SACC 2014中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2014
发现架构之美

层次分类中结构关系的挖掘

鞠奇

2014.9.17

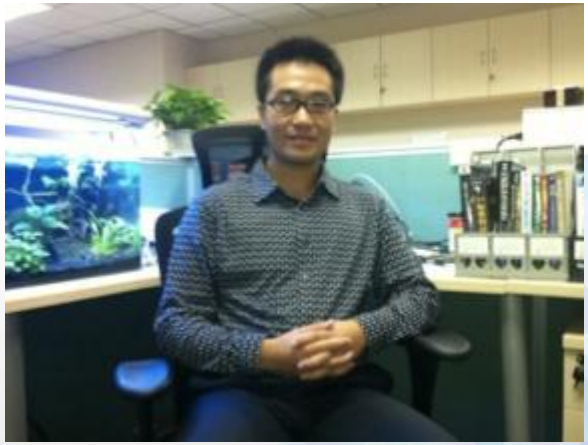
2013.4 意大利TRENTO大学 博士

IT技术男

乒乓球

管理，历史

2013.6 当当网 算法研究员



提纲

- 背景及意义
- 传统层次分类算法
- 重排序模型（acl 2012，Qi）
- 局部渐增式重排序模型
- 总结&下一部分工作

2014.3　CTO俱乐部 营销技术经验点滴谈
（图书评论处理，推荐算法，相似品牌）

目录体系对于算法处理很重要

更新当当网目录体系，为实际应用铺平道路
（reranker from acl 2012，Qi）

层次目录实例


搜狐 SOHU.com
搜狗 输入法 浏览器 地图 邮件 微博 博客 BBS 我说两句 微门户 大视野 视频 校友录 游戏 新天龙 幻想神域 手机搜狐 听书 手机游戏 e购房
新闻 军事 文化 历史 读书 原创 评论 体育 NBA CBA 中超 高尔夫 财经 理财 股票 基金 IT 数码 手机 汽车 二手车 房产 二手房 家居
时尚 女人 男人 奢侈品 美容 美食 星座 母婴 健康 旅游 教育 出国 高考 公益 绿色 播客 娱乐 美剧 音乐 彩票 城市
资讯 中国 国际 台海 航空航天
策划 军情站 兵工厂 军博馆
历史 战史 人物 野史
社区 军情观察 军事历史 军事酷图


58.com 同城
100%房东直租
同城搜索
免费发布信息
修改/删除信息
租房 鱼缸 淘宝美工 夜班 日结 团购门票 租房贷款 同城交友
首页 招聘 租房 二手房 二手车 二手市场 宠物狗 本地服务 手机58 58帮帮 团购 微站
快速放贷 先行赔付 我要推广
北京房产 毕业生租房特供
房屋出租 100%个人房源
整租 / 合租 求租
二手房
商铺出租 商铺出售
生意转让 写字楼租售
短租/日租公寓
厂房/仓库/土地/车位
二手房产经纪公司
链家地产 我爱我家
车辆买卖及服务
二手车 全国二手车 豪华车
3万以下 3-5万
5-8万 8-10万
SUV · MPV 准新车
按品牌
大众 本田 丰田 奥迪
宝马 现代 别克 雪佛兰
日产 奔驰 福特 马自达
铃木 奇瑞 夏利 比亚迪
厂商认证二手车
诚新二手车 品荐二手车
丰田二手车 品质二手车
二手市场
苹果专区 新品半价
手机 • 数码 • 手机号
二手手机 苹果 / 三星 / 小米
台式电脑 显示器 / 外设
笔记本 macbook / IBM
平板电脑 三星 / iPad
数码产品 相机 / mp4 / Xbox
手机号码 通讯业务
家具 • 家电 • 车辆
二手家电 空调 / 冰箱 / 电风扇
二手家具 沙发 / 床 / 桌
摩托车 雅马哈 / 本田
自行车 电动车
百货 • 办公 • 设备
家居日用 厨具 / 布艺 / 凉席
母婴玩具 婴儿床 / 车 / 泳池
服装箱包 手表 / 裙子 / 鞋
成人用品 美容保健
文体户外 健身 / 乐器 / 棋牌
艺术收藏 古玩 / 把件
办公设备 办公桌 / 打印机
二手设备 机床 / 工程机械
北京招聘 北京求职简历
包吃住专区 加薪再拿红包 淘宝店招聘
五险一金专区 周末双休专区 招聘会
知名企业招聘
京东商城 百度 星巴克
宜信 到家美食会 德邦物流
热招职位
销售 客服 司机 营业员
文员 助理 普工 导购员
前台 编辑 电工 收银员
会计 网管 焊工 理货员
出纳 幼教 木工 店长
教师 护士 操作工 服务员
医生 保洁 保安 厨师
翻译 保姆 库管 配菜
网编 月嫂 快递员 学徒
导游 家政 搬运工 杂工
签证顾问 房产 钟点工 洗碗工
人事专员 保险 洗车工 面点师
行政/后勤 足疗师 摄影师 送餐员
平面设计 发型师 汽车修理 防损员
网页设计 美容师 汽车美容 大堂经理
装潢设计 化妆师 理赔专员 更多职位»
本地服务大全 2014年最赚钱的创业项目
58招标：免费登记需求，让商家主动来找您
3~5万个人贷款 白手创业买车买房
家政服务
加盟生活服务沙龙，锁定无限财富
搬家 保姆/月嫂 钟点工 保洁
疏通 回收 生活配送 洗衣店
食品 礼品 鲜花绿植 租车
房屋维修 家具维修 家电 / 手机维修
电脑维修 开锁换锁 空调维修 / 移机
婚庆摄影 2014"中国好商家"火热进行中
婚庆 摄影摄像 礼仪庆典
婚车租赁 婚纱摄影 儿童摄影
装修建材
登记需求，免费领取3份户型设计»
家庭装修 工装服务 家装设计
建材 家具 家纺家饰
商务服务 58印刷广告，精致每一处细节
工商注册 财务会计 商标专利
投资担保 保险 法律 / 咨询
物流 / 快递 货运专线 起名
网站建设 网络布线 广告传媒

