



# 基于机器学习的排序技术



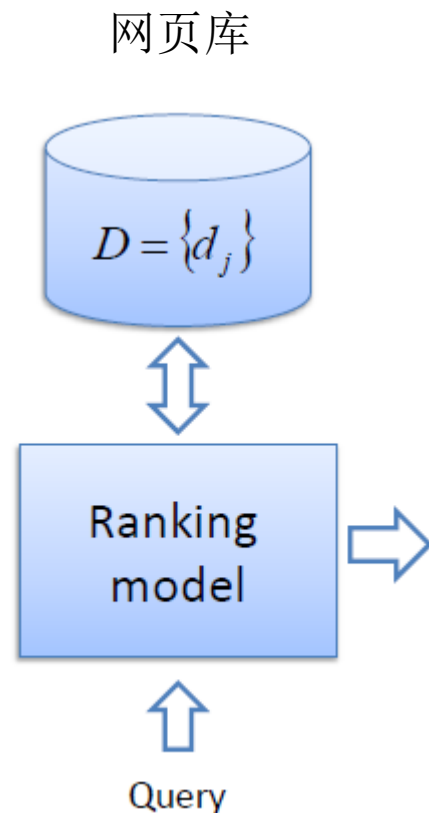
蒋龙

2012年12月29日

## 一、排序技术介绍

## 二、Learning to Rank介绍

# 网页搜索排序



Baidu 百度 新闻 网页 贴吧 知道 MP3 图片 视频 地图 更多▼

互联网广告

百度一下

[万通网维网络广告一条龙](http://www.beijingweb.net) [www.beijingweb.net](http://www.beijingweb.net)

网络广告专业机构,域名空间,网页设计,全面搜索引擎服务

[互联网营销 找三打哈 百万推手共同帮你推广](http://www.sandaha.com) [www.sandaha.com](http://www.sandaha.com)

三打哈网—中国专业的推广服务平台,聚集了百万推手人才,为您提供论坛发帖,

[互联网广告](#) [百度百科](#)

概念 [互联网广告](#)就是在网络上做的广告。利用网站上的广告横幅、文本链接、多媒体的方法,在互联网刊登或发布广告,通过网络传递到互联网用户的一种...共10次编辑

[概念](#) - [起源](#) - [分类](#) - [互联网广告企业](#)

[baike.baidu.com/view/1239656.htm](http://baike.baidu.com/view/1239656.htm) 2011-12-23

[中国网络广告公司TOP 50排行榜](#) [互联网](#) [科技时代](#) [新浪网](#)

2010年8月5日... 本次针对网络广告营销公司的调查和评选主要由《[互联网周刊](#)》编辑部设置具体指标,本刊联合市场研究公司、数据公司、广告主、门户网站和清华大学新媒体传播...

[tech.sina.com.cn/i/2010-08-05/18374512650 ...](http://tech.sina.com.cn/i/2010-08-05/18374512650...) 2010-8-5 - [百度快照](#)

[移动互联网广告](#) [百度知道](#)

目前中国移动[互联网广告](#)事业高速发展,最近几年是发展迅猛期。国内已形成规模的四大移动

[互联网广告](#)平台:一、赢告。先天优势,客户以及合作伙伴强大,有新浪、凤凰、...

[zhidao.baidu.com/question/289420932.html](http://zhidao.baidu.com/question/289420932.html) 2011-7-18

[互联网广告要多少钱](#) 4个回答 2011-5-14

[福州有什么好的广告公司、网络公司、互联网公司的。](#) 5个回答 2012-1-17

[更多知道相关问题>>](#)

- 文档检索
- 商品评分
- 推荐
- 协同过滤
- 关键词抽取

- 测试数据集
  - 随机选择一批query，每个query下选择一批文档
  - 为每个query下的文档标注相关性分数
- 计算量化的评测效果
  - 计算每个query下的效果指标
  - 计算整个测试集的效果指标

