

机器学习及排序学习基础

武威

微软亚洲研究院自然语言计算组副研究员

wuwei@microsoft.com

2012年11月

提纲

- 什么是排序学习 (Learning to Rank) (L2R)
- 如何学习一个排序
 - 排序学习的两条线
 - 排序生成
 - 排序整合
- 参考资料

你能得到什么

- 为什么要“学习”排序
- 什么是排序学习
- 排序学习的基本流程是什么
- 基本的排序学习算法

- 什么是排序学习
- 如何学习一个排序
 - 排序学习的两条线
 - 排序生成
 - 排序整合
- 参考资料

无处不在的排序



Microsoft Excel - 2010~2011第一学期年三年级期末考试...

文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 工具(T) 数据(D) 窗口(W) 帮助(H)

宋体 10 B I U

P17

2010-2011学年第一学期期末考试学生成绩明细表

准考证号	姓名	班级	地理	历史	数学	英语	语文	政治	总计	班级排名	年级排名
2010080090	刘文浩	三年级三班	86	79	81	85	95	86	512		
2010080053	王毓晨	三年级二班	87	92	73.5	110	87	67	516.5		
2010080064	冯栋	三年级三班	86	78	41.5	82.5	85	83	456		
2010080045	王睿	三年级二班	95	94	65	95.5	89	68	506.5		
2010080115	刘涛	三年级三班	66	96	71	90.5	89	61	473.5		
2010080109	刘茜	三年级三班	62	74	107.5	55	69	55	422.5		
2010080075	田志远	三年级三班	36	64	79	80.5	88	58	405.5		
2010080091	刘长波	三年级三班	63	61	88.5	72	79	65	428.5		
2010080003	丁蒙	三年级一班	64	80	25	100	92	59	420		
2010080101	刘健	三年级三班	56	98	57	82	77	67	437		
2010080004	丁雅琳	三年级一班	62	58	58.5	80.5	83	58	400		
2010080029	王阳阳	三年级一班	42	88	68	63.5	90	63	414.5		
2010080119	刘敬	三年级三班	55	67	40	100	85	60	407		



Web IMAGES VIDEOS MAPS MORE

bing

MS Beta

排序学习

1,850,000 RESULTS

排序学习 - 搜搜百科

baike.soso.com/v6556301.htm - Translate this page

一种比较新的网页排序方法，将机器学习的方法加入到网页排序中，分为三种：点方式，对方式和列表方式。重要的算法有 ...

排序学习——中国科学院自动化研究所

www.ia.cas.cn/jhzc/gjllhcxm/200910/20091015... - Translate this page

项目名称：排序学习 合作机构：微软（中国）有限责任公司 起始时间：2007-12-1 ~ 2008-12-31. 项目负责人：王珏

学习排序 百度文库

wenku.baidu.com/view/584972e9b8f67c1cfad6b8a1.html - Translate this page

学习排序 - 学习排序 - 教学目标 1、通过观察发现图形之间的排列规律，并能按规律接着往下排。 2、理解排序的含义。

排序学习模型_Yode_新浪博客

blog.sina.com.cn/s/blog_4c98b96001000813.html - Translate this page

排序学习旨在为目标对象按照某种规律确定一个顺序，它可以看成是连接回归问题和分类问题的桥梁。排序学习在 ...

基于神经网络的网页排序学习算法研究_百度文库

wenku.baidu.com/view/65af100bf78a6529647d5351.html - Translate this page

基于神经网络的网页排序学习算法研究 - 中山大学 硕士学位论文 基于神经网络的网页排序学习算法研究 姓名：吴桂 ...

Baidu 百度 新闻 网页 贴吧 知道 MP3 图片 视频 地图 文库 更多

排序学习

百度一下

排序学习 - 下载频道 - CSDN.NET

c# 源代码 教程 实例 将好几个排序的算法进行比较。对算法学习非常有帮助上传者：bacteria1 987 上传时间：2010-09-08 下载次数：150 学习 合并重复行 定义新的列为 ...

download.csdn.net/tag/排序学习 2012-9-22 - 百度快照

排序学习 百度文库

排序学习 - 排序学习 李巧兰 学号：学号：1102121363 2012-3-5 一、排序学习的定义 二、排序学习的目的 三、排序学习的分类及特点 四、排序学习 ... 33页 浏览1次

wenku.baidu.com/view/...ab003dc.html 2012-3-6

排序学习.doc 34页 浏览39次

学习排序.doc 2页 浏览91次

座位排序学习.doc 4页 浏览1次

更多文库相关文档>>

排序学习模型_Yode_新浪博客

排序学习旨在为目标对象按照某种规律确定一个顺序，它可以看成是连接回归问题和分类问题的桥梁。排序学习在信息检索中有着非常广泛的应用，在用户提交查询后，搜索引擎 ...

blog.sina.com.cn/s/...1000813.html 2012-8-28 - 百度快照

排序学习 - 搜搜百科

一种比较新的网页排序方法，将机器学习的方法加入到网页排序中，分为三种：点方式，对方式和列表方式。重要的算法有RankNet，Ranking SVM都是比较经典的对方是 ...

baike.soso.com/v6556301.htm 2012-9-16 - 百度快照

排序算法学习小结 - 能巴 - 博客园

排序算法学习小结 参考维基百科 0. 一些基本概念~（体会）空间换时间是进一步提高算法速度的最终方法，比如希尔排序和基数排序就是通过占用额外空间来获得比快速排序 ...

www.cnblogs.com/.../2009225.html 2012-10-5 - 百度快照

Google

排序学习

搜索

找到约 33,300,000 条结果 (用时 0.22 秒)

网页

排序学习 - 搜搜百科

baike.soso.com/v6556301.htm - 网页快照

一种比较新的网页排序方法，将机器学习的方法加入到网页排序中，分为三种：点方式，对方式和列表方式。重要的算法有RankNet，Ranking SVM都是比较经典的对方 ...

视频

排序学习简介_李航博士_新浪博客

blog.sina.com.cn/s/blog_7ad48fee0100ynd0.html - 网页快照

排序学习简介_李航博士_新浪博客_李航博士 ... 排序学习简介. (2011-10-18 14:31:54). 转载▼ 标签：机器学习 信息检索 分类：科研介绍 ...

更多

网页

排序学习——中国科学院自动化研究所

www.ia.cas.cn/jhzc/gjllhcxm/...h20091015_2552455.html - 网页快照

2009年10月15日 - 排序学习 ... 排序学习旨在解决这样的问题，即用户可以得到所输入查询的准确信息。决定学习模型的精度有两个因素：即训练集合的制备以及学习 ...

所有中文网页

简体中文网页

翻译的外文网页

更多搜索工具

用于信息检索的代价敏感排序学习算法研究 - Microsoft Research

research.microsoft.com/en-us/people/.../phdthesis_lean2rank.pdf

文件格式: PDF/Adobe Acrobat

的学习方法融入到排序支持向量机的学习算法中，提出了代价敏感排序学习算 ... 与排序支持向量机相比，本文所提出的代价敏感排序学习算法能够降低发生在 ...

快速排序学习1 - 随感而发 - C++博客

www.cppblog.com/shongbee2/archive/2009/.../08085.html - 网页快照

2009年4月23日 - 今天我学习了快速排序，顾名思义，快速排序的速度是很快的，平均复杂度是nlogn，我也不知道是怎么算出来的，反正T(n) = 2T(n/2) + o(n) 这样怎么 ...