当当网目录体系


全部商品详细分类
新品闪购
尾品汇
图书
数字馆
服装
运动户外
孕婴童
家居
电器城
当当超市
图书/童书/数字馆
服饰/内衣
鞋靴/箱包
运动户外
孕/婴/童
家居/家纺/汽车
家具/家装/康体
美妆/个人护理/成人
食品/茶酒/宠物
珠宝饰品/手表眼镜
手机/数码
电脑办公
家用电器
婴儿服饰 >
上衣 | 裤子 | 套装 | 爬服/哈衣
学步鞋 | 婴儿礼盒 | 凉席/睡袋
童装 >
T恤/polo衫 | 裤子 | 套装 | 外套
裙子 | 羽绒服/棉服 | 毛衣/针织衫
衬衫 | 卫衣 | 内衣/家居服
马甲/背心 | 亲子装 | 配饰用品
礼服/演出服 | 泳衣
童鞋 >
运动鞋 | 休闲鞋 | 皮鞋
帆布鞋/板鞋 | 学步鞋 | 凉鞋
靴子/雪地靴 | 棉鞋 | 拖鞋/洞洞鞋
孕产妇服饰及用品 >
孕妇装 | 孕产妇内衣 | 防辐射服
孕妈用品 | 产后恢复 | 妈咪包
母婴用品 >
奶粉 | 辅食 | 营养品 | 孕妈营养
尿裤 | 孕婴洗护 | 喂养 | 孕妈用品
暖奶/消毒 | 奶瓶 | 水杯/水壶
婴儿床 | 清洁 | 护理/安全
出行用品 | 寝居
品牌
美素 | 美赞臣 | 惠氏 | 雅培
帮宝适 | 好奇 | 花王/大王
妈咪宝贝 | 嘉宝 | 亨氏 | 贝亲
喜多 | NUK | 新安怡 | 笑巴喜
良良 | 贝儿欣 | 贝贝帕克
威尔贝鲁
玩具 >
积木拼插 | 早教益智 | 模型 | 娃娃
婴幼玩具 | 遥控电动 | 毛绒 | 乐器
书包 | 童车 | 户外 | 创意
绘画手工
童书 >
0-2岁 | 3-6岁 | 7-10岁 | 11-14岁
科普/百科 | 中国儿童文学
外国儿童文学 | 精装图画书
平装图画书 | 婴儿读物 | 幼儿启蒙
益智游戏 | 玩具书 | 动漫/卡通
少儿英语 | 励志/成长 | 进口儿童书
少儿期刊 | 阅读工具书 | 更多
品牌
蒲公英童书馆 | 步印童书馆
爱心树童书 | 接力出版社
乐乐趣童书 | 双螺旋童书馆
歪歪兔童书 | 神奇校车
国际大奖小说 | 丁丁历险记
小牛顿科学馆 | 可爱的鼠小弟
贝贝熊 | 第一次发现
童装馆
母婴馆
玩具馆