- 相关度绝对分数
  - 二值标注：相关VS不相关
  - 多级相关性
    - perfect>excellent>good>fair>bad
    - $5 > 4 > 3 > 2 > 1$
- 相对排序
  - $A > B$  or  $A < B$
- 完全排序
  - 对一个Query下的所有文档排序

- MAP (Mean Average Precision)

- 二值标注：相关，不相关

- Precision at position  $k$  for query  $q$ :

$$P@k = \frac{\#\{\text{relevant documents in top } k \text{ results}\}}{k}$$

- Average precision for query  $q$ :

$$AP = \frac{\sum_k P@k \cdot l_k}{\#\{\text{relevant documents}\}}$$



$$AP = \frac{1}{3} \cdot \left( \frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$$

- MAP: averaged over all queries.

- NDCG
  - 可用于多级标注

$$N(n) = \underbrace{Z_n}_{\text{Normalization}} \underbrace{\sum_{j=1}^n}_{\text{Cumulating}} \underbrace{(2^{r(j)} - 1)}_{\text{Gain}} \underbrace{/\log(1+j)}_{\text{Position discount}}$$



- Mean reciprocal rank
  - 只考虑排位最高的相关文档

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

- 依赖Query

- 布尔模型

- 向量空间模型

$$w_{t,d} = \text{tf}_{t,d} \cdot \log \frac{|D|}{|\{d' \in D \mid t \in d'\}|}$$

- TF\*IDF

- 概率检索模型

$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

- BM25

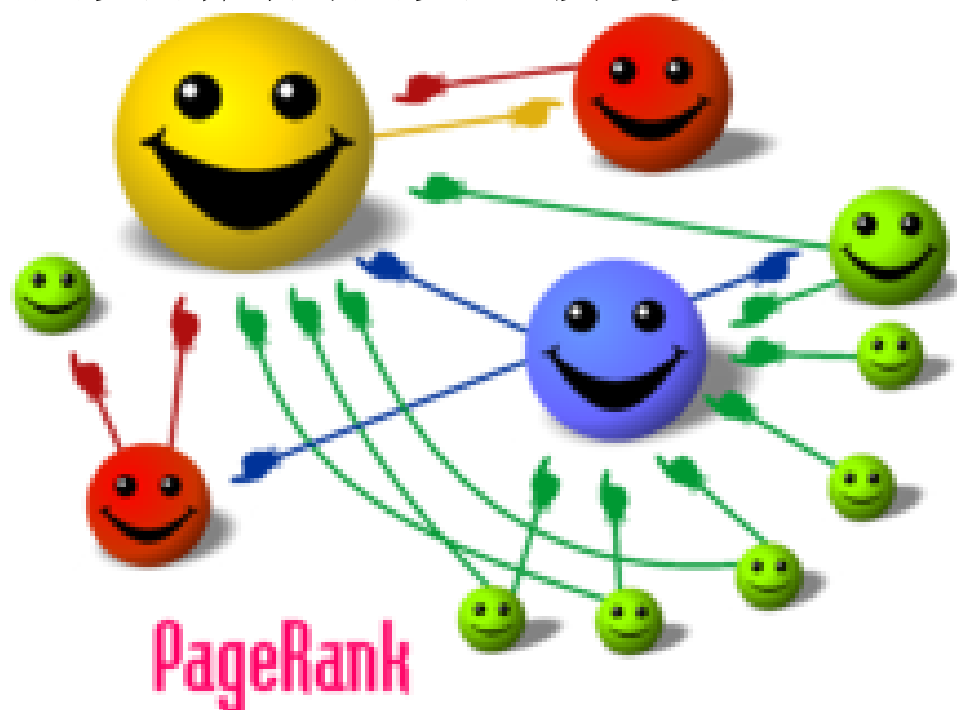
$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

- 统计语言模型

$$\arg \max_D P(D \mid Q) = \arg \max_D P(Q \mid D)P(D)$$

$$P(w \mid D) = \lambda \frac{c(w, D)}{\sum_{w' \in D} c(w', D)} + (1 - \lambda) \frac{c(w, C)}{\sum_{w' \in V} c(w', C)}$$