Department
Cell Phones & Accessories
Unlocked Cell Phones
Cases
Protective Skins
Data Cables
No-Contract Cell Phones
Electronics
Electronics Accessories & Supplies
MP3 Player Accessories
+ See All 33 Departments

"iphone 4s"
Related Searches: iPhone 4, iPhone 4s unlocked, iPhone 5.
Showing 1 - 16 of 392,031 Results


iPhone 4 / 4S Anti-Glare, Anti-Scratch, Anti-Fingerprint - Matte Finishing Screen Protector by Generic
\$0.60
In Stock
More Buying Choices
\$0.01 new (124 offers)
\$0.30 used (2 offers)
★★★★☆ (1,017)
Cell Phones & Accessories: See all 174,771 items

Shipping Option (What's this?)
Free Super Saver Shipping

Cell Phones & Accessories
Phones with Plans: AmazonWireless
Mobile Broadband: AmazonWireless
No-Contract Phones
Unlocked Phones
Cases & Covers
Headsets
All Accessories

Listmania!


stuff you need to be like me (for independant girls) By Elisabeth amy: A list by Shelagh Skeen

Apple iPhone 4S 16GB - AT&T - Black
~~\$699.00~~ **\$564.99**
Only 2 left in stock - order soon
More Buying Choices
\$564.99 new (17 offers)
\$369.99 used (73 offers)
★★★★☆ (84)
Electronics: See all 185,679 items


STREET STYPER.COM
iPHONE: A list by anthony german
Create a Listmania! list

Apple iPhone 4S 16GB (White)
\$599.00
Only 16 left in stock - order soon.
More Buying Choices
\$450.00 new (32 offers)
\$380.00 used (106 offers)
★★★★☆ (33)
Trade in this item for an Amazon.com Gift Card
Cell Phones & Accessories: See all 174,771 items

Search Listmania!


pur

Apple iPhone 4S 16GB Black - FACTORY UNLOCKED by Apple Computer
~~\$649.00~~ **\$598.88**
Order in the next **57 hours** and get it by Tuesday, Oct 16
Only 1 left in stock - order soon.
More Buying Choices
\$650.00 new (56 offers)
\$418.99 used (54 offers)
★★★★☆ (28)
Eligible for FREE Super Saver Shipping.
Cell Phones & Accessories: See all 174,771 items

找到相关产品 1 款 (宝贝 9544 件), [查看所有宝贝](#)

[所有分类](#) [手机](#) [Apple/苹果 iPhone4S](#)

Apple/苹果 iPhone4S

 [查看全部大图](#)

参考价: 约 **4268.00元**  周销量 **12337**件 类目排名3位

网络类型: CDMA2000(3G) 操作系统: iPhone
主屏尺寸: 3.5英寸 主屏分辨率: 960×640像素
手机CPU: 双核1G 运行内存RAM: 512M
厚度: 超薄(小于9mm) 上市时间: 2011年

购买过的人说 [更多](#)
 温良大薯
唯一没弄明白的是插美国卡能否在美国用。其实美版有锁的反而好,假如有一半时间在美国的话。
[查看宝贝](#)

热卖商家 **产品详情**

所有宝贝 **天猫** **二手** **我的搜索** 1/100

本分类下搜索 ☐ 假一赔三 ☐ 电器城 ☐ 正品保障 ☐ 消费者保障 ☐ 7天退换 ☐ 折扣促销 ☐ 旺旺在线

默认 ☒ 人气 ☐ 销量 ☐ 信用 ☐ 价格 ☐ 合并卖家

iPhone 4S **Apple/苹果 iPhone4S 16G 32G 64G 4S** **¥ 4360.00** 上海 最近79人成交81笔  七天退换
运费: 22.00 [43条评论](#)
行货另有电信版16G 64G
天猫 TMALL.COM
苏宁易购专营店 [和我联系](#) 

【广东联通旗舰店】Apple/苹果 iPhone 4S 16G 手机 智能 0元购机 **¥ 4488.00** 广东 广州 最近9人成交10笔  七天退换
运费: 0.00 [765条评论](#)
天猫 TMALL.COM
广东联通官方旗舰店 [和我联系](#) 

Apple/苹果 iPhone 4S 原装正品 未激活 未拆封 手机5 无锁 **¥ 3700.80** 广东 深圳 2067人成交 | 9992条评论  消费者保障
运费: 20.00 
马兴通 [和我联系](#) 
原装(2424人) 七天退换
包装很到位(231人)

返200红包 Apple/苹果 iPhone4S iPhone 4s 联通版 16G 32SG 64G **¥ 4148.00** 上海 最近1616人成交1652笔  七天退换
运费: 22.00 [6053条评论](#)
天猫 TMALL.COM
彪彪数码专营店 [和我联系](#) 

什么是排序学习

搜索答案

我要提问

我要回答



百度知道 > 搜索结果

全部回答

待完善问题

排序: 相关性 | 最新提问

[关于生态学问题: 什么是“排序”, 直接排序...](#)

及环境特征是什么生物群落的分类与排序 4. 《中国了解景观生态学要解决的关键问题, 它与个体、种群、群落和生态系统生态学之间 ...能问的具体些吗 ...

溱海屋市 - 2011-12-06 21:10 - 最佳答案者: 待定 - [教育/科学](#) > [理工学科](#) > [生态学](#)

[土木工程的学习应该注重哪些方面? 着重程度...](#)

要看你主要想从事什么和你的一些大致的想法了比如你毕业了直接工作, 可能工程上的一些东西你就要多学学, 比如概预算啊工程管理之类的如果你打算考研, 那么一些...

匿名 - 2011-05-12 12:20 - 最佳答案者: [wang508710348](#) - [电脑/网络](#) > [程序设计](#)

[小学英语\(听音, 排序\)是什么意思](#)

几组图或句子放在一起, 他报单词或含这个单词的句子, 然后你按那个顺序标号, 就好了。望楼主采纳。。。 ...就是说让你听那道题的录音, 然后根据听到内容...

[1161487633](#) - 2011-09-17 20:14 - 最佳答案者: [连旧](#) - [教育/科学](#) > [学习帮助](#)

[数据结构课到底学的是什?](#)

数据结构还会涉及到数据结构的一些应用比如查找、排序、文件等。数据结构是学习...什么是数据结构 1.2 基本概念和术语 1.3 抽象数据类型的表现与实现 1.4...

[734720752](#) - 2010-05-23 17:43 - 最佳答案者: [L121000](#) - [电脑/网络](#) > [程序设计](#) > [其他编程语言](#)

[星战前夜eve技能怎么排序自动学习](#)

LS说的很对。CCP将会在虫洞或者疆域开放技能训练序列, 可以安排24小时内的...那是下个虫洞版本的了, 现在你就等等吧 ...虫洞版本有 ...学习技能的...

[tab198548](#) - 2010-05-30 15:15 - 最佳答案者: [ak47degame](#) - [游戏](#) > [网络游戏](#)

Search

what is learning to rank

Search Y! Answers

Sort by: [Relevance](#) | [Newest](#) | [Most Answers](#)



TIPS for my suggestion guide for AF Basic?

...." then Sir with **what** you have **to** say for example. Sir Trainee Whoever...leave out the first Sir otherwise you **are** calling the T.I. A trainee... your way. You can **learn ranks** it will also help you some in...
☆ In Military - Asked by northgwinnett02 - 1 answer - 3 years ago



Tell me about **what to** except at Navy bootcamp?

...officer chief. **Was** it easy **to learn ranks**??? I don't no I tend **to** over think things and when they actually happen they **are** not as bad as I thought...of the day I just want **to do what is** asked of ...
☆ In Military - Asked by Anthony - 2 answers - 11 months ago



What would Cronkite think of the current crop of news readers?

... his way up through the **ranks, learning** as he went, keeping his bias **to** himself, old-school journalism. **What** did he think of over...Maybe the same disdain JFK, who **was** a WWII hero and spent...
1☆ In Politics - Asked by Dcntamcn - 11 answers - 3 years ago



What's the deal?

...would just have **to** take some sort of class **to learn rank** structure and other stuff like that. I...39;t give the details. I **was** hoping some of you could maybe...
☆ In Military - Asked by ? - 10 answers - 6 years ago



Any tips on preparing for air force boot camp?

Like should I **learn all ranks. Learn** any workouts they do besides.... And anything else I need **to** know. And how hard **is it** to take apart and put back together...year so I have time jus not sure **what** all I need **to** know.
1☆ In Military - Asked by Marc - 4 answers - 1 year ago

