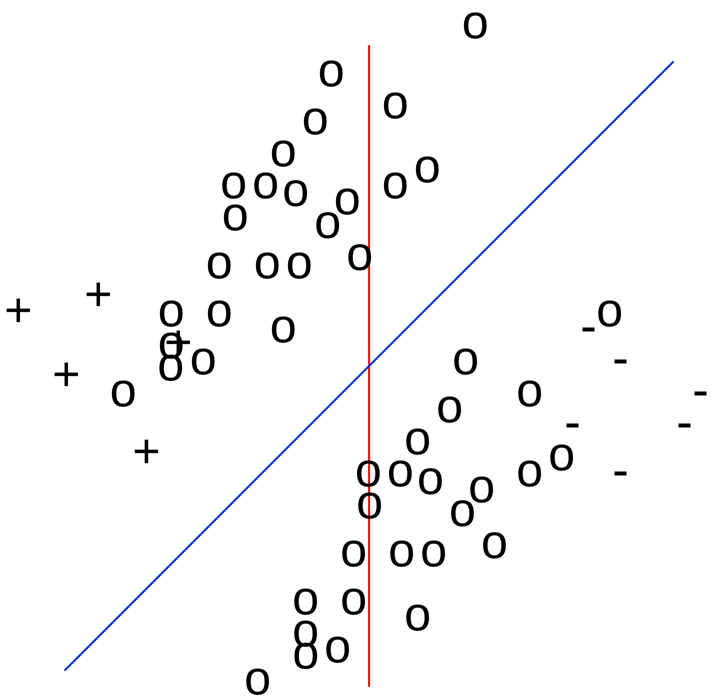
TREC'03 Data Ablation



Low Density Separation Assumption in Classification

Classifier cuts through low density region,

revealed by clusters of data



Algorithms:

Transductive SVM [Joachim'99]

Boosting with Pseudo-Margin [Bennett'02]

$$\min \sum_{i \in labeled} \underbrace{\exp(-y_i F(x_i))}_{} + \sum_{i \in unlabeled} \underbrace{\exp(-|F(x_i)|)}_{}$$

margin= "distance" to hyperplane

pseudo margin= distance to hyperplane assuming correct prediction

Low Density Separation in Ranking

Test Query1

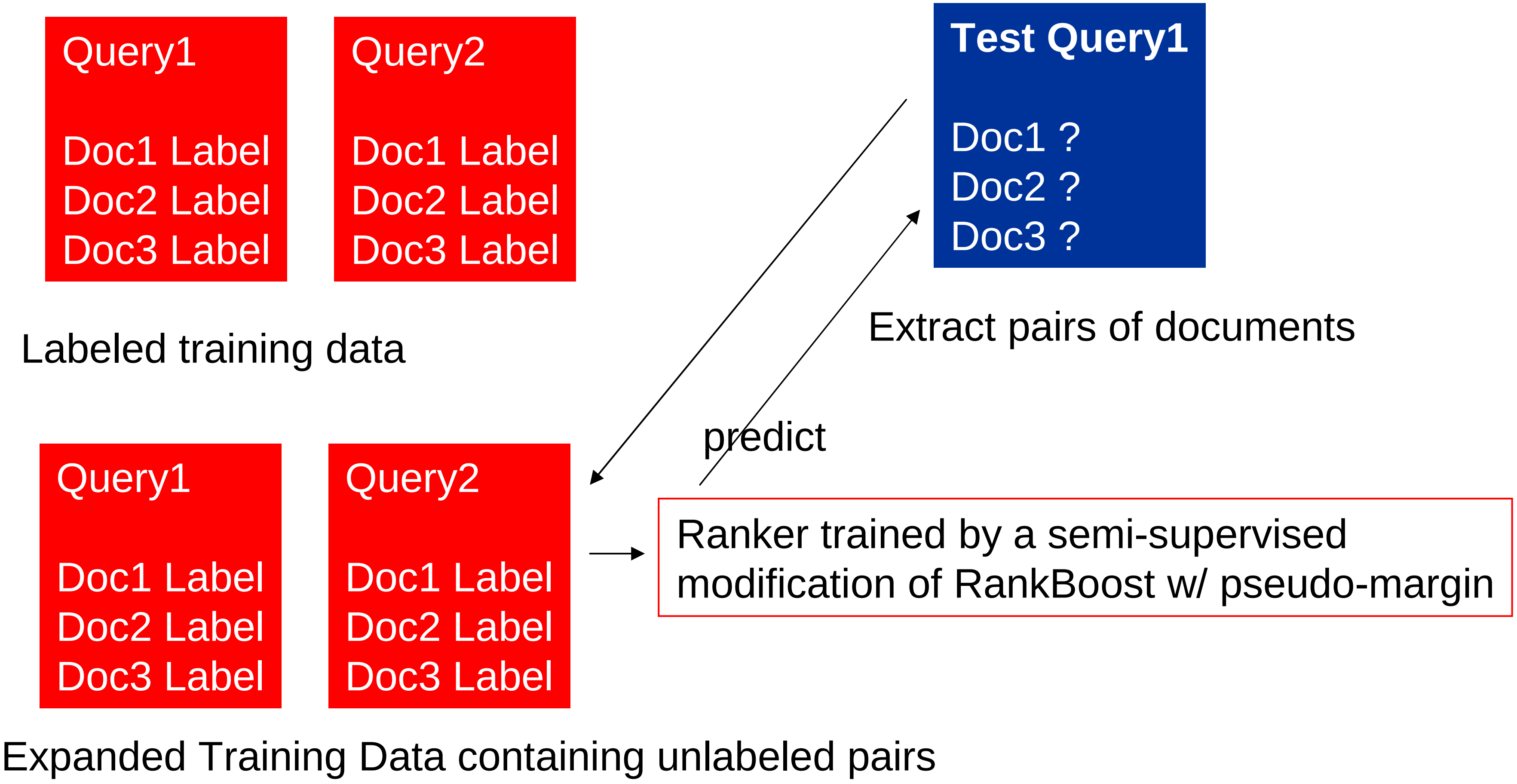
Doc1 ?
Doc2 ?
Doc3 ?