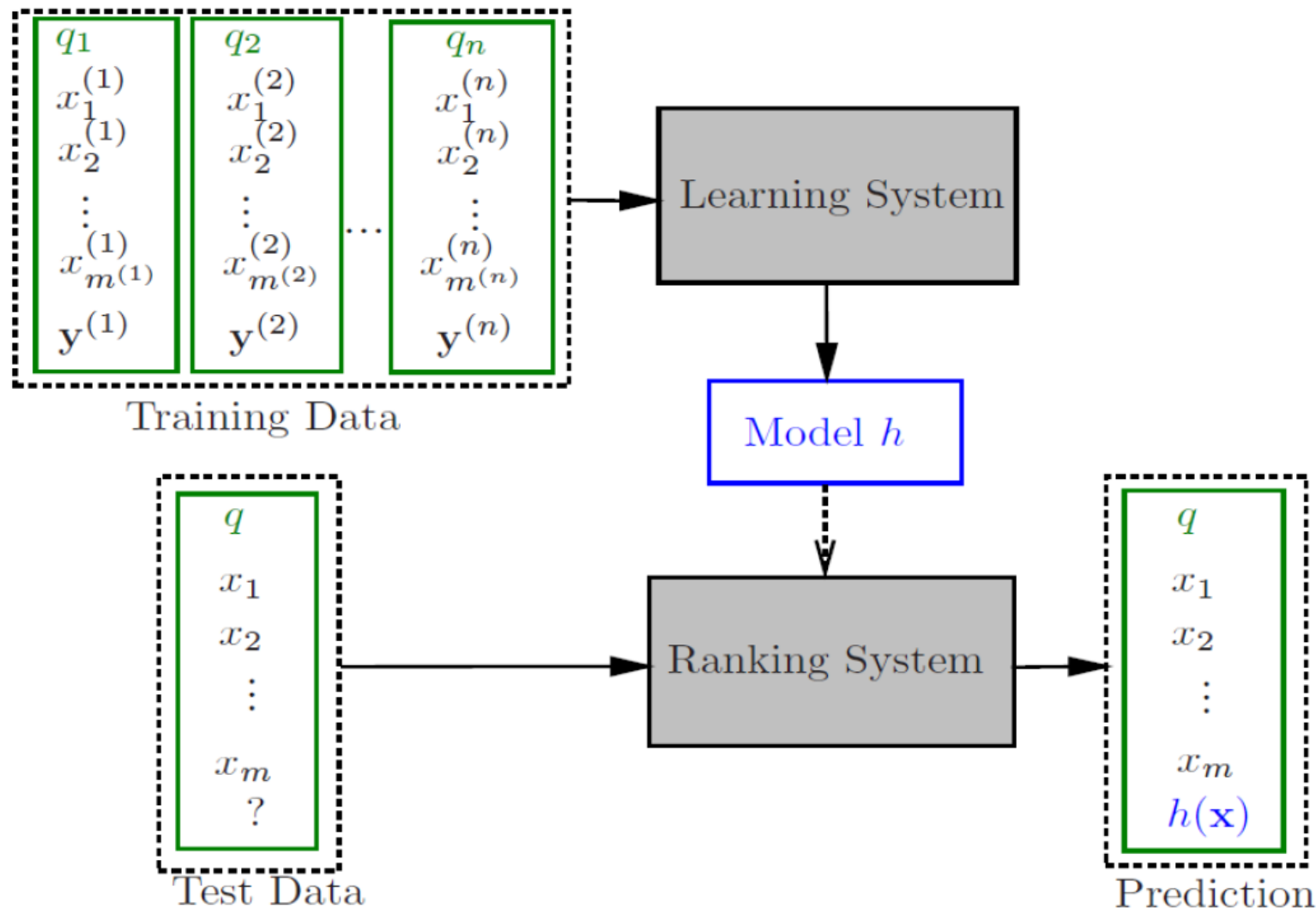
- Query无关
  - Pagerank
    - 被很多网页引用的网页比较重要
    - 被重要网页引用的网页比较重要