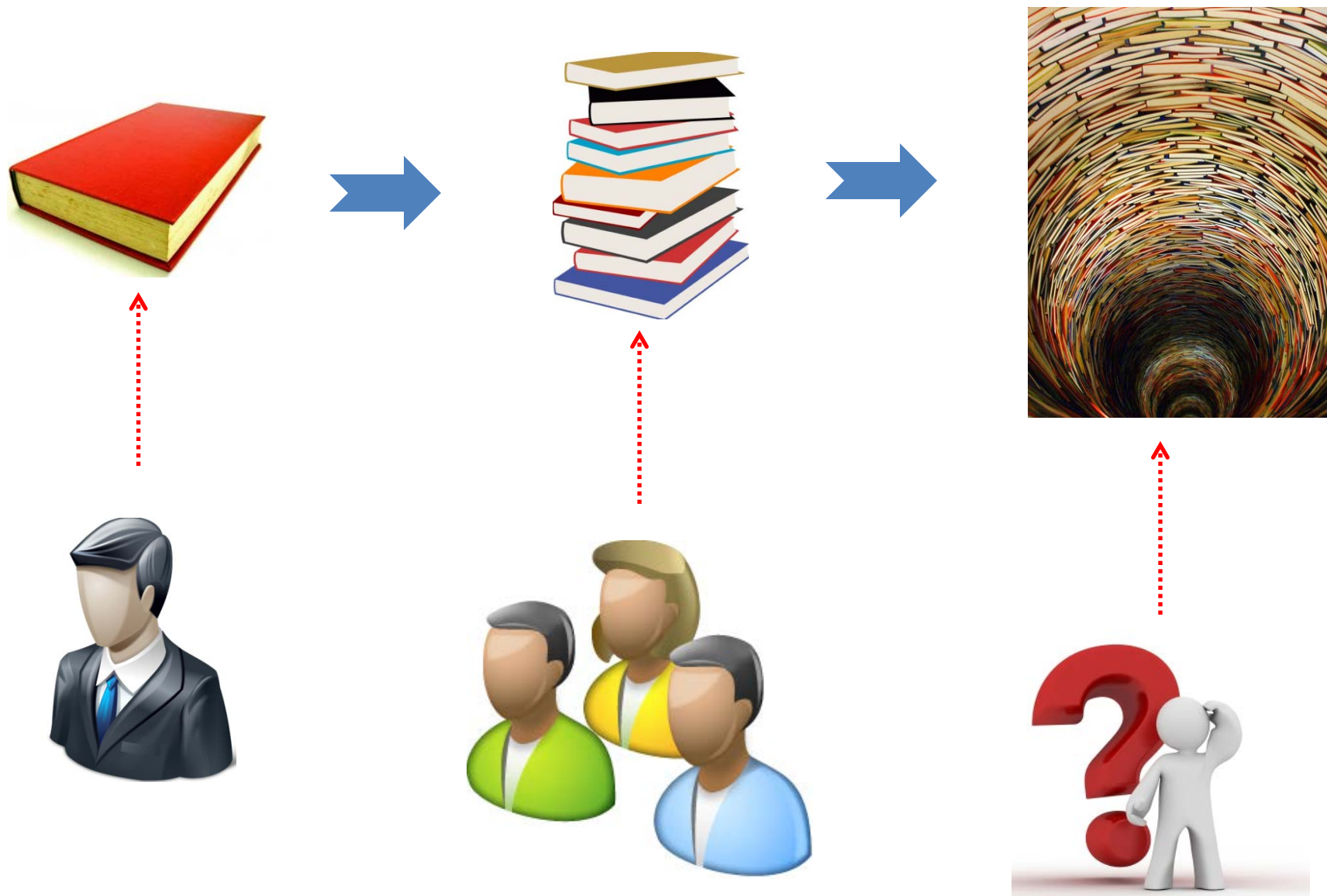


Are You Looking For A Psn Black Ops Clan?

...your skill and shape it properly. **What** more could you want? The setup of our clan **is** very simple and straight forward. It gives easy **to learn ranks** and leadership positions...
☆ In Video & Online Games - Asked by Chris - 1 answer - 1 year ago

其他例子

- 统计机器翻译 (Machine Translation)
- 图片搜索 (Image Search)
- 专家搜索 (Expert Search)
- 推荐系统 (Recommender System)
-



近年排序学习的发展趋势

- 排序学习已经被成功地应用到了实际网络搜索中
 - Google
 - Bing
 - 百度
 - 有道
 -
- 在SIGIR, ICML, NIPS等信息检索、机器学习的国际会议上有超过100篇的文章
- 每年信息检索的顶级会议SIGIR上有两个session
- Yahoo Learning to Rank Challenge
- 排序学习的标准数据集 LETOR
 - <http://research.microsoft.com/en-us/um/beijing/projects/letor/>



机器学习基础



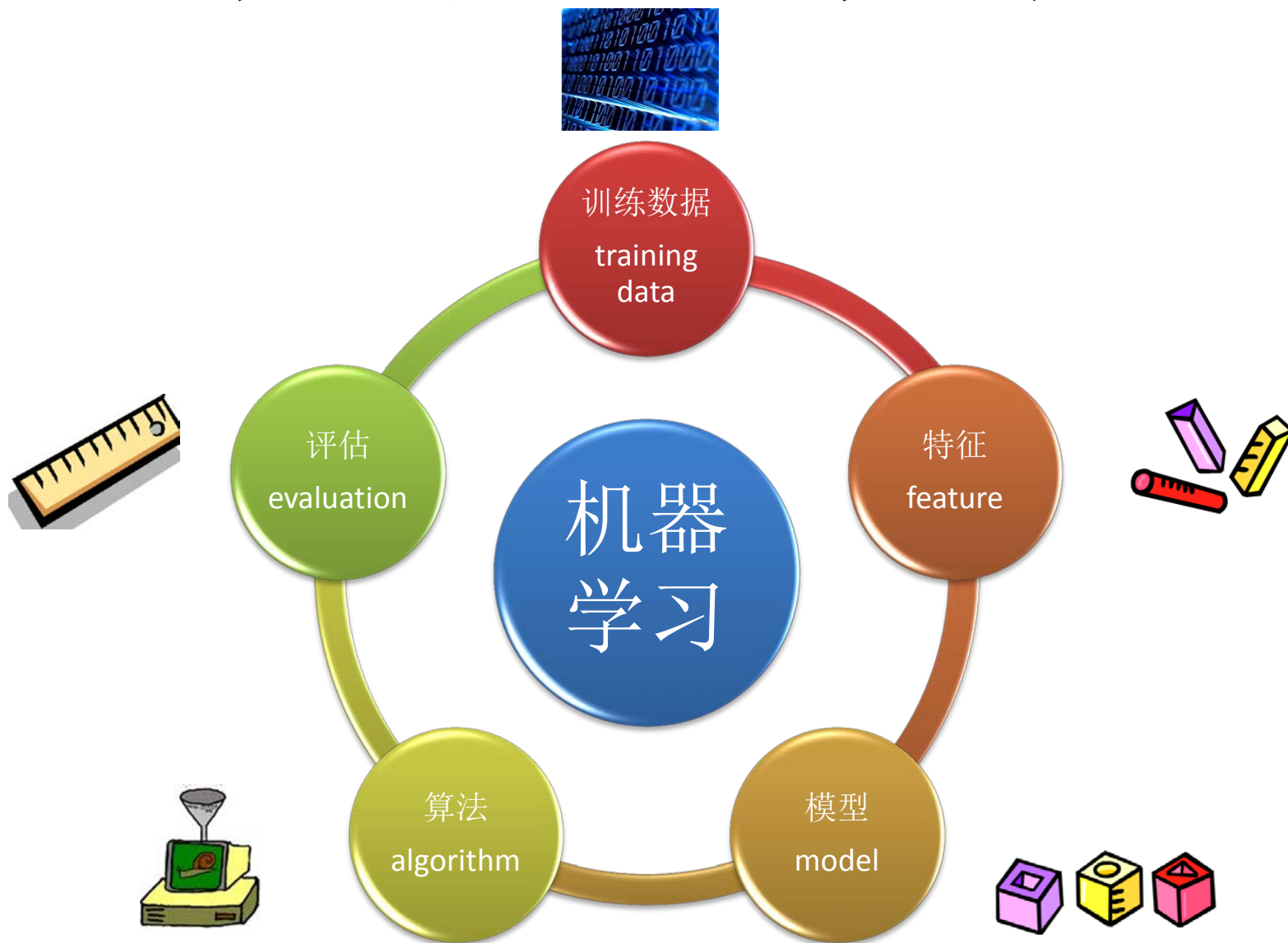
机器学习的定义

- *A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P** , if its performance at tasks in T , as measured by P , **improves with experience E***

---- by Tom Mitchell in *Machine Learning*

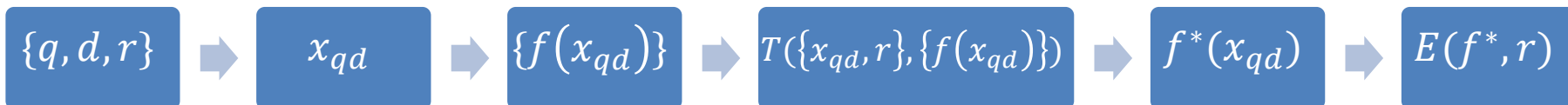
- 排序学习
 - 任务：对一组文档进行排序
 - 评估指标：相关文档应该排在前面
 - 经验：人工标注
 - 目标：利用人工标注设计算法以抓住隐藏在数据中的规律从而实现对任意查询请求给出反映相关性的文档排序

机器学习的基本要素



机器学习的基本步骤

训练数据生成 → 特征抽取 → 模型构建 → 算法设计 → 模型选择 → 效果评估



机器学习分类

- 监督学习 (Supervised Learning)

- 标注数据 → 模型



- 半监督学习 (Semi-supervised Learning)