- 1 vs 2: F(Doc1)>>F(Doc2) or F(Doc2)>>F(Doc1)
- 2 vs 3: F(Doc2)>>F(Doc3) or F(Doc3)>>F(Doc2)
- 1 vs 3: F(Doc1)>>F(Doc3) or F(Doc3)>>F(Doc1)

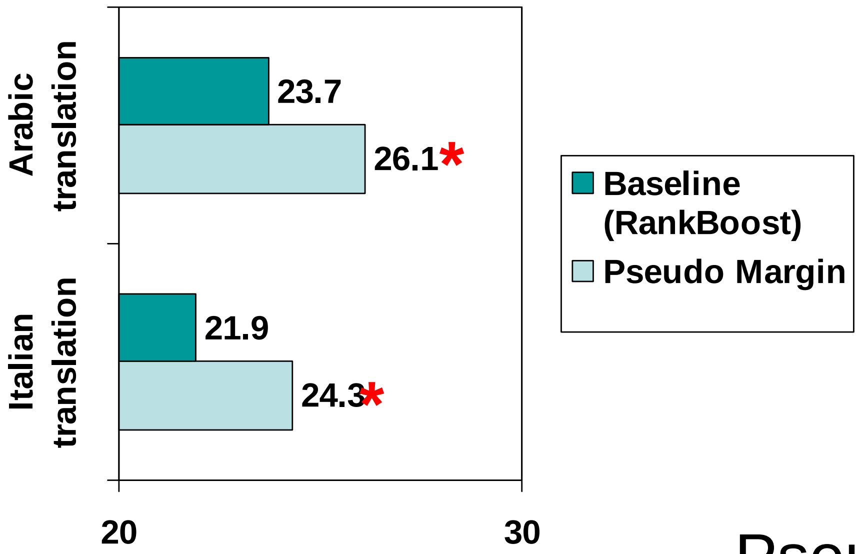
- Define Pseudo-Margin on unlabeled document pairs

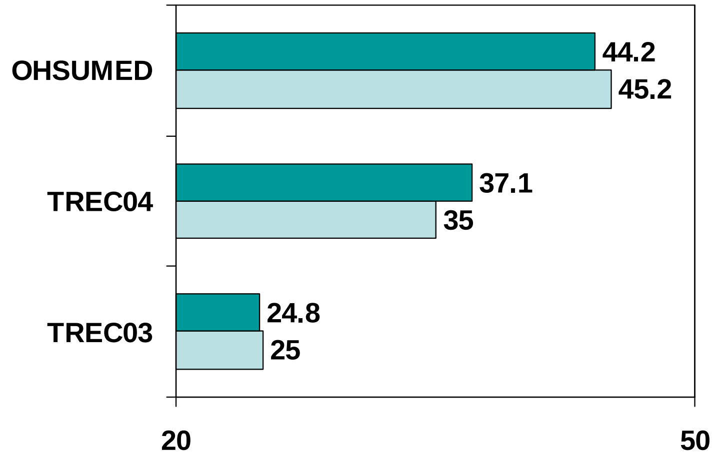
$$\sum_{(i,j)\in labeled} \exp(-(F(x_i)-F(x_j))) + \sum_{(i,j)\in unlabeled} \exp(-|F(x_i)-F(x_j)|)$$