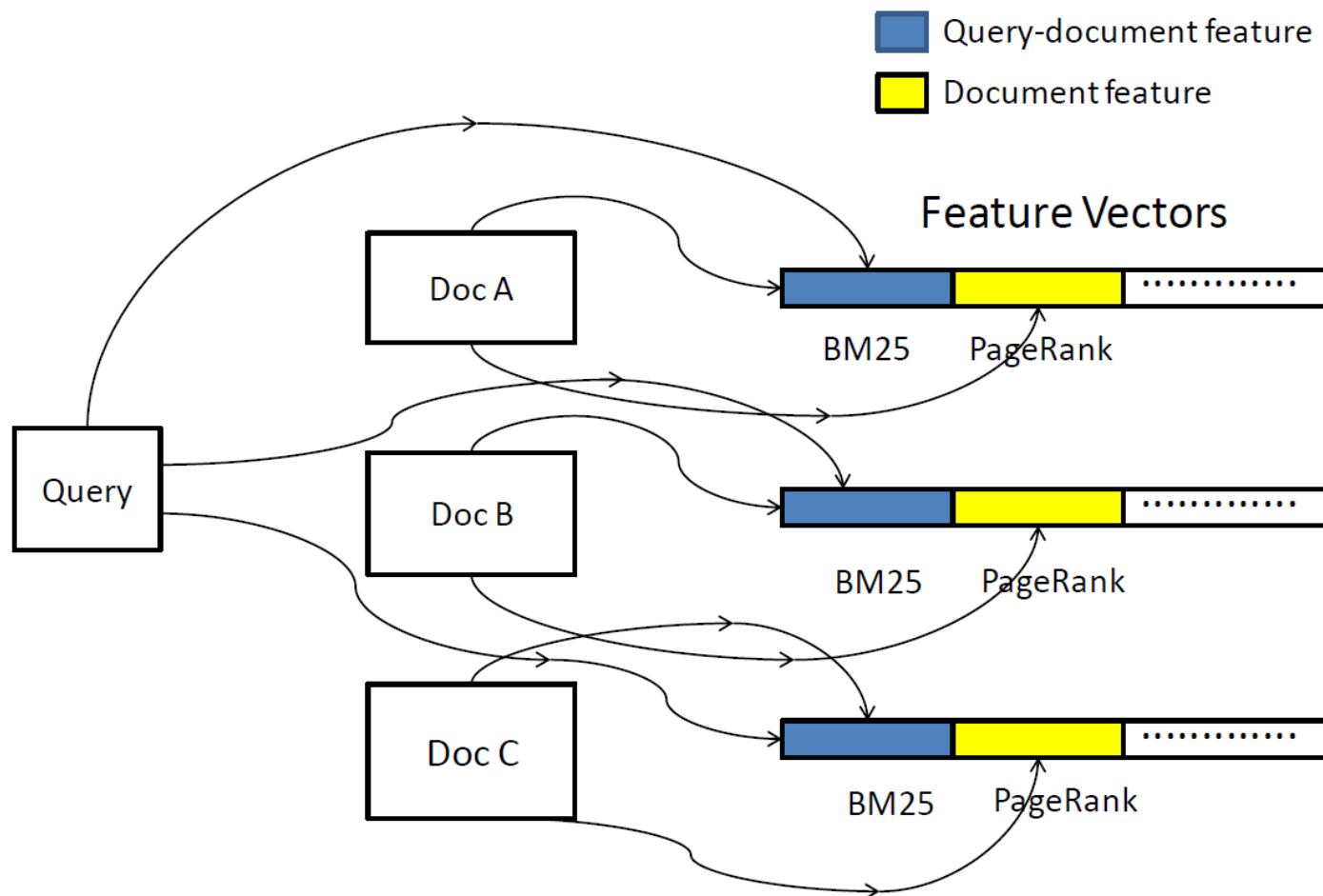


## 一、排序技术介绍

## 二、Learning to Rank介绍

# Learning to Rank





Category	Algorithms
Pointwise Approach	<p><b>Regression:</b> Least Square Retrieval Function (TOIS 1989), Regression Tree for Ordinal Class Prediction (Fundamenta Informaticae, 2000), Subset Ranking using Regression (COLT 2006), ...</p> <p><b>Classification:</b> Discriminative model for IR (SIGIR 2004), McRank (NIPS 2007), ...</p> <p><b>Ordinal regression:</b> Pranking (NIPS 2002), OAP-BPM (EMCL 2003), Ranking with Large Margin Principles (NIPS 2002), Constraint Ordinal Regression (ICML 2005), ...</p>
Pairwise Approach	<p>Learning to Retrieve Information (SCC 1995), Learning to Order Things (NIPS 1998), Ranking SVM (ICANN 1999), RankBoost (JMLR 