存在的问题

- 商品运营事业部，根据供应链，人工建立（机械表）

  当当网 > 手表眼镜 > 手表 > 日韩品牌表

  当当网 > 手表眼镜 > 手表 > 国产品牌表

  当当网 > 手表眼镜 > 手表 > 时尚品牌表

  保温杯：户外运动，日用家居等

  羽绒服：户外运动，服装等

- 信息被零碎化，没有从用户选购商品角度考虑

存在的问题

- 商品事业部基于商家，供销数据，专业知识等
- 技术基于用户选购角度的商品目录体系
- 影响搜索，广告，推荐等实际应用

其他问题

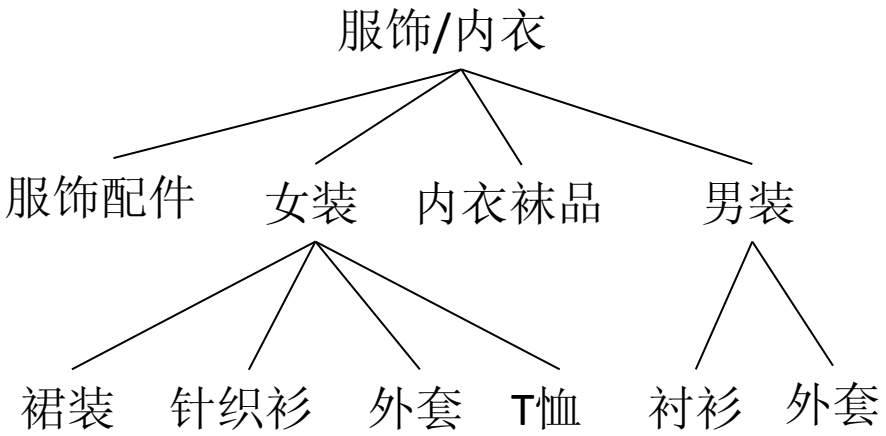
- 内部人员编辑内容
  - 内容多面性
  - 信息零碎化
  - 大数据，耗费人力
  - 实时性难以得到满足

- 统一的分类模型
  - 基于用户，保留信息整体性
  - 节省事业部人力
  - 无缝隙和技术开发对接
  - 保证实时性，高效

提纲

- 背景及意义
- 传统层次分类方法
- 重排序算法
  - 扩展结果产生
  - 最优结果选择
  - 结果的结构化表示
  - 正负样本构建
  - Reranker训练
  - Reranker测试效果及性能
- 局部渐增式重排序模型
- 总结及下一步工作