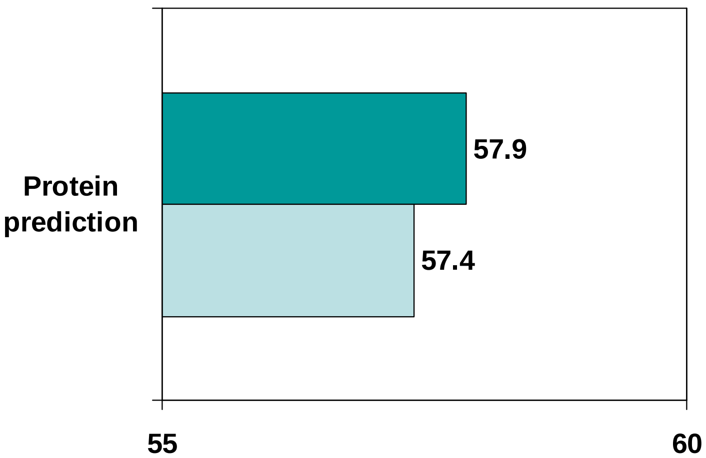
Pseudo Margin Method



Evaluation (Pseudo Margin)







- Pseudo Margin improves for Machine Translation
- Degrades for other tasks

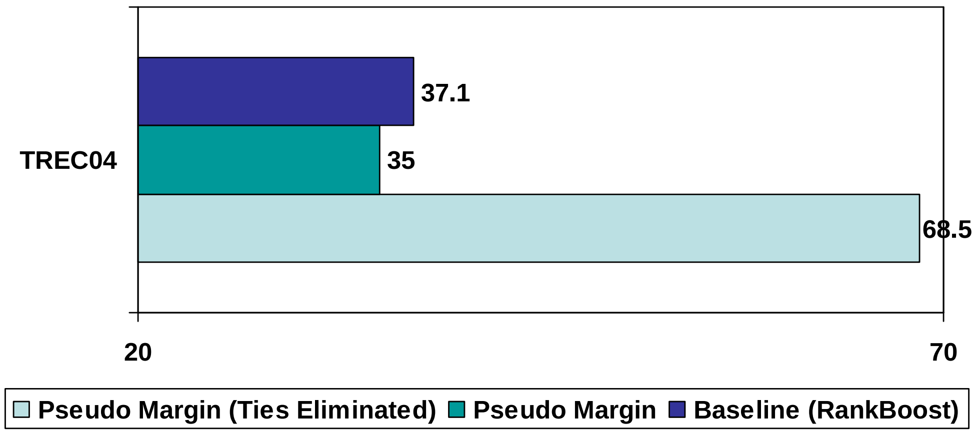
Analysis: Tied Ranks and Low Density Separation

Test Query1

Doc1 ?
Doc2 ?
Doc3 ?

- 1 vs 2: F(Doc1)>>F(Doc2) or F(Doc2)>>F(Doc1)
- Ignores the case F(Doc1)=F(Doc2)

- But most documents are tied in Information Retrieval!
- If tied pairs are eliminated from semi-cheating experiment, Pseudo Margin improves drastically



Outline

1. Problem Setup
2. Investigating the Manifold Assumption
3. Local/Transductive Meta-Algorithm
    1. Change of Representation Assumption
    2. Covariate Shift Assumption
    3. Low Density Separation Assumption
4. Summary

Contribution 1

Investigated 4 assumptions on how unlabeled data helps ranking

- Ranker Propagation:
  - assumes ranker vary smoothly over manifold on lists
- Feature Generation method:
  - use on unlabeled test data to learn better features
- Importance Weighting method:
  - select training data to match the test list's distribution
- Pseudo Margin method:
  - assumes rank differences are large for unlabeled pairs

Contribution 2

Comparison on 3 applications, 6 datasets

	Information Retrieval	Machine Translation	Protein Prediction
Ranker Propagation	=	IMPROVE	BEST
Feature Generation	IMPROVE	DEGRADE	=
Importance Weighting	BEST	=	=
Pseudo Margin	=	BEST	=

Future Directions

- Semi-supervised ranking works! Many future directions are worth exploring:
  - Ranker Propagation with Nonlinear Rankers
  - Different kinds of List Kernels
  - Speed up Local/Transductive Meta-Algorithm
  - Inductive semi-supervised ranking algorithms
  - Statistical learning theory for proposed methods