- 标注数据+未标注数据 → 模型



- 无监督学习 (Unsupervised Learning)

- 未标注数据 → 模型



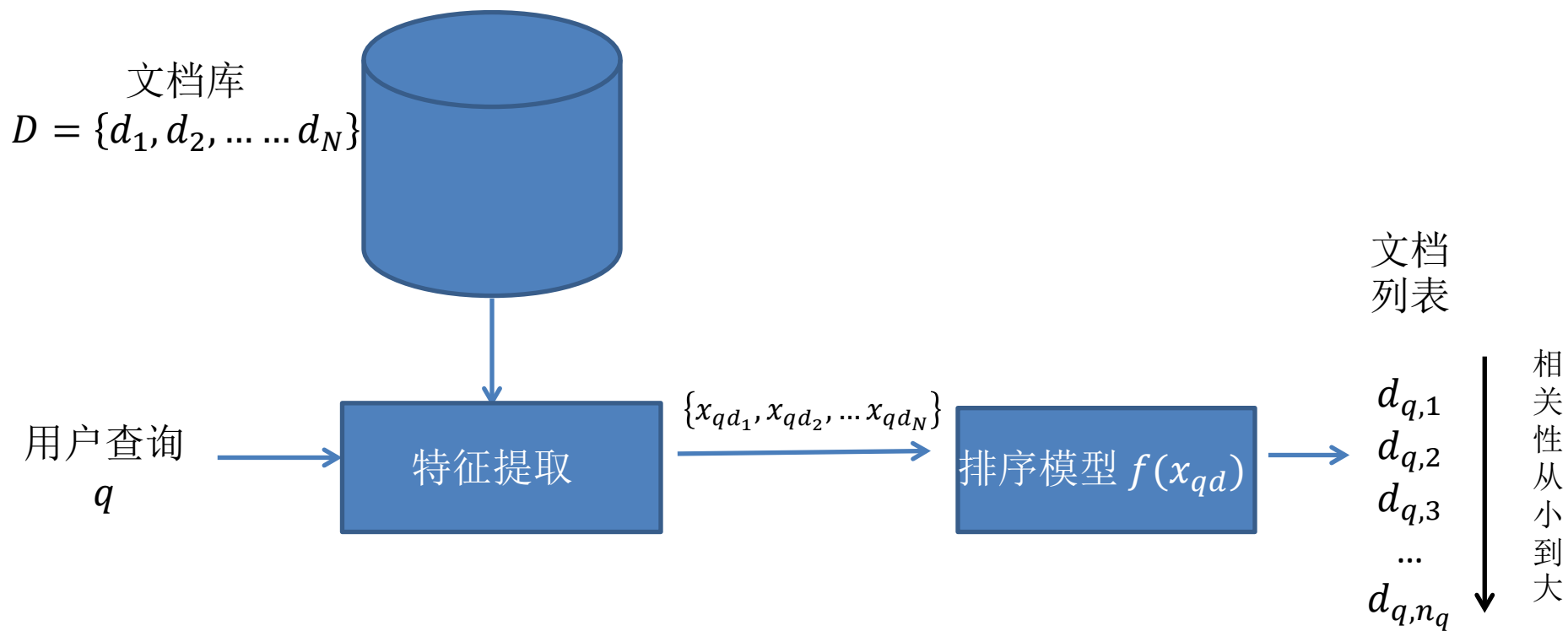
什么是排序学习

排序学习的定义

- 广义
 - 排序学习指**机器学习**中任何用于**排序**的技术
- 狭义
 - 排序学习指在**排序生成** (ranking creation) 和**排序整合** (ranking aggregation) 中用于构建**排序模型**的**机器学习方法**

[Hang Li, Learning to Rank for Information Retrieval and Natural Language Processing]

排序学习



- 什么是排序学习
- 如何学习一个排序
 - 排序学习的两条线
 - 排序生成
 - 排序整合
- 参考资料

- 什么是排序学习
- 如何学习一个排序
 - 排序学习的两条线
 - 排序生成
 - 排序整合
- 参考资料

横线

- 排序生成

- 输入：查询文档对以及标注数据
- 输出：一个排序模型，可以为新查询排列文档

- 排序整合

- 输入：一组排序列表和标注数据
- 输出：一个排序模型，可以整合不同的排序列表

排序生成

Learning to Rank

[Learning to rank - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Learning_to_rank ▾ [翻译此页](#)

[Applications](#) · [Feature vectors](#) · [Evaluation measures](#) · [Approaches](#) · [History](#)

Learning to rank or **machine-learned ranking** (MLR) is a type of supervised or semi-supervised machine **learning** problem in which the goal is to automatically ...

[Yahoo! Learning to Rank Challenge](#)

learningtorankchallenge.yahoo.com ▾ [翻译此页](#)

Learning to Rank Challenge is closed! Close competition, innovative ideas, and fierce determination were some of the highlights of the first ever Yahoo!

[Search Engine Optimization Tips, Learn SEO, SEO Tools, Guide...](#)

searchandrunk.com ▾ [翻译此页](#)

Welcome to SearchandRank.com where you can **Learn** Search Engine Optimization SEO and on how **to rank** on Google, Yahoo and Bing. From what Meta-Descriptions, ...

[Learning to rank - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Learning_to_rank ▾ [翻译此页](#)

[Applications](#) · [Feature vectors](#) · [Evaluation measures](#) · [Approaches](#) · [History](#)

Learning to rank or **machine-learned ranking** (MLR) is a type of supervised or semi-supervised machine **learning** problem in which the goal is to automatically ...

[Yahoo! Learning to Rank Challenge](#)

learningtorankchallenge.yahoo.com ▾ [翻译此页](#)

Learning to Rank Challenge is closed! Close competition, innovative ideas, and fierce determination were some of the highlights of the first ever Yahoo!

[Search Engine Optimization Tips, Learn SEO, SEO Tools, Guide...](#)

searchandrunk.com ▾ [翻译此页](#)

Welcome to SearchandRank.com where you can **Learn** Search Engine Optimization SEO and on how **to rank** on Google, Yahoo and Bing. From what Meta-Descriptions, ...

[Learning to rank - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Learning_to_rank ▾ [翻译此页](#)

[Applications](#) · [Feature vectors](#) · [Evaluation measures](#) · [Approaches](#) · [History](#)

Learning to rank or **machine-learned ranking** (MLR) is a type of supervised or semi-supervised machine **learning** problem in which the goal is to automatically ...

[Search Engine Optimization Tips, Learn SEO, SEO Tools, Guide...](#)

searchandrunk.com ▾ [翻译此页](#)

Welcome to SearchandRank.com where you can **Learn** Search Engine Optimization SEO and on how **to rank** on Google, Yahoo and Bing. From what Meta-Descriptions, ...

[Yahoo! Learning to Rank Challenge](#)

learningtorankchallenge.yahoo.com ▾ [翻译此页](#)

Learning to Rank Challenge is closed! Close competition, innovative ideas, and fierce determination were some of the highlights of the first ever Yahoo!

[Search Engine Optimization Tips, Learn SEO, SEO Tools, Guide...](#)

searchandrunk.com ▾ [翻译此页](#)

Welcome to SearchandRank.com where you can **Learn** Search Engine Optimization SEO and on how **to rank** on Google, Yahoo and Bing. From what Meta-Descriptions, ...

[Yahoo! Learning to Rank Challenge](#)

learningtorankchallenge.yahoo.com ▾ [翻译此页](#)

Learning to Rank Challenge is closed! Close competition, innovative ideas, and fierce determination were some of the highlights of the first ever Yahoo!