2003), LDM (SIGIR 2005), RankNet (ICML 2005), Frank (SIGIR 2007), MHR(SIGIR 2007), GBRank (SIGIR 2007), QBRank (NIPS 2007), MPRank (ICML 2007), IRSVM (SIGIR 2006), ...</p>
Listwise Approach	<p><b>Listwise loss minimization:</b> RankCosine (IP&amp;M 2008), ListNet (ICML 2007), ListMLE (ICML 2008), ...</p> <p><b>Direct optimization of IR measure:</b> LambdaRank (NIPS 2006), AdaRank (SIGIR 2007), SVM-MAP (SIGIR 2007), SoftRank (LR4IR 2007), GPRank (LR4IR 2007), CCA (SIGIR 2007), ...</p>


- 把排序问题分别转化为
  - 分类
    - 二值分类：相关VS不相关
    - 多值分类：perfect>excellent>good>fair>bad
  - 回归
    - 为每个文档计算一个相关性分数（连续值）



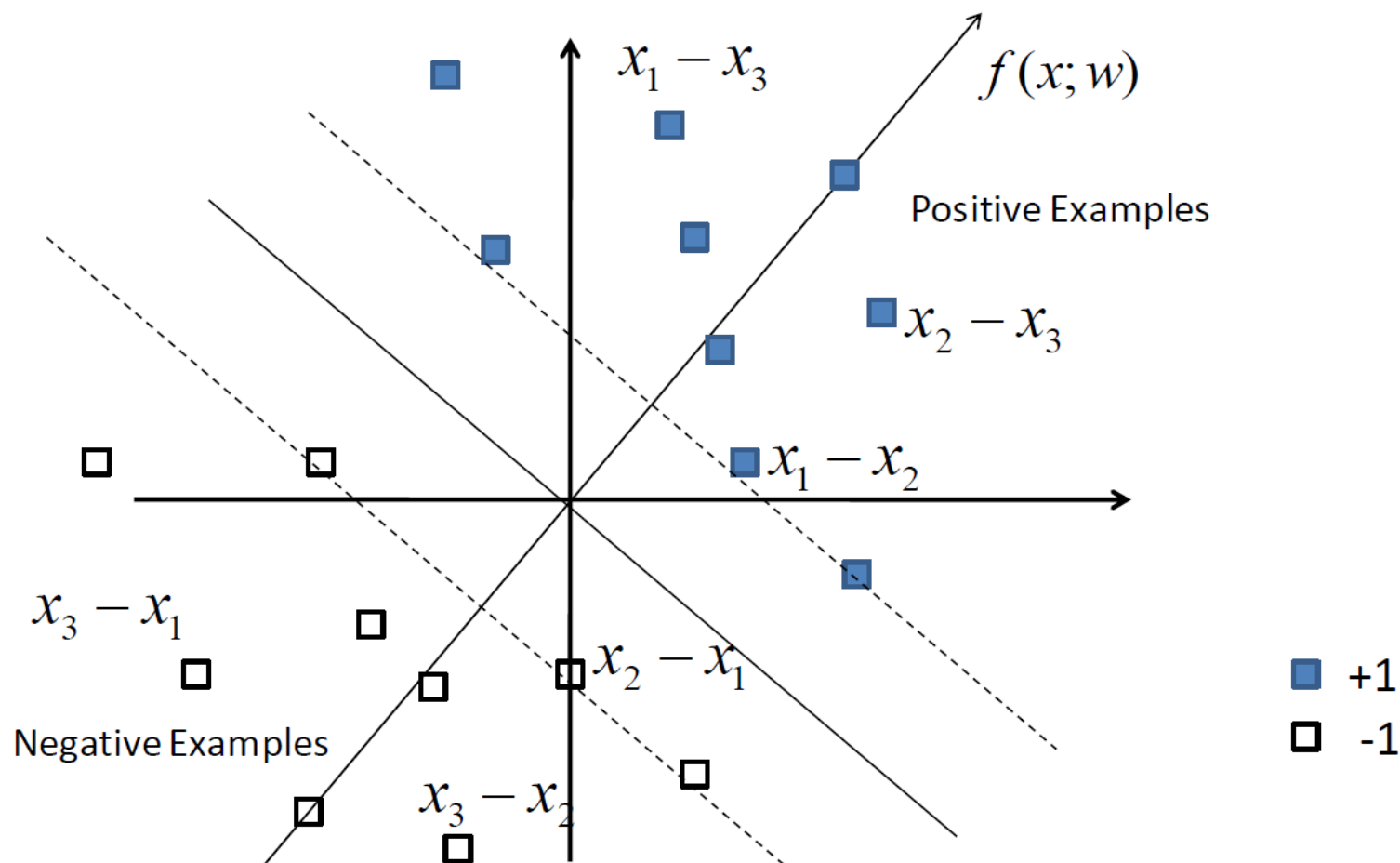
- 对排序问题的特性缺乏考虑
  - 排序的query依赖性
    - 同一文档对某些query是perfect，但对于其他query可能是bad
  - 排序的位置相关性
    - 排在前面的文档的相关性对整体效果的影响更大