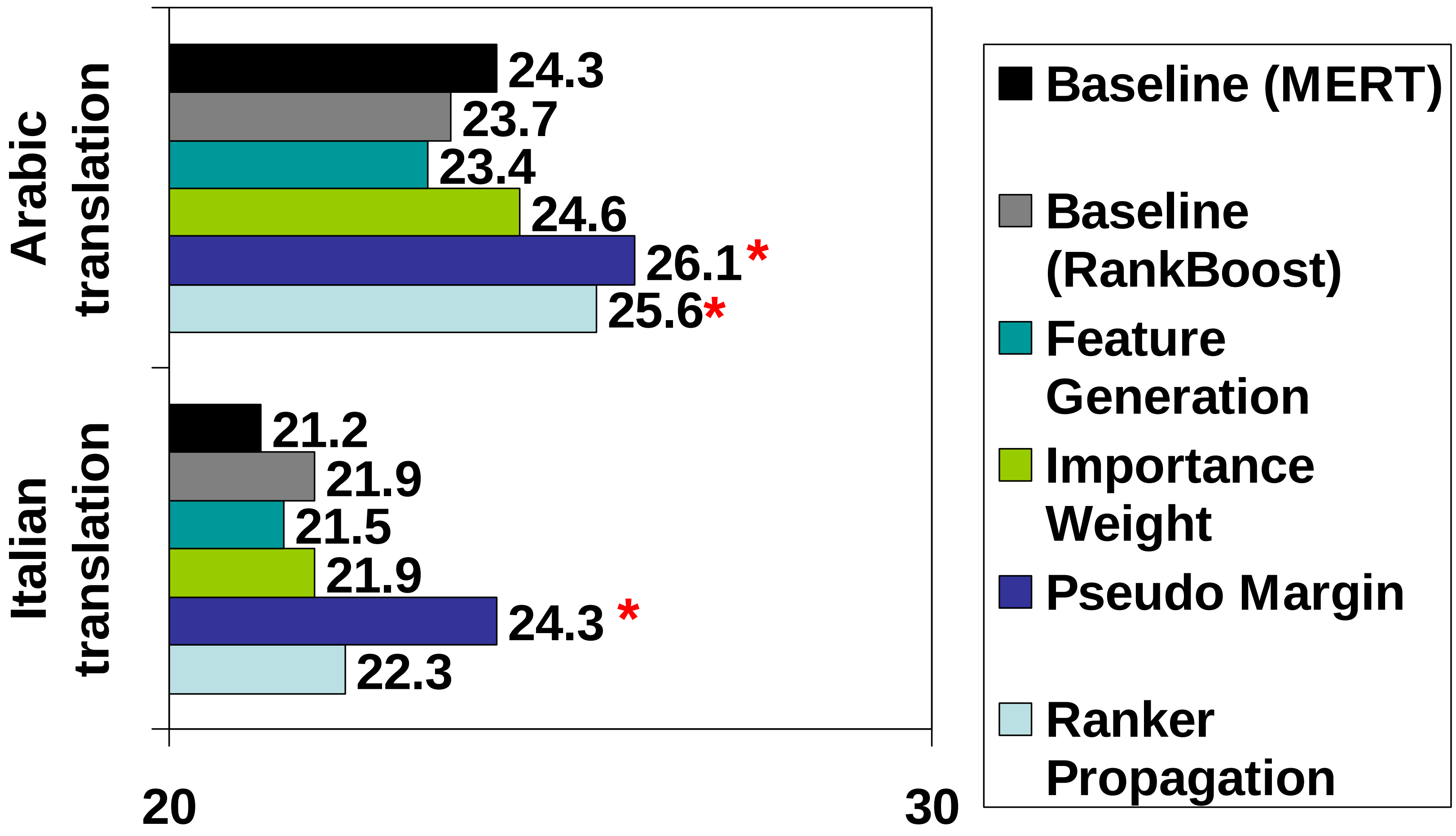
Thanks for your attention!

- Questions? Suggestions?