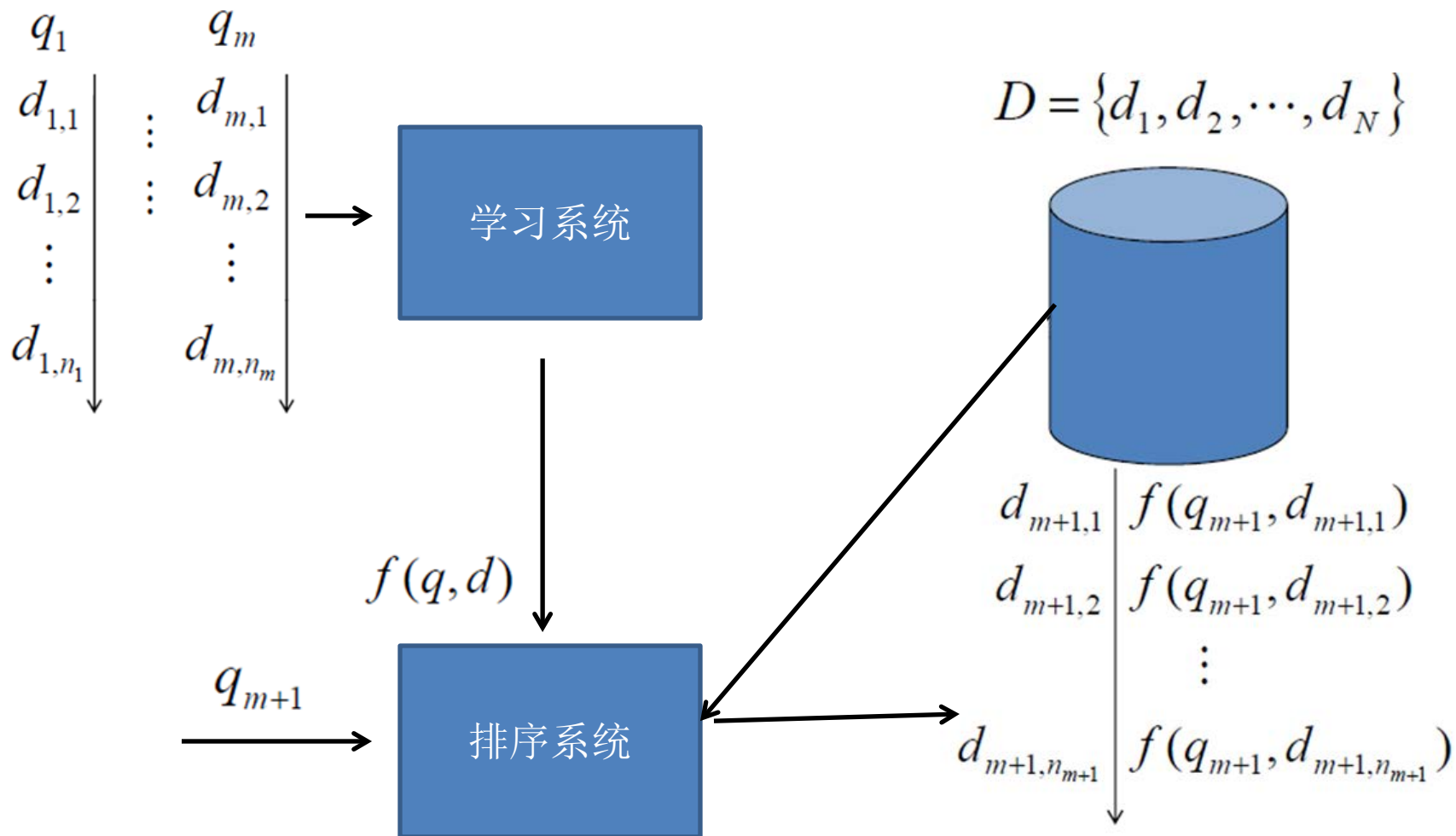
[Learning to rank - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Learning_to_rank ▾ [翻译此页](#)

[Applications](#) · [Feature vectors](#) · [Evaluation measures](#) · [Approaches](#) · [History](#)

Learning to rank or **machine-learned ranking** (MLR) is a type of supervised or semi-supervised machine **learning** problem in which the goal is to automatically ...





排序整合



[Learning to Rank 小结 - Searcher's Log](#)

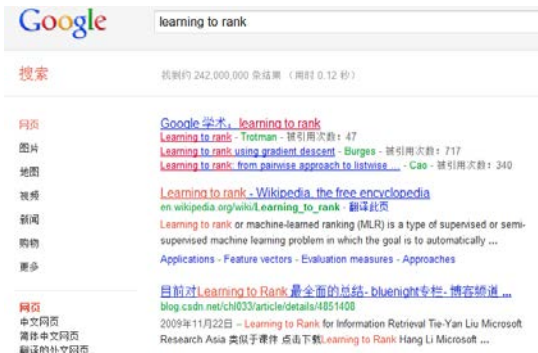
于是 Learning to Rank (LTR) 就被引入了进来。LTR 的核心是想用机器学习来解决排序的问题。目前被广泛运用在 信息检索 (IR)、自然语言处理 (NLP) 和数据挖掘 (Data Mining) 等领域。
[blog.crackcell.com/...o_2_ltr.html 2012-8-14 - 百度快照](#)

[learning to rank - bluenight 专栏 - 博客频道 - CSDN.NET](#)

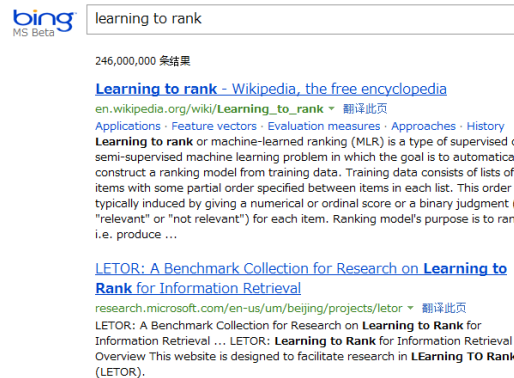
“Yahoo 发起了一项学习排序竞赛 (Learning to Rank Challenge) 作为 CML 2010 大会的一部分。任何人可以以个人名义或组队 (最多 10 人) 参赛。竞赛 3 月 1 日开始，至 5 月 31 日结束。”
[blog.csdn.net/ch033/.../5364615 2012-8-12 - 百度快照](#)

[Learning to rank - Wikipedia, the free encyclopedia](#)

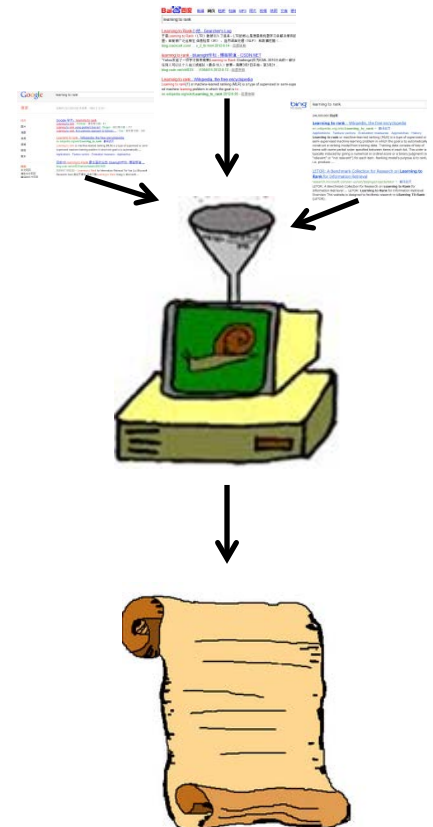
Learning to rank [1] or machine-learned ranking (MLR) is a type of supervised or semi-supervised machine learning problem in which the goal is to ...
[en.wikipedia.org/wiki/Learning_to_rank 2012-8-28 - 百度快照](#)