扁平分类

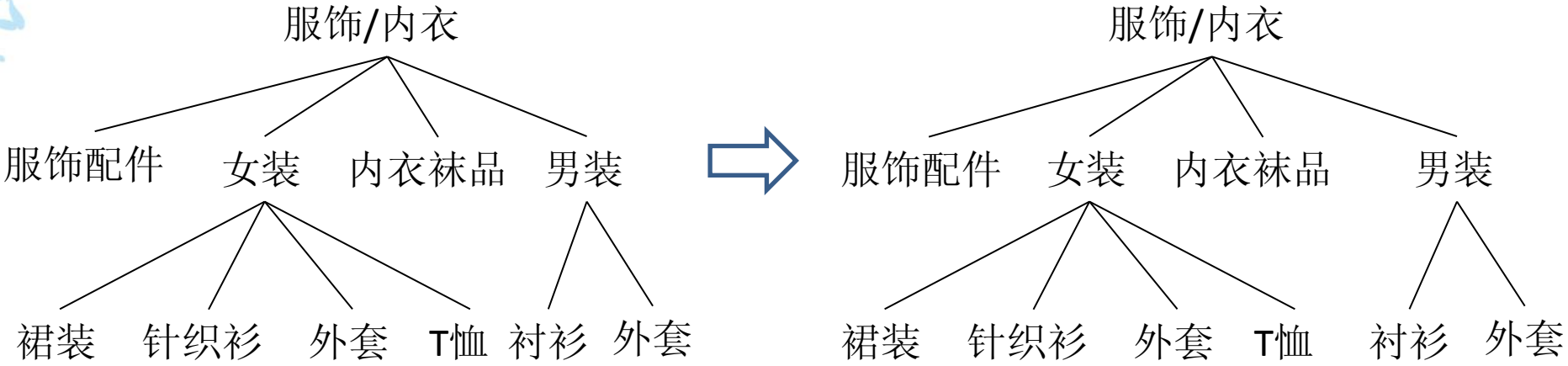


二分类模型

↓ ……

服饰/内衣 服饰配件 女装 内衣袜品 男装 裙装 针织衫 外套 T恤 衬衫 外套

自顶向下分类



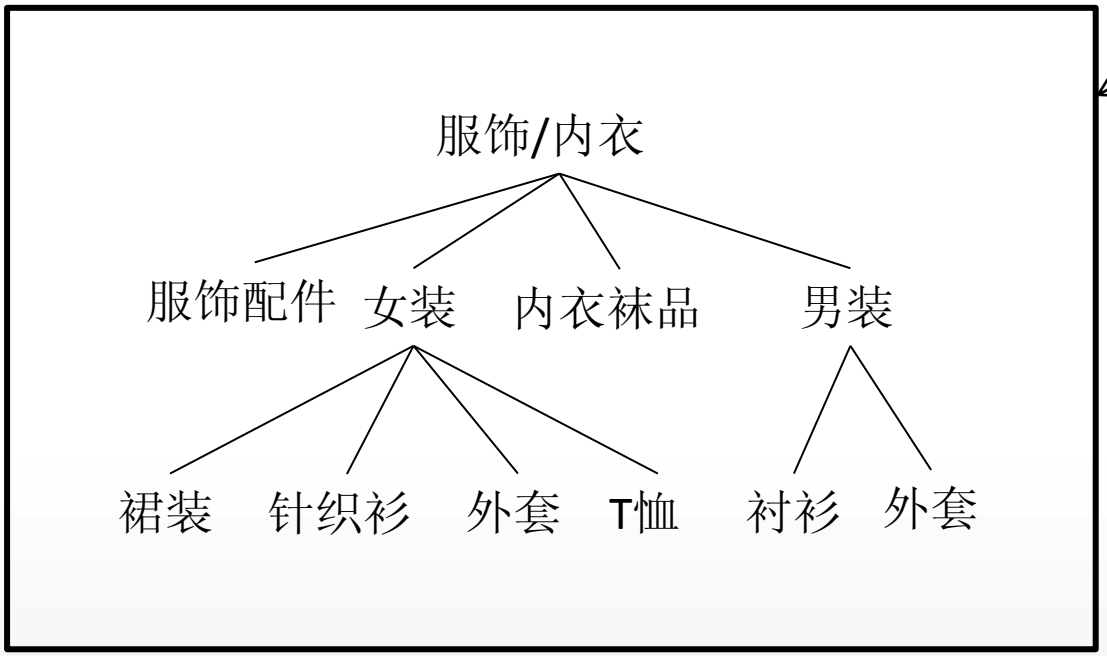
非叶子节点类：2个二分类模型，其中：

--- 子树模型：判断商品是否属于以这个节点为根的子树；

--- 自身模型：判断item是否属于这个节点类本身；

叶子节点类：只有自身模型。

层次整体分类


全局model
服饰/内衣
服饰配件
女装
内衣袜品
男装
裙装
针织衫
外套
T恤
衬衫
外套


秋冬
新品
¥99.00
¥259.00
(3.8折)
唐狮男装 秋装新款拼接针织绞花长袖
九月力促 满200减40 满300减80

传统分类算法缺点

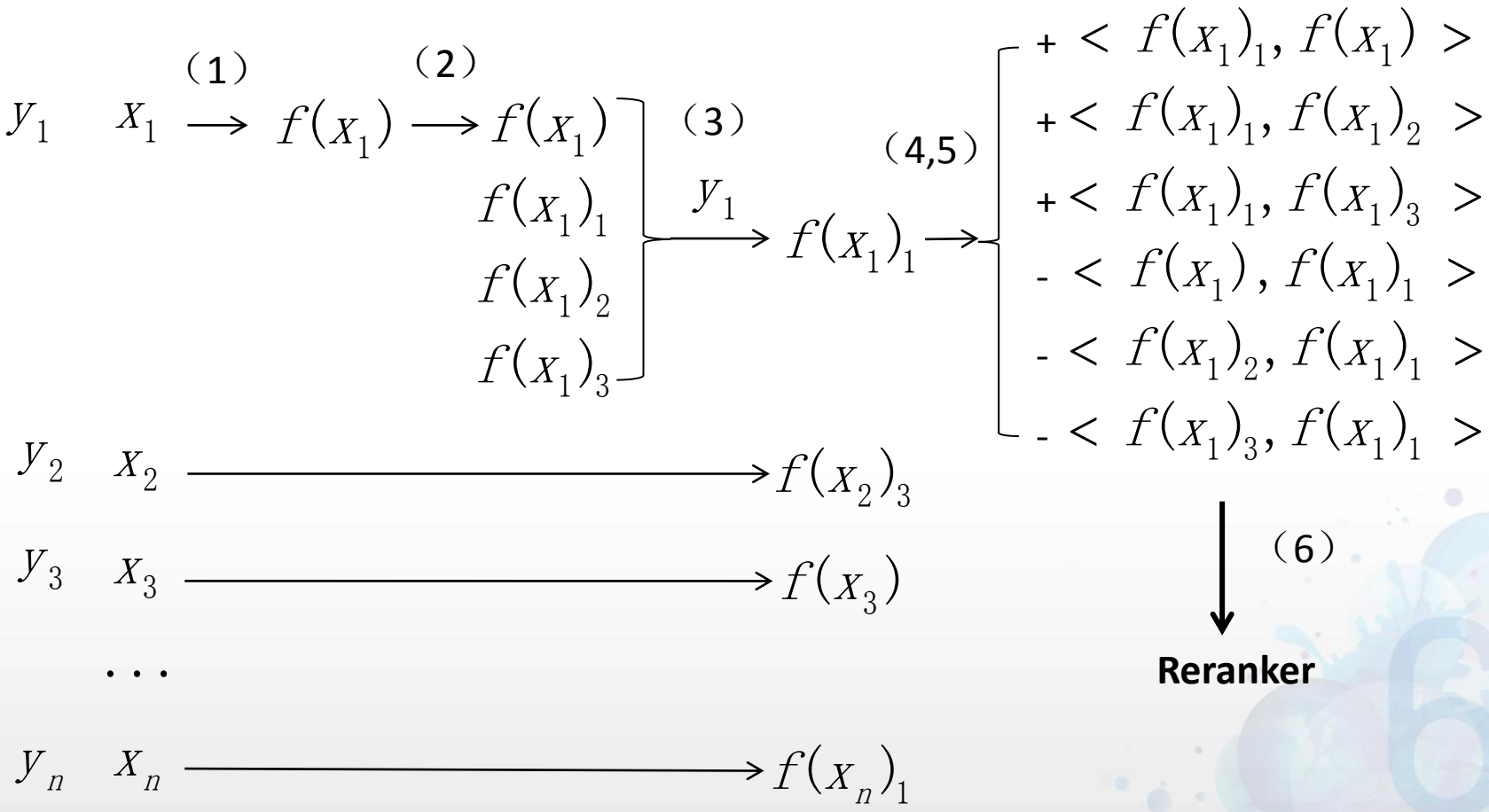
- 相对较低准确性（没有考虑类别依赖关系）
  - 扁平分类
  - 自顶向下的层次分类
- 性能低（考虑了依赖关系）
  - 整体分类