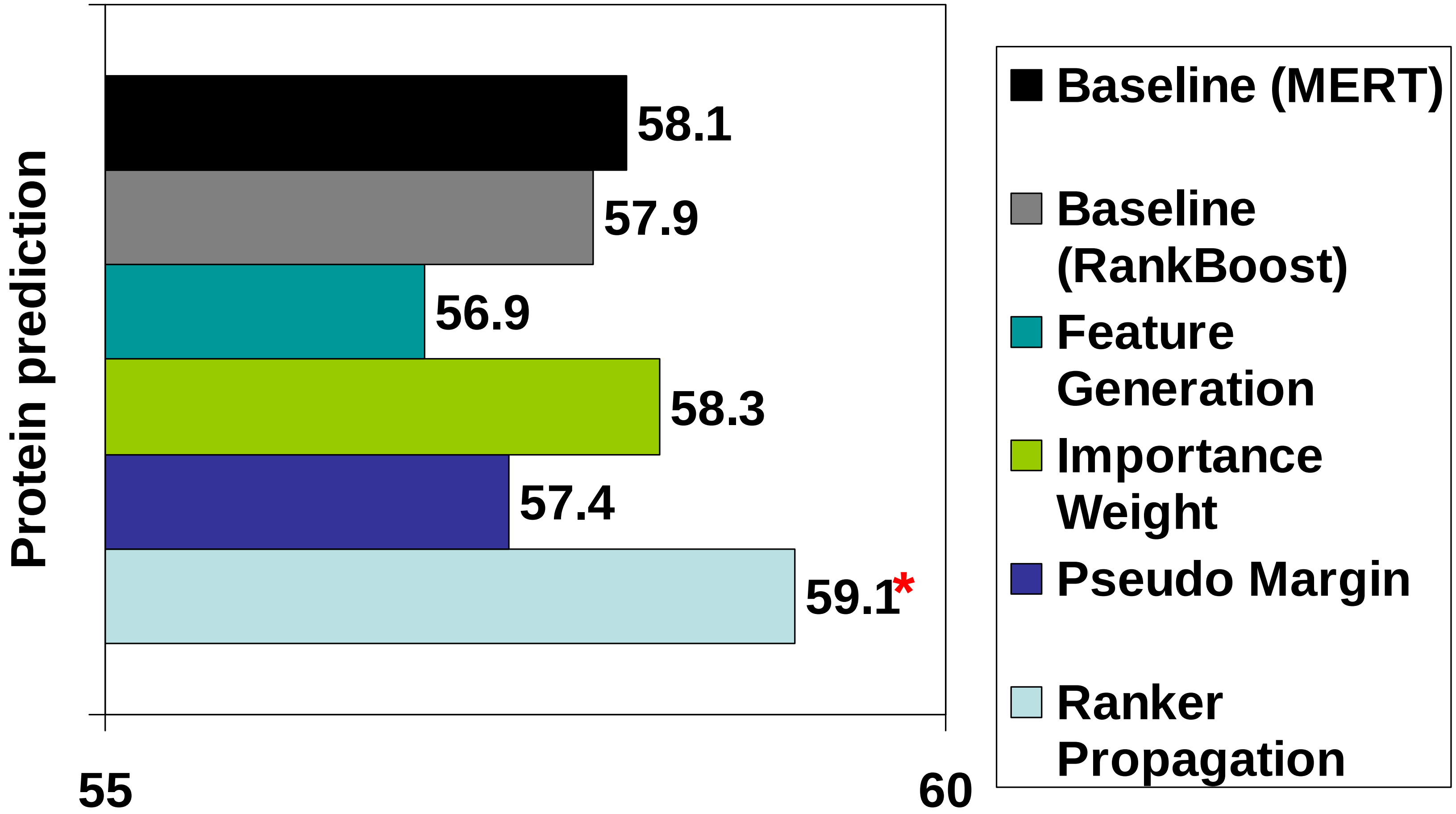
- Acknowledgments:
  - NSF Graduate Fellowship (2005-2008)
  - RA support from my advisor's NSF Grant IIS-0326276 (2004-2005) and NSF Grant IIS-0812435 (2008-2009)

- Related publications:
  - Duh & Kirchhoff, *Learning to Rank with Partially-Labeled Data*, ACM SIGIR Conference, 2008
  - Duh & Kirchhoff, *Semi-supervised Ranking for Document Retrieval*, under journal review