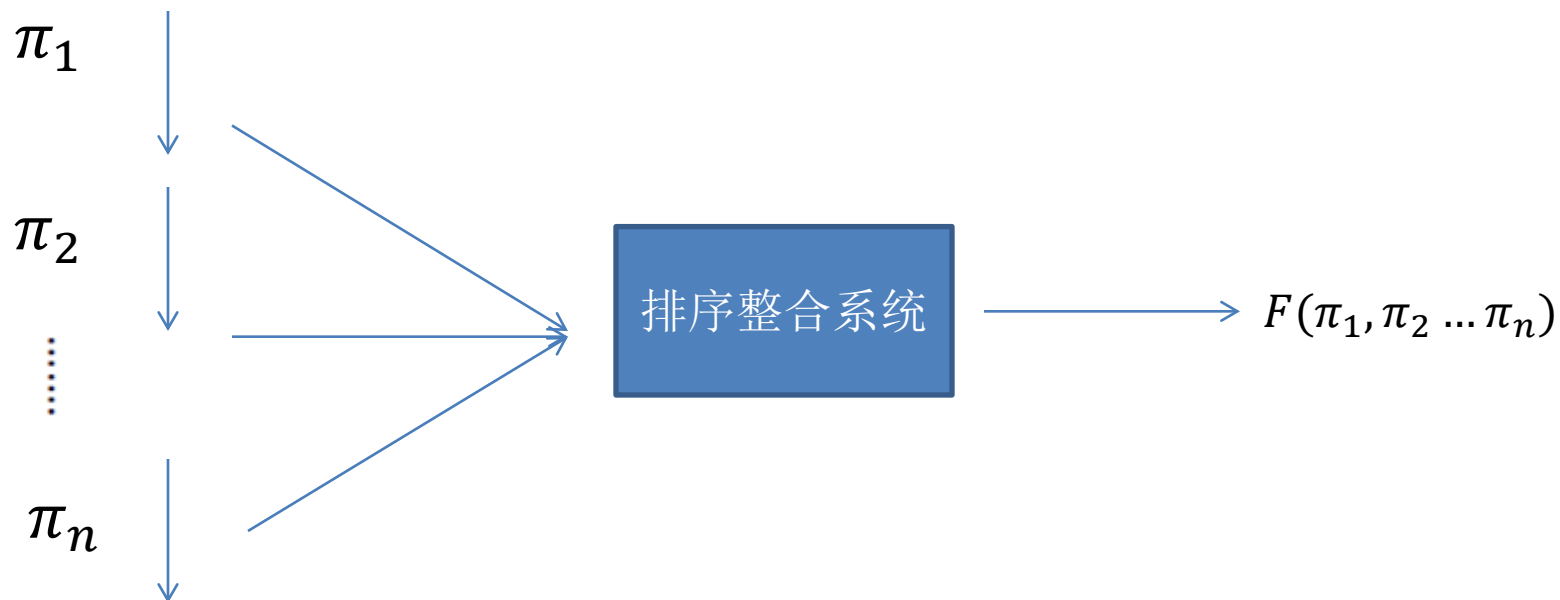
Learning to Rank



听谁的？



排序整合（续）



纵线

- 训练数据怎么来?
 - 单点标注 (pointwise)
 - 两两标注 (pairwise)
 - 列表标注 (listwise)
- 模型如何学习?
 - 监督学习 (supervised learning)*
 - 无监督学习 (unsupervised learning)
 - 多作为特征使用
 - 半监督学习 (semi-supervised learning)
- 如何评估学习效果?
 - Mean Average Precision (MAP) *
 - Normalized Discount Cumulative Gain (NDCG)*
 - Mean Reciprocal Rank (MRR)
 - Winners Take All (WTA)
 - Kendall's Tau
 -

- 什么是排序学习
- 如何学习一个排序
 - 排序学习的两条线
 - 排序生成
 - 排序整合
- 参考资料

标注数据生成 (Data Labeling)

- 显式标注
 - 对每一个查询，人工检查文档的相关性
 - 代价高（钱+时间）
 - 噪声大
- 隐式标注
 - 从用户点击记录(query log)里抽取数据标注
 - 用户总是习惯于从上到下浏览搜索结果
 - 如果用户跳过了排在前面的文档而点击浏览了排在后面的文档，那么排在后面的文档就比排在前面的文档更相关
 - 用户点击有比较大的噪声
 - 只有头查询(head query)才存在用户点击

人工标注 (Human Labeling)

- 训练+评估
- 单点标注
 - 为每个查询文档对打上绝对标签
 - 二元标注：相关 v.s. 不相关 $\{1,0\}$
 - 五级标注：完美(Perfect), 出色(Excellent), 好(Good), 一般(Fair), 差(Bad)
 - 一般+差=不相关
 - 好处：工作量相对较小($O(n)$)
 - 坏处：定多少级合适？不同的人对相同的级，相同的文档能否有相同的理解？---> 噪声大，一致性差
- 两两标注
 - 对于一个查询 q ，文档 d_1 是否比文档 d_2 更相关 $(q, d_1) >? (q, d_2)$
 - 好处：不同的标注人员容易达成一致
 - 坏处：工作量大，($O(n^2)$)，最好 ($O(n \log n)$)
- 列表标注
 - 给定一个查询，为所有文档给出标准排序
 - 不常见

特征提取 (Feature Extraction)

- 如何表示一对查询文档?
 - 查询文档中共同出现的词的个数
 - 点击次数
 - BM25
 - [S.E.Robertson & S.Walker Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval]
 - PageRank
 - [L.Page et al., The pagerank citation ranking: Bringing order to the web]
 - 编辑距离 (edit distance)
 - [Tao Tao and Chengxiang Zhai. An exploration of proximity measures in information retrieval]
 - 网页质量
 - [Bendersky et al. Quality-biased ranking of web documents]
 -

BM25

$$\text{BM25}(q, d) = \sum_{w \in q \cap d} \text{idf}(w) \frac{(k+1)\text{tf}(w)}{\text{tf}(w) + k((1-b) + b \frac{dl}{\text{avgdl}})}$$

Diagram illustrating the BM25 formula with annotations:

- 词** (Term): Points to the summation index w .
- 逆文档频率** (Inverse Document Frequency): Points to $\text{idf}(w)$.
- 词频** (Term Frequency): Points to $\text{tf}(w)$.
- 文档长度** (Document Length): Points to dl .
- 平均文档长度** (Average Document Length): Points to avgdl .

$$\text{idf}(w) = \log\left(\frac{\#\{d\}}{\#\{d | t \in d\}}\right)$$

PageRank

度量文档的重要程度

一个文档的重要程度由链向它的文档的重要程度决定

互联网图上的随机游动

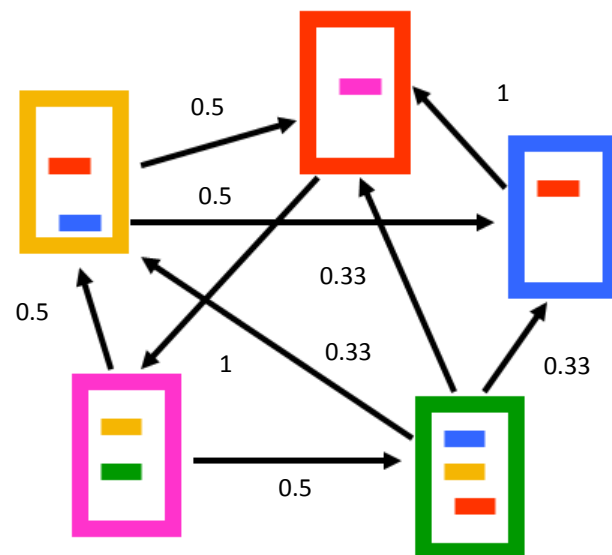
文档得分，停留概率

$$P(d) = \alpha \sum_{d_i \in M(d)} \frac{P(d_i)}{L(d_i)} + (1 - \alpha) \frac{1}{N}$$

链向 d 的文档集合

文档 d_i 的出度

随机跳转




评估准则 (Evaluation Measure)

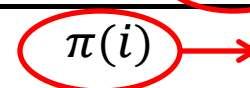
- MAP (Mean Average Precision)

- 给定一个排序，相关文档排得越靠前，排序结果越好
- 计算排在相关文档前面的相关文档比例

- $$P(i) = \frac{\sum_{\pi(k) \leq \pi(i)} y_{q,k}}{\pi(i)}$$



1 相关
0 不相关



文档*i*的排序

- $$AP(q) = \frac{\sum_{i=1}^{n_q} P(i) y_{q,i}}{\sum_{i=1}^{n_q} y_{q,i}},$$

- $$MAP = \frac{\sum_q AP(q)}{\#q}$$

- (1,0,1,1,0,0)