考虑依赖关系，高性能的算法

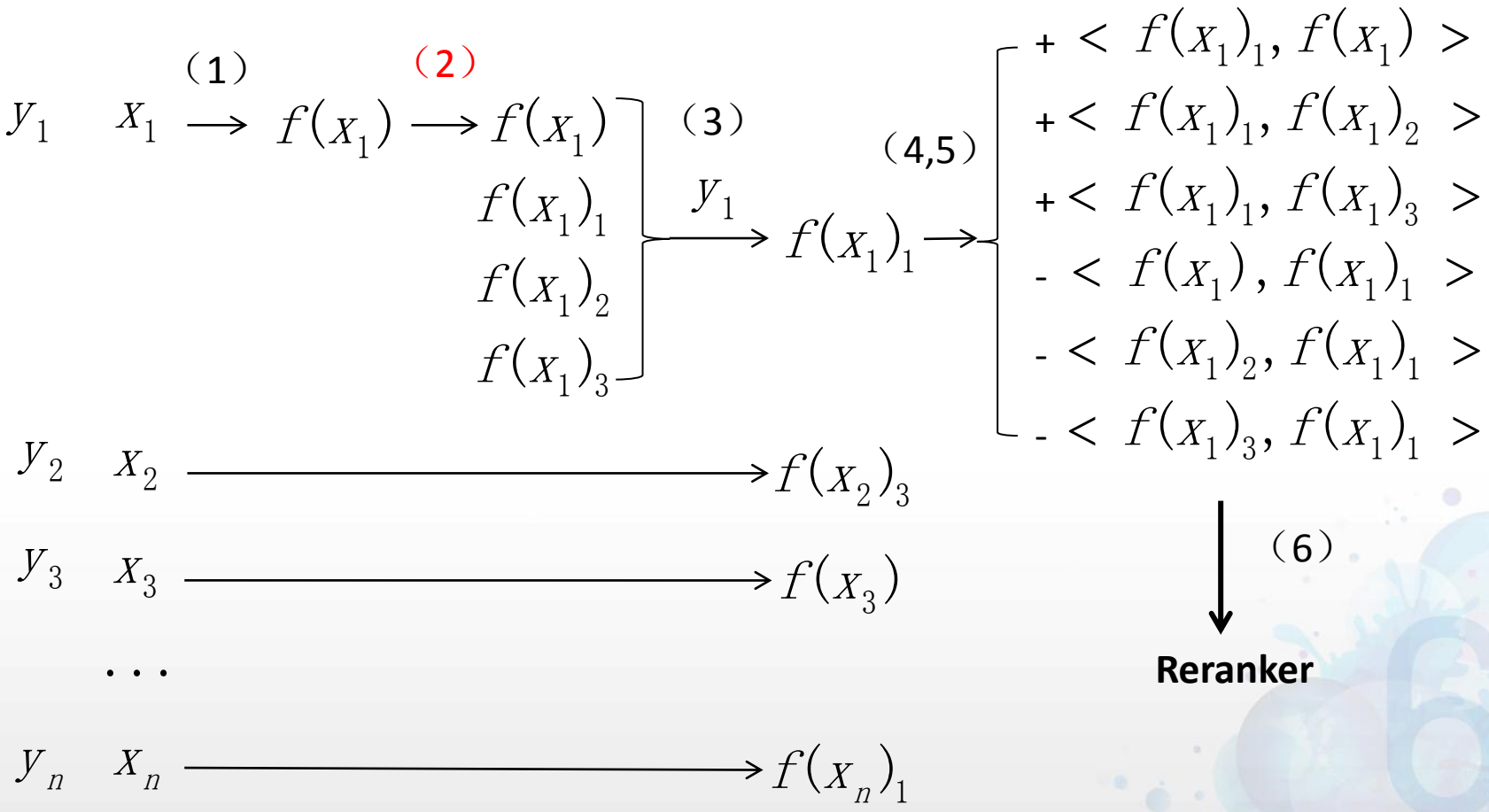
提纲

- 背景及意义
- 传统层次分类方法
- 重排序算法
  - 扩展结果产生
  - 最优结果选择
  - 结果的结构化表示
  - 正负样本构建
  - Reranker训练
  - Reranker测试效果及性能
- 局部渐增式重排序模型
- 总结及下一步工作