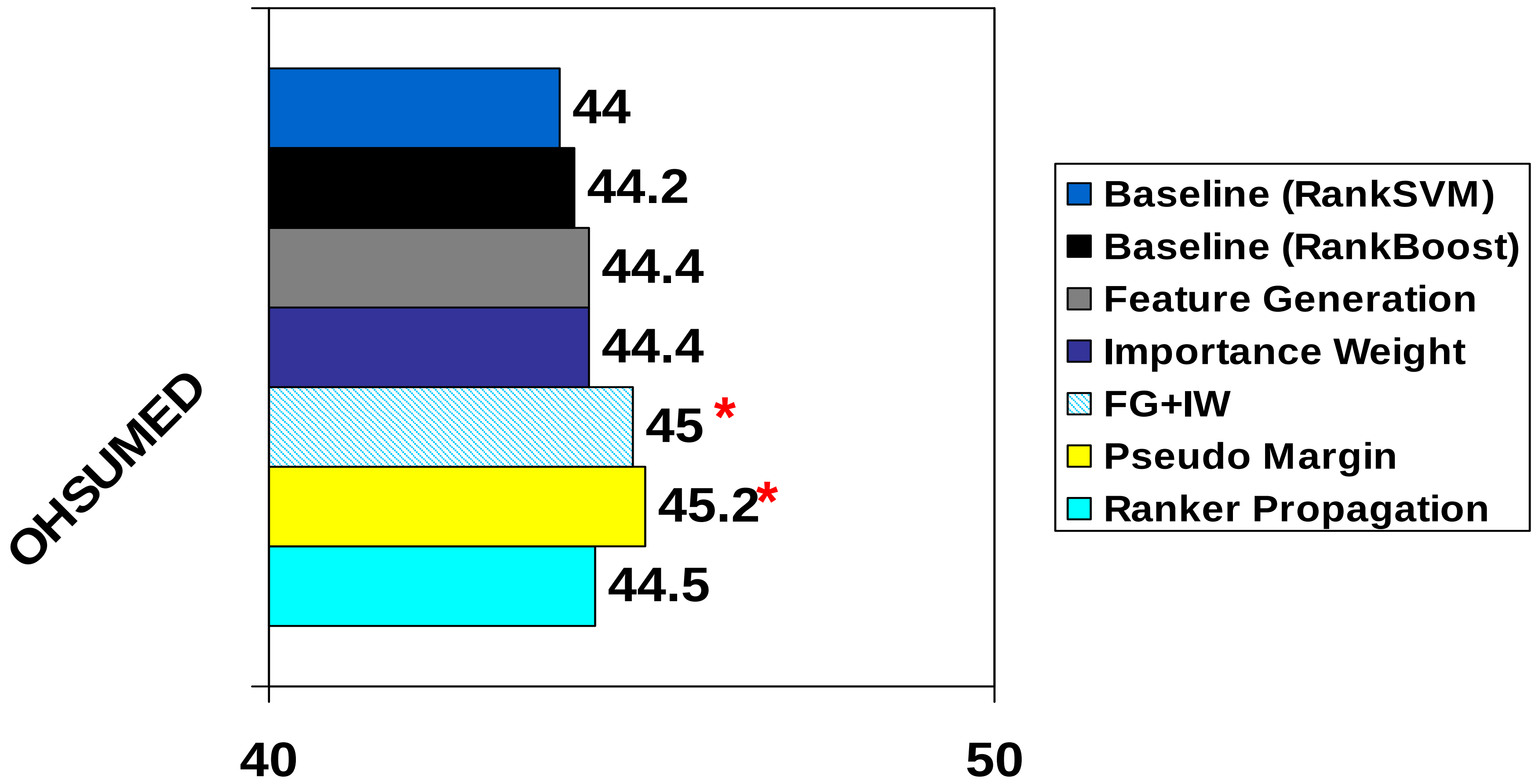
Machine Translation: Overall Results



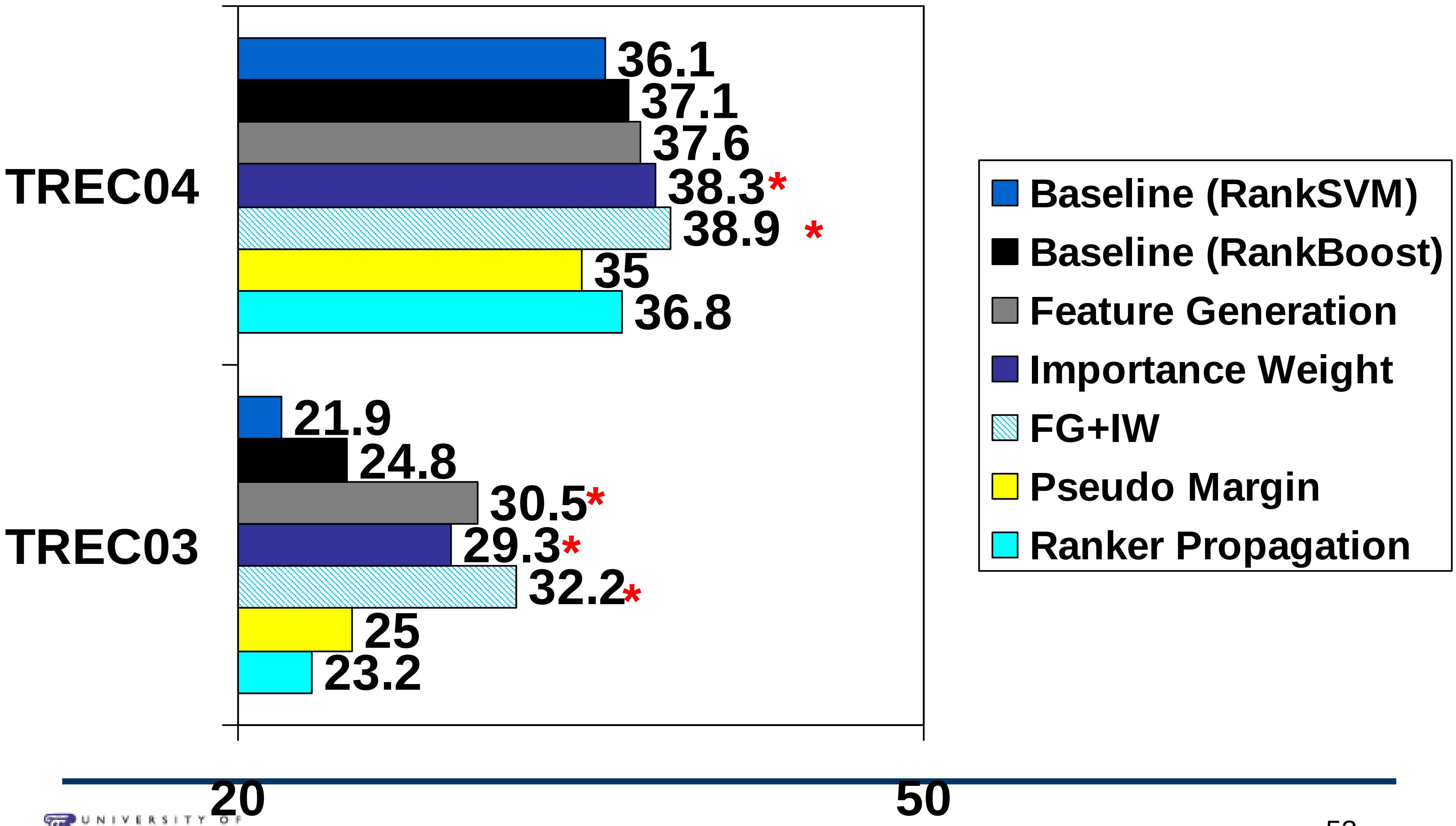
Protein Prediction: Overall Results



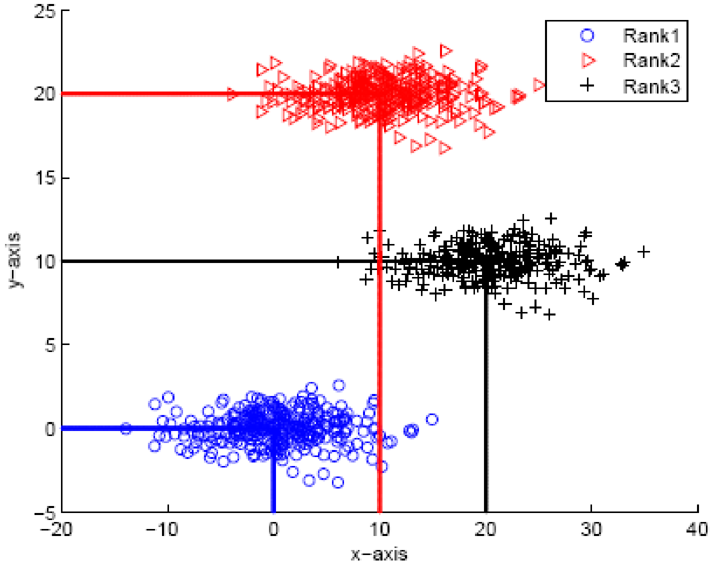
OHSUMED: Overall Results



TREC: Overall Results



Supervised Feature Extraction for Ranking



Linear Discriminant Analysis (LDA)

$$\arg\max_{\alpha} \frac{\alpha^T B \alpha}{\alpha^T W \alpha}$$

B: between-class scatter
W: within-class scatter

RankLDA $\arg\max_{\alpha} \frac{\alpha^T \tilde{B} \alpha}{\alpha^T W \alpha}$

$$s.t. \quad \alpha^T B_{13} \alpha > \alpha^T B_{12} \alpha$$

$$\alpha^T B_{13} \alpha > \alpha^T B_{23} \alpha$$

OHSUMED
Baseline: 44.2
Feature Generation:44.4
w/ RankLDA: 44.8

KLIEP Optimization

$$KL(p_{test}(x)//w(x)*p_{train}(x)) = \int p_{test}(x)\log\frac{p_{test}(x)}{w(x)*p_{train}(x)}dx$$

$$= \int p_{test}(x)\log\frac{p_{test}(x)}{p_{train}(x)}dx - \int p_{test}(x)\log w(x)dx$$

$$\mathcal{O}_{KLIEP} = \int p_{test}(x)\log w(x)dx$$

$$\approx \frac{1}{U_{pair}}\sum_{u=1}^{U_{pair}}\log w(x_u)$$