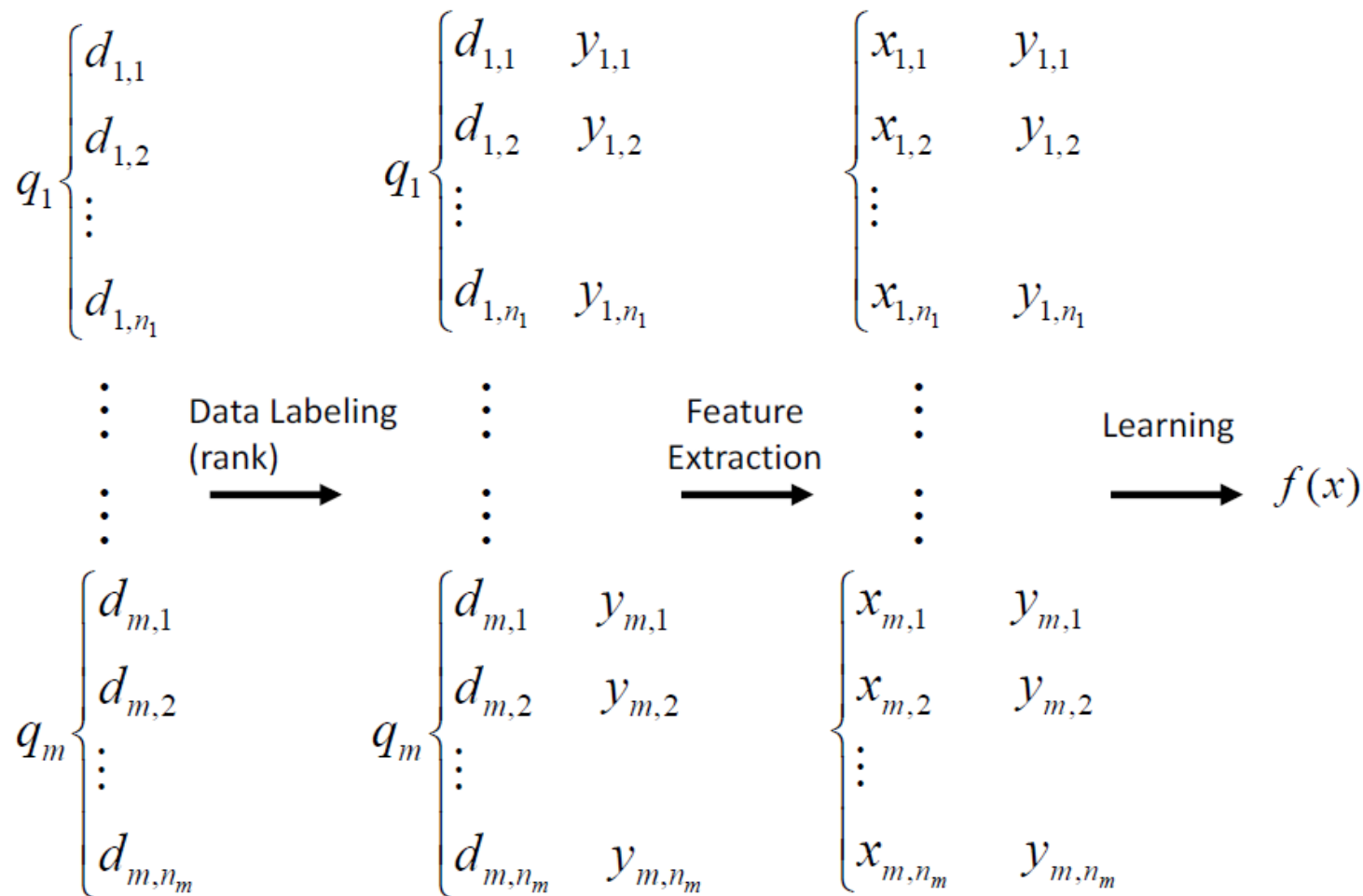
- $P(i) = (1, -, 0.67, -, 0.75, -, -)$

- $AP = (1 + 0.67 + 0.75) / 3 = 0.81$

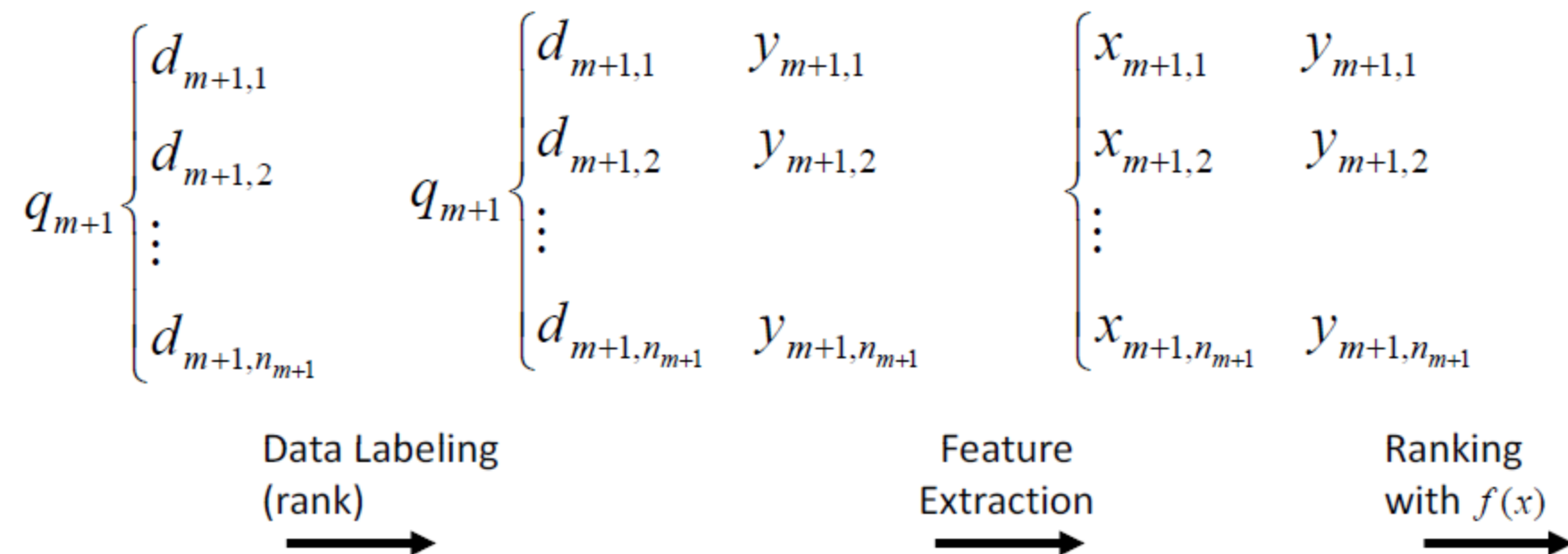
评估准则（续）

- NDCG (Normalized Discount Cumulative Gain)
 - 当标注数据有多级的时候，级高的文档应该排在前面
 - Perfect>Excellent>Good>Fair>Bad
 - $DCG = \sum_{i=1}^n \frac{2^{y_{q,i}} - 1}{\log(\pi(i) + 1)}$
 - $NDCG = \frac{DCG}{DCG_{max}}$
 - (2,3,2,3,1,1)
 - $2^{y_{q,i}} - 1 = (3, 7, 3, 7, 1, 1)$
 - $1/\log(\pi(i) + 1) = (1, 0.63, 0.5, 0.43, 0.39, 0.36)$
 - DCG=12.67
 - DCG_max=14.95
 - NDCG=12.67/14.95=0.85

训练



预测



排序模型

- 基于数据点的方法 (pointwise)
 - 输入
 - 单个查询文档对: $(x_{1,1}, y_{1,1}), (x_{1,2}, y_{1,2}) \dots \dots (x_{m,n_m}, y_{m,n_m})$
 - 完全忽略相同查询下文档间的关系
 - 标注(label) 转化成数字
 - Perfect->5, Excellent->4, Good->3, Fair->2, Bad->1
 - 输出
 - 排序函数 $f(x)$, 对于给定查询文档对, 能够计算出得分(score)
 - 代表模型
 - subset ranking
 - [David Cossock and Tong Zhang, Subset ranking using regression]
 - McRank
 - [Ping Li, Christopher Burges, and Qiang Wu. Macrank: Learning to rank using multiple classification and gradient boosting]
 - Prank
 - [Koby Crammer and Yoram Singer. Pranking with ranking]

排序模型（续）

- 基于数据对的方法 (pairwise)

- 输入

- 同一查询下的一对文档 $(x_{1,1}, x_{1,2}, y_{1,2}^1), \dots, (x_{m,n_m-1}, x_{m,n_m}, y_{n_m-1,n_m}^m)$
 - 标注是两个文档的相对关系，如果文档 $x_{i,j}$ 比文档 $x_{i,k}$ 更相关，那么 $y_{ijik}^i = 1$
 - 部分保留了同一查询下文档间的关系

- 输出

- 排序函数 $f(x)$ ，对于给定查询文档对，能够计算出得分(score)

- 代表模型

- **Ranking SVM**

- [Ralf Herbrich, Thore Graepel, and Klaus Obermayer, Large margin rank boundaries for ordinal regression]

- RankBoost

- [Yoav Freund, Raj D. Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences]

- RankNet

- [Chris Burges, et al. Learning to rank using gradient descent]

排序模型（续2）

- 基于列表的方法 (Listwise)

- 输入

- 一个查询下的整个列表

$$\begin{cases} x_{1,1} & y_{1,1} \\ x_{1,2} & y_{1,2} \\ \vdots & \\ x_{1,n_1} & y_{1,n_1} \end{cases}$$

- 输出

- 排序函数 $f(x)$ ，对于给定查询文档对，能够计算出得分(score)

- 代表模型

- Lambda Rank

- [Chris Burges et al. Learning to rank with nonsmooth cost functions]

- ListNet

- [Zhe Cao et al. Learning to rank: from pairwise approach to listwise approach]

- ListMLE

- [Fen Xia et al. Listwise approach to learning to rank: theory and algorithm]

- AdaRank

- [Jun Xu and Hang Li, Adarank: a boosting algorithm for information retrieval]