Reranker--训练

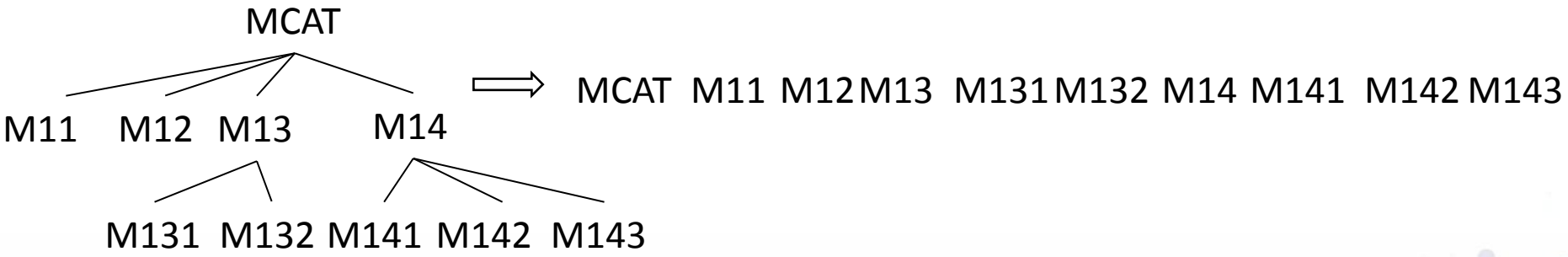

(1)
(2)
(3)
(4,5)
(6)
y_1 x_1 $f(x_1)$ $f(x_1)$ $f(x_1)_1$ $f(x_1)_2$ $f(x_1)_3$
y_1 $f(x_1)_1$
+ < $f(x_1)_1$, $f(x_1)$ >
+ < $f(x_1)_1$, $f(x_1)_2$ >
+ < $f(x_1)_1$, $f(x_1)_3$ >
- < $f(x_1)$, $f(x_1)_1$ >
- < $f(x_1)_2$, $f(x_1)_1$ >
- < $f(x_1)_3$, $f(x_1)_1$ >
y_2 x_2 $f(x_2)_3$
y_3 x_3 $f(x_3)$
. . .
y_n x_n $f(x_n)_1$
Reranker

Reranker--训练



扁平分类结果扩展

- SVM产生分类概率；
- 联合概率P：所以类别概率的乘积；
- X_1： M12, M132



$$P = (1 - p_{MCAT}) \times (1 - p_{M11}) \times p_{M12} \times (1 - p_{M_{13}}) \times (1 - p_{M131}) \times p_{M132}$$
$$\times (1 - p_{M14}) \times (1 - p_{M141}) \times (1 - p_{M142}) \times (1 - p_{M143})$$

扁平分类结果扩展

- $f(x_1)$：原始输出(最大P)：　P= 0.4040

MCAT　M11　M12　M13　M131　M132　M14　M141　M142　M143

$p_{MCAT}=0.003$ $p_{M11}=0.006$ $p_{M12}=0.453$ $p_{M13}=0.006$ $p_{M131}=0.023$ $p_{M132}=0.779$ $p_{M14}=0.009$ $p_{M141}=0.001$ $p_{M142}=0.004$ $p_{M142}=0.001$

- $f(x_1)_1$：在第1个基础上改变M12 (第二大P)：　Pro= 0.3346

MCAT　M11　M12　M13　M131　M132　M14　M141　M142　M143

$p_{MCAT}=0.003$ $p_{M11}=0.006$ $p_{M12}=0.453$ $p_{M13}=0.006$ $p_{M131}=0.023$ $p_{M132}=0.779$ $p_{M14}=0.009$ $p_{M141}=0.001$ $p_{M142}=0.004$ $p_{M142}=0.001$