- 对任意两个需要排序的对象A,B，决定其相对顺序A>B还是B>A
- 可以转化为二值分类问题来解决

$$q \leftrightarrow \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \downarrow$$



$$\left\{ \begin{array}{l} (x_1, x_2, +1), (x_2, x_1, -1), \dots, \\ (x_2, x_m, +1), (x_m, x_2, -1) \end{array} \right\}$$



$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \sum_{u,v: y_{u,v}^{(i)}=1} \xi_{u,v}^{(i)}$$

$$w^T (x_u^{(i)} - x_v^{(i)}) \geq 1 - \xi_{u,v}^{(i)}, \text{ if } y_{u,v}^{(i)} = 1.$$

$$\xi_{uv}^{(i)} \geq 0, i = 1, \dots, n.$$

$x_u - x_v$  as positive instance of learning

Use SVM to perform binary classification on these instances, to learn model parameter  $w$

- 多层平面排序(multi-hyperplane ranker)
  - 区分对待不同文档对
    - $(1,2)(1,3),\dots,(4,5)$
  - 为每一种文档对建立分类模型
  - 综合考虑所有分类模型给出的分数进行最终的排序

- Q1:100文档
  - ->~10000文档对
- Q2:10000文档
  - ->~100000000文档对
- 模型更偏向对应文档多的Query

- 为每个query设定一个权重

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \boxed{\frac{1}{\tilde{m}^{(i)}}} \sum_{u,v} \xi$$

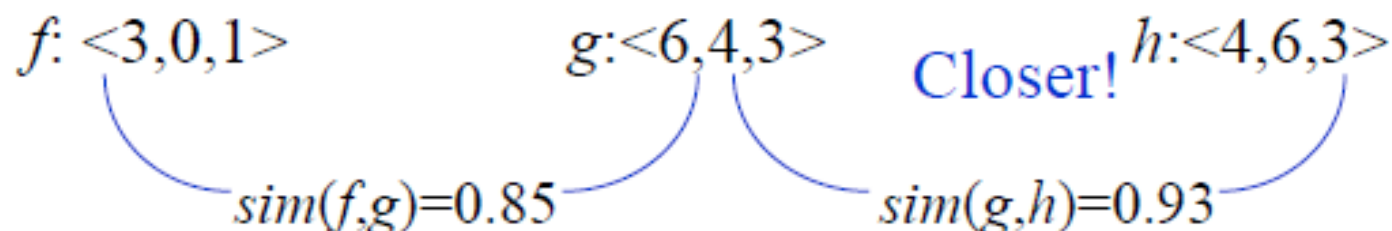
Query-level normalizer

An example:

- function  $f$ :  $f(A)=3, f(B)=0, f(C)=1$  ACB
- function  $h$ :  $h(A)=4, h(B)=6, h(C)=3$  BAC
- ground truth  $g$ :  $g(A)=6, g(B)=4, g(C)=3$  ABC

Question: which function is closer to ground truth?

- Based on pointwise similarity:  $\text{sim}(f,g) < \text{sim}(g,h)$ .
- Based on pairwise similarity:  $\text{sim}(f,g) = \text{sim}(g,h)$
- Based on cosine similarity between score vectors?



However, according to NDCG,  $f$  should be closer to  $g$ !



- 为每一种排序结果给出一个分数，最后采用分数最高的排序

$$P(\pi) = \prod_{i=1}^n \frac{s_{\pi(i)}}{\sum_{j=i}^n s_{\pi(j)}}$$

$$P(ABC) = \frac{s_A}{s_A + s_B + s_C} \cdot \frac{s_B}{s_B + s_C} \cdot \frac{s_C}{s_C}$$

$P(\text{A ranked No.1})$

$P(\text{B ranked No.2} \mid \text{A ranked No.1})$

$P(\text{C ranked No.3} \mid \text{A ranked No.1, B ranked No.2})$

- 损失：用模型输出和人工标注定义的两个分布之间的KL距离衡量

$$L(f; \mathbf{x}, \pi_y) = D(P_y(\pi) \parallel P(\pi \mid \varphi(f(\mathbf{x}))))$$

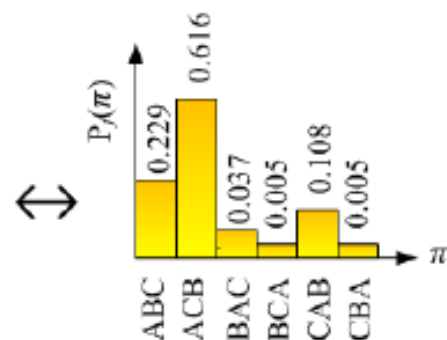
Probability distribution defined  
by the ground truth

Probability distribution defined  
by the model output

- 梯度下降法求解
- 效果最好，但复杂度高

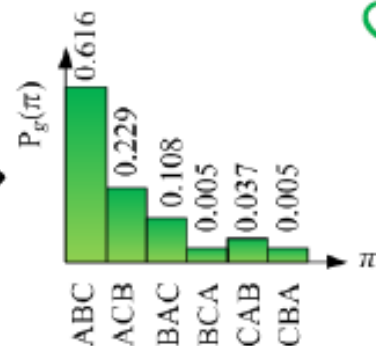
$\varphi = \exp$

$f: f(A) = 3, f(B) = 0, f(C) = 1;$   
Ranking by  $f$ : ABC



Using **KL-divergence**  
to measure difference  
between distributions

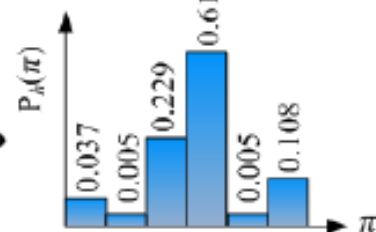
$g: g(A) = 6, g(B) = 4, g(C) = 3;$   
Ranking by  $g$ : ABC



Closer!

$dis(f, g) = 0.46$

$h: h(A) = 4, h(B) = 6, h(C) = 3;$   
Ranking by  $h$ : ACB



$dis(g, h) = 2.56$

- 基于Ranking SVM的网页排序实践
  - 数据集: letor4.0
    - 5-fold交叉验证
  - 46个特征

谢谢!