$$= \frac{1}{U_{pair}}\sum_{u=1}^{U_{pair}}\log\sum_{b=1}^{B}\beta_b\psi_b(x_u)$$

constraints that $\beta \geq 0$

$$1 = \int w(x)p_{train}(x)dx \approx \frac{1}{L_{pair}}\sum_{x=1}^{L_{pair}}\sum_{b}^{B}\beta_b\psi(x_l)$$

List Kernel Proof: Symmetricity

Proposition 8.3.1. *The function $K(x,y)$ in Algorithm 10 is symmetric, i.e. $K(x,y) = K(y,x)$.*

Proof.

$$
\begin{aligned}
K(x,y) &= \frac{\sum_{m=1}^{M} \lambda_x^m \lambda_y^{a(m)} \cdot |< u_x^m, u_y^{a(m)} >|}{(||\lambda_x|| \cdot ||\lambda_y||)} \\
&= \frac{\sum_{m=1}^{M} \lambda_y^{a(m)} \lambda_x^m \cdot |< u_y^{a(m)}, u_x^m >|}{(||\lambda_y|| \cdot ||\lambda_x||)} \\
&= \frac{\sum_{m=1}^{M} \lambda_y^m \lambda_x^{a^{-1}(m)} \cdot |< u_y^m, u_x^{a^{-1}(m)} >|}{(||\lambda_y|| \cdot ||\lambda_x||)} \\
&= K(y,x)
\end{aligned}
$$

List Kernel Proof: Cauchy-Schwartz Inequality

Proposition 8.3.2. *The function $K(x,y)$ in Algorithm 10 is satisfies the Cauchy-Schwartz Inequality, i.e. $K(x,y)^2 \leq K(x,x)K(y,y)$.*