- $f(x_1)_2$：基于第二个(第三大P)：　Pro= 0.095

MCAT　M11　M12　M13　M131　M132　M14　M141　M142　M143

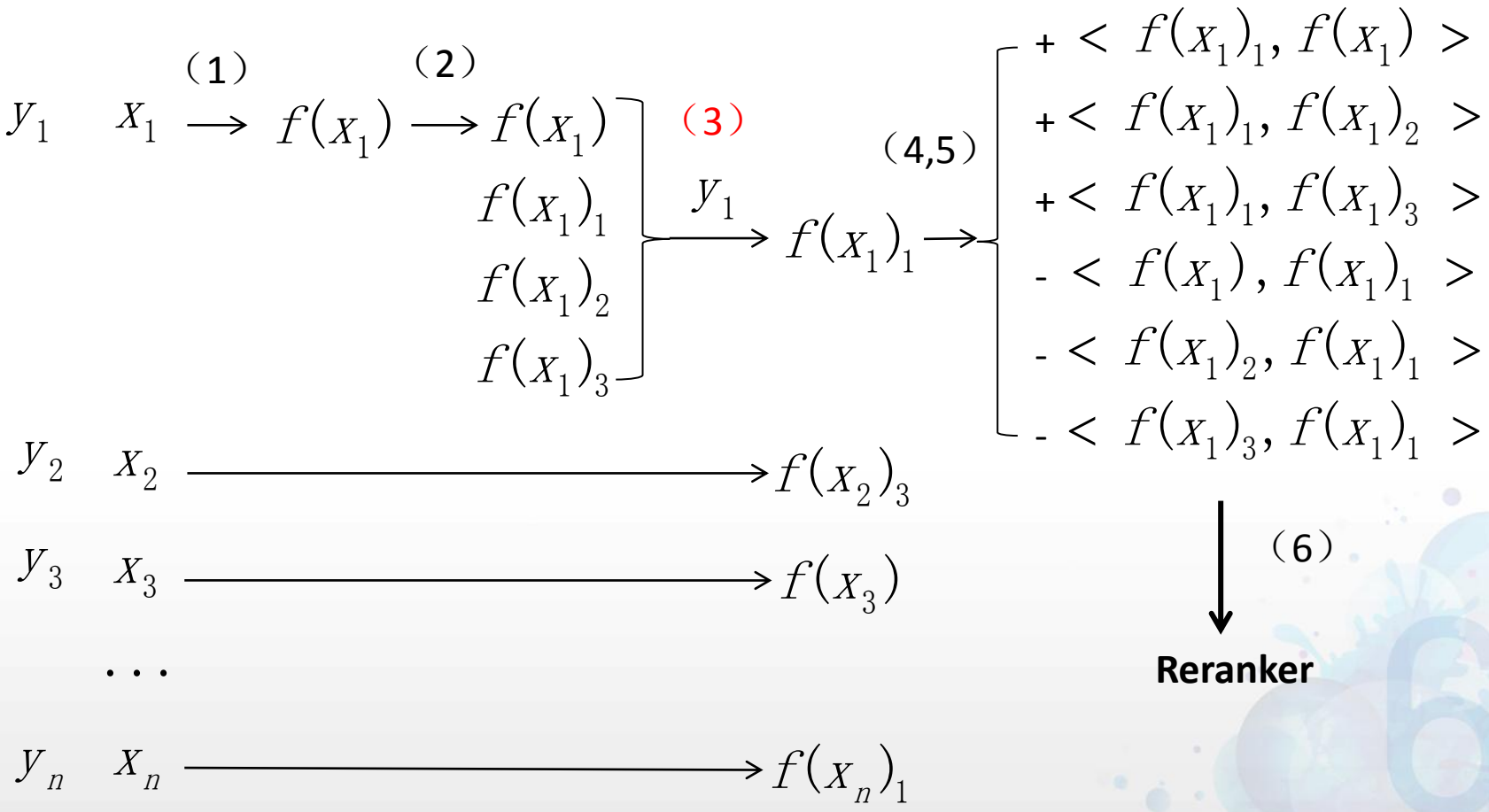
$p_{MCAT}=0.003$ $p_{M11}=0.006$ $p_{M12}=0.453$ $p_{M13}=0.006$ $p_{M131}=0.023$ $p_{M132}=0.779$ $p_{M14}=0.009$ $p_{M141}=0.001$ $p_{M142}=0.004$ $p_{M142}=0.001$

- $f(x_1)_3$：基于第一个(第三大P)：　Pro= 0.009

MCAT　M11　M12　M13　M131　M132　M14　M141　M142　M143

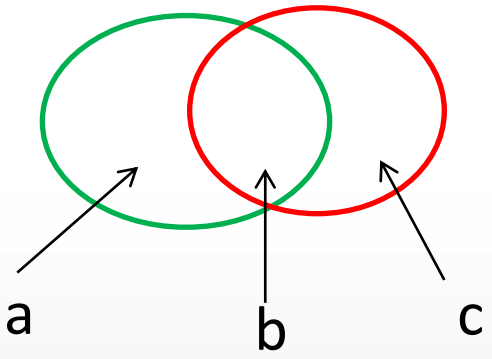
$p_{MCAT}=0.003$ $p_{M11}=0.006$ $p_{M12}=0.453$ $p_{M13}=0.006$ $p_{M131}=0.023$ $p_{M132}=0.779$ $p_{M14}=0.009$ $p_{M141}=0.001$ $p_{M142}=0.004$ $p_{M142}=0.001$

Reranker--训练


（1）
（2）
（3）
（4,5）
（6）
y_1 x_1 $f(x_1)$ $f(x_1)$ $f(x_1)_1$ $f(x_1)_2$ $f(x_1)_3$
y_1 $f(x_1)_1$
+ < $f(x_1)_1$, $f(x_1)$ >
+ < $f(x_1)_1$, $f(x_1)_2$ >
+ < $f(x_1)_1$, $f(x_1)_3$ >
- < $f(x_1)$, $f(x_1)_1$ >
- < $f(x_1)_2$, $f(x_1)_1$ >
- < $f(x_1)_3$, $f(x_1)_1$ >
y_2 x_2 $f(x_2)_3$
y_3 x_3 $f(x_3)$
. . .
y_n x_n $f(x_n)_1$
Reranker