- SVMMap

- [Yisong Yue et al. A support vector method for optimizing average precision]

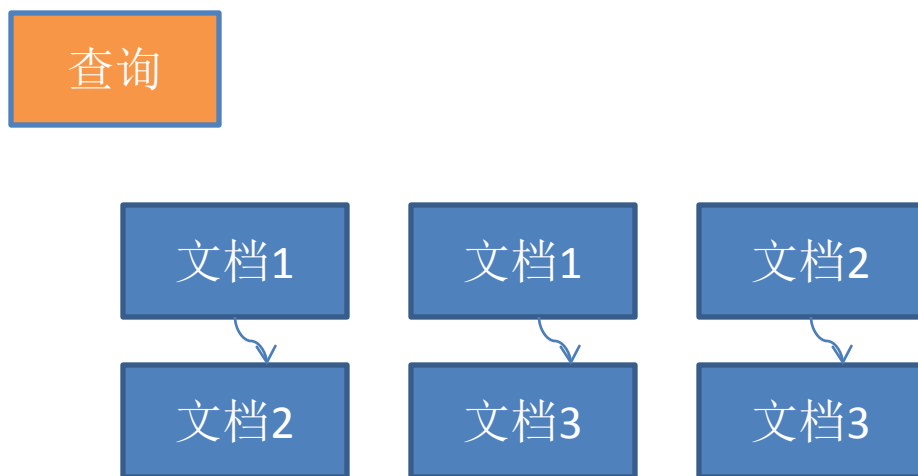
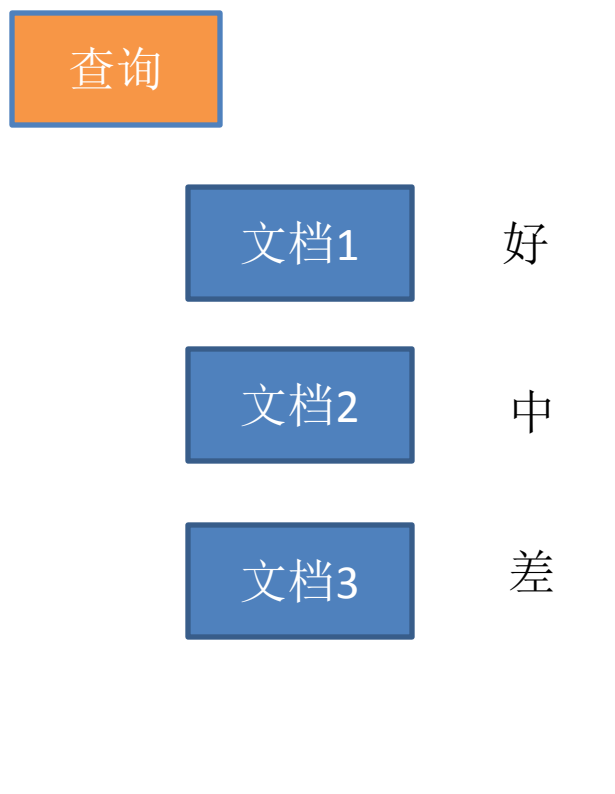
Pointwise v.s. Pairwise v.s. Listwise

	pointwise	pairwise	listwise
信息完全度	不完全	部分完全	完全
输入	(x, y)	(x_1, x_2, y)	$(x_1, x_2, \dots \dots x_n, \pi)$
输出	$f(x)$	$f(x)$	$f(x)$
样本复杂度	$\mathcal{O}(n)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n!)$
表现	差	中	好

Ranking SVM

输入

- 对于每一个查询，将文档列表转换成文档对



算法设计

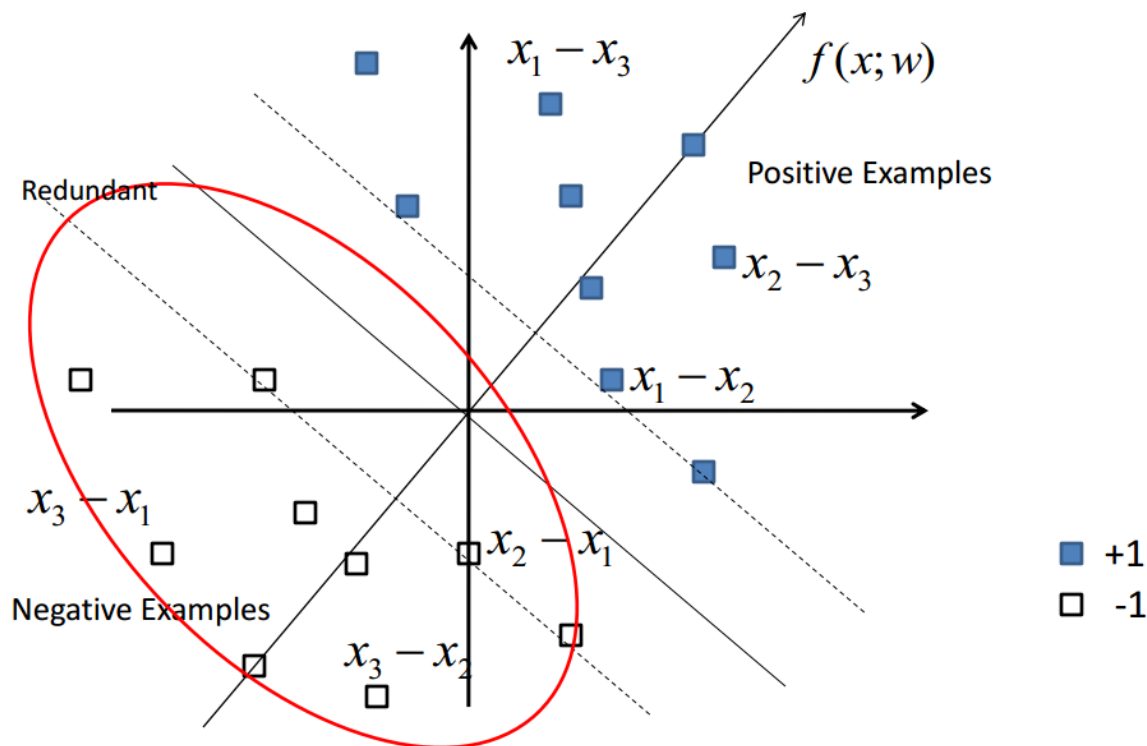
- 目标：学习一个（线性）排序函数 $f(x) = w \cdot x$
- 直观：如果 x_i 比 x_j 更相关，那么 $f(x_i) > f(x_j)$
- 数学表示： $\langle w, x_i - x_j \rangle > 0$
- 转化成分类问题
 - $(x_i - x_j, y)$ $x_i > x_j, y = 1$, 否则 $y = -1$

利用SVM求解排序问题

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

$$y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \geq 1 - \xi_i \quad i = 1, \dots, N$$

$$\xi_i \geq 0$$



工具

- *SVMLight*
 - <http://svmlight.joachims.org/>

- 什么是排序学习
- 如何学习一个排序
 - 排序学习的两条线
 - 排序生成
 - 排序整合
- 参考资料

代表模型

- 无监督
 - Borda Count
 - [Javed A. Aslam and Mark Montague, Models for metasearch]
 - Markov Chain
 - [Cynthia Dwork et al. Rank aggregation methods for the web]
- 监督
 - Cranking
 - [Guy Lebanon and John Lafferty, Cranking: combining rankings using conditional probability models on permutations]

Borda Count

- 给定一组排序列表，计算排在每个文档后面文档的个数和

基本排序 $\pi_1, \pi_2, \dots \dots \pi_n$

计算文档 x 在
 π_i 上的得分 $S_i(x) = \#\{\pi_i(k) > \pi_i(x)\}$

排序函数 $f(x) = \sum_{i=1}^n S_i(x)$

Borda Count (续)

- 基本排序
 - $(A,B,C), (A,C,B), (B,A,C)$
- 计算每个排序上的得分
 - $(S_1(A), S_1(B), S_1(C)) = (2,1,0)$
 - $(S_2(A), S_2(B), S_2(C)) = (2,0,1)$
 - $(S_3(A), S_3(B), S_3(C)) = (1,2,0)$
- 计算排序函数值
 - $f(A) = 2+2+1=5$
 - $f(B) = 1+0+2=3$
 - $f(C) = 0+1+0=1$

- 什么是排序学习
- 如何学习一个排序
 - 排序学习的两条线
 - 排序生成
 - 排序整合
- 参考资料

参考文献

- Hang Li, Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies*, 4(1):1-113, 2011
- Tieyan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*. 3(3):225-331, 2009

谢谢