Proof. First, we show that $K(x,x) = 1$:

$$\begin{aligned} K(x,x) &= \frac{\sum_{m=1}^{M} \lambda_x^m \lambda_x^{a(m)} \cdot |< u_x^m, u_x^{a(m)} >|}{(||\lambda_x|| \cdot ||\lambda_x||)} \\ &= \frac{\sum_{m=1}^{M} \lambda_x^m \lambda_x^m \cdot |< u_x^m, u_x^m >|}{(||\lambda_x|| \cdot ||\lambda_x||)} \\ &= \frac{\sum_{m=1}^{M} \lambda_x^m \lambda_x^m}{(||\lambda_x|| \cdot ||\lambda_x||)} \\ &= \frac{||\lambda||^2}{(||\lambda_x|| \cdot ||\lambda_x||)} = 1 \end{aligned}$$

The second step follows from the fact that maximum bipartite matching would achieve $a(m) = m \; \forall m$ since $< u_x^m, u_x^m >= 1$ and $< u_x^m, u_x^{m'} >= 0$ for any $m \neq m'$. The third step is a result of $< u_x^m, u_x^m >= 1$.

Next we show that $K(x,y)^2$ is bounded by 1. Note that $< u_x^m, u_y^{a(m)} > \leq 1$, so that $K(x,y) \leq \frac{\sum_{m=1}^{M} \lambda_x^m \lambda_y^{a(m)}}{(||\lambda_x|| \cdot ||\lambda_y||)} \leq 1$ where the last inequality follows from applying Cauchy-Schwartz to the vectors of eigenvalues. □

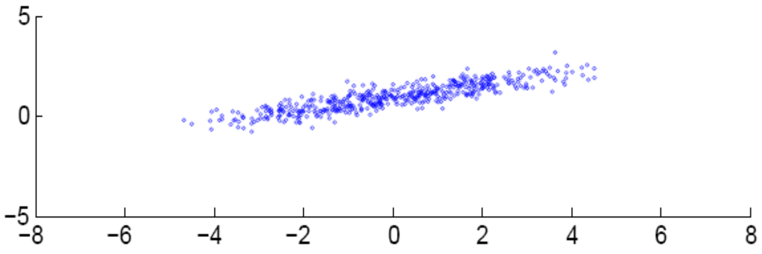
List Kernel Proof: Mercer's Theorem

Theorem 8.3.3 (Mercer's Theorem, c.f. [123]). *Every positive (semi) definite, symmetric function is a kernel: i.e., there exists a feature mapping* ϕ *such that it is possible to write:* $K(x,y) =< \phi(x), \phi(y) >$.

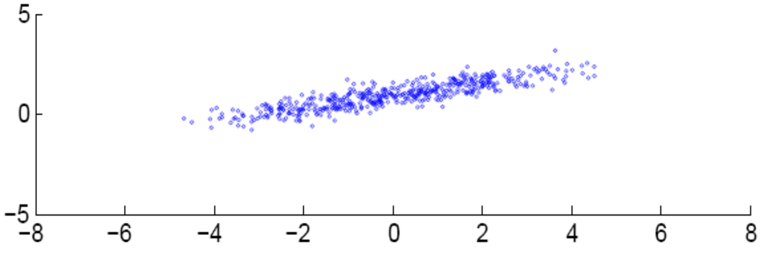
Proposition 8.3.4. *The function* $K(x,y)$ *in Algorithm 10 satisfies the Mercer Theorem.*

Proof. We have already proved that $K(x,y)$ is symmetric. To see that it is positive semi-definite, we just need to observe that $K(x,y) \geq 0$ for any x,y. We prove this by contradiction: Suppose $K(x,y) < 0$ for some x,y. This implies that $\sum_{m=1}^{M} \lambda_x^m \lambda_y^{a(m)} \cdot | < u_x^m, u_y^{a(m)} > |$ is negative. However, by construction, we will only obtain non-negative eigenvalues λ_x from PCA. Further, the absolute value operation $| < u_x^m, u_y^{a(m)} > |$ ensures non-negativity. Thus, the statement that $K(x,y) < 0$ for some x,y is false. $\square$

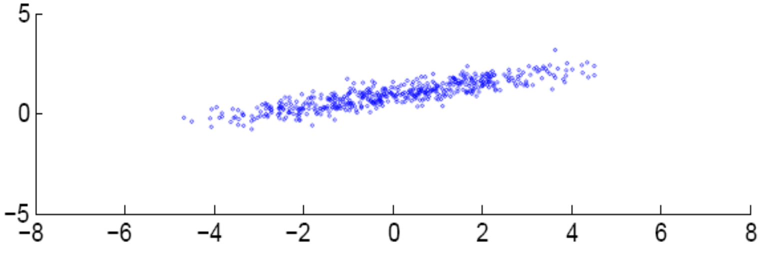
Invariance Properties for Lists



Shift-invariance



Scale-invariance



Rotation-invariance