最优扩展结果选择

- 比较 Y_1 与$f(x_1)$，$f(x_1)_1$，$f(x_1)_2$，$f(x_1)_3$
- Pr, Re和F1值



标注值

预测值

Pr = b/(b+c)

Re = b/(a+b)

F1=2*Pr*Re/(Pr+Re)

最优扩展结果选择

- 由下表得知，对于x_1，$f(x_1)_1$ 在4个中是最好的

	Precision	Recall	F1
$f(x_1)$	1/1	1/2	0.667
$f(x_1)_1$	2/2	2/2	1
$f(x_1)_2$	1/1	1/2	0.667
$f(x_1)_3$	1/2	1/2	0.5

性能空间

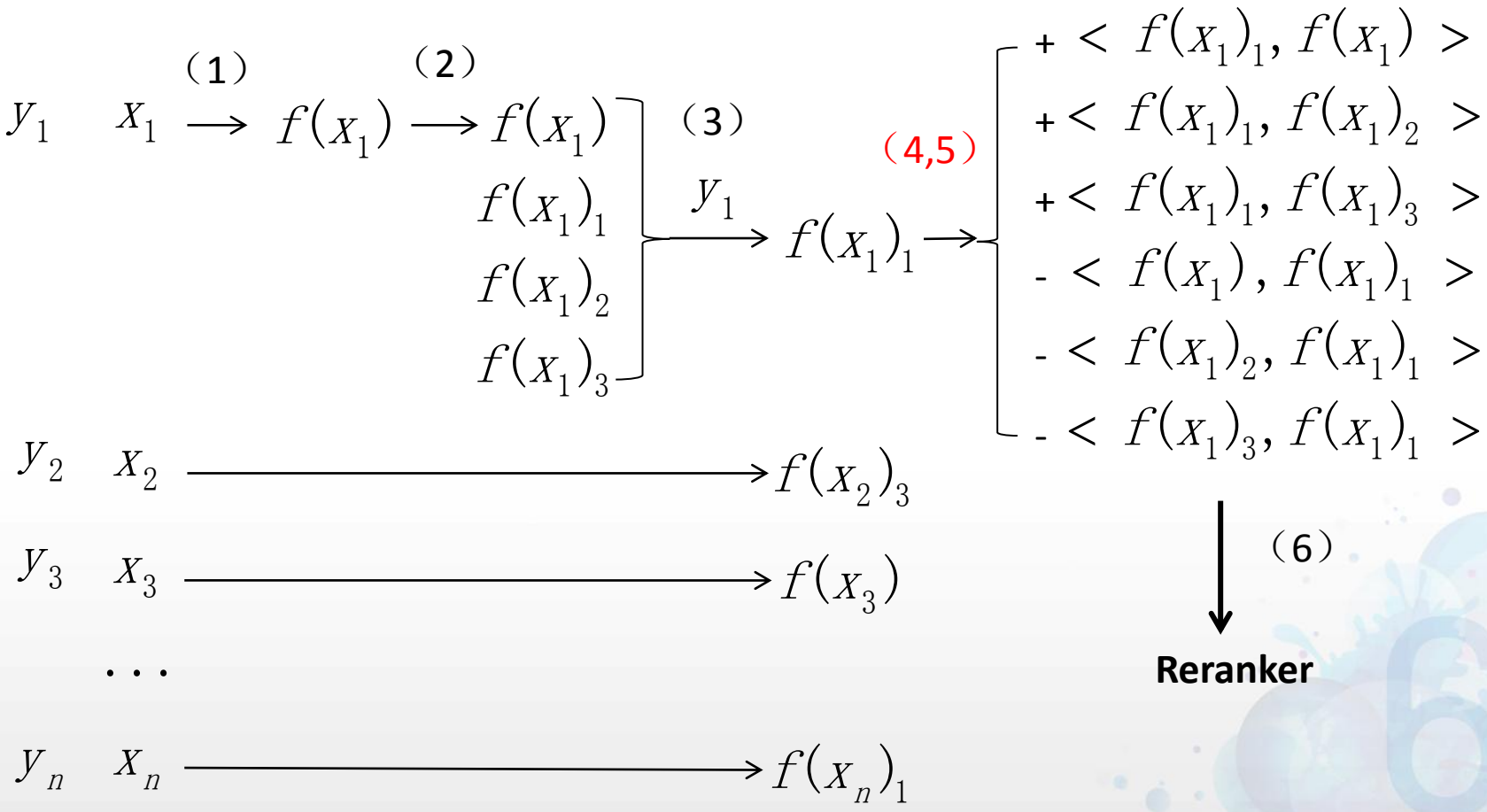
- 对于每一个x，假设总能选择最好的

k	Flat Generation	
	Micro-F_1	Macro-F_1
1	0.640	0.408
2	0.758	0.504
4	0.821	0.566
8	0.858	0.610
16	0.898	0.658

Micro-F1 = 2*Pr_总* Re_总/(Pr_总 + Re_总)
Macro_F1 = average(sum(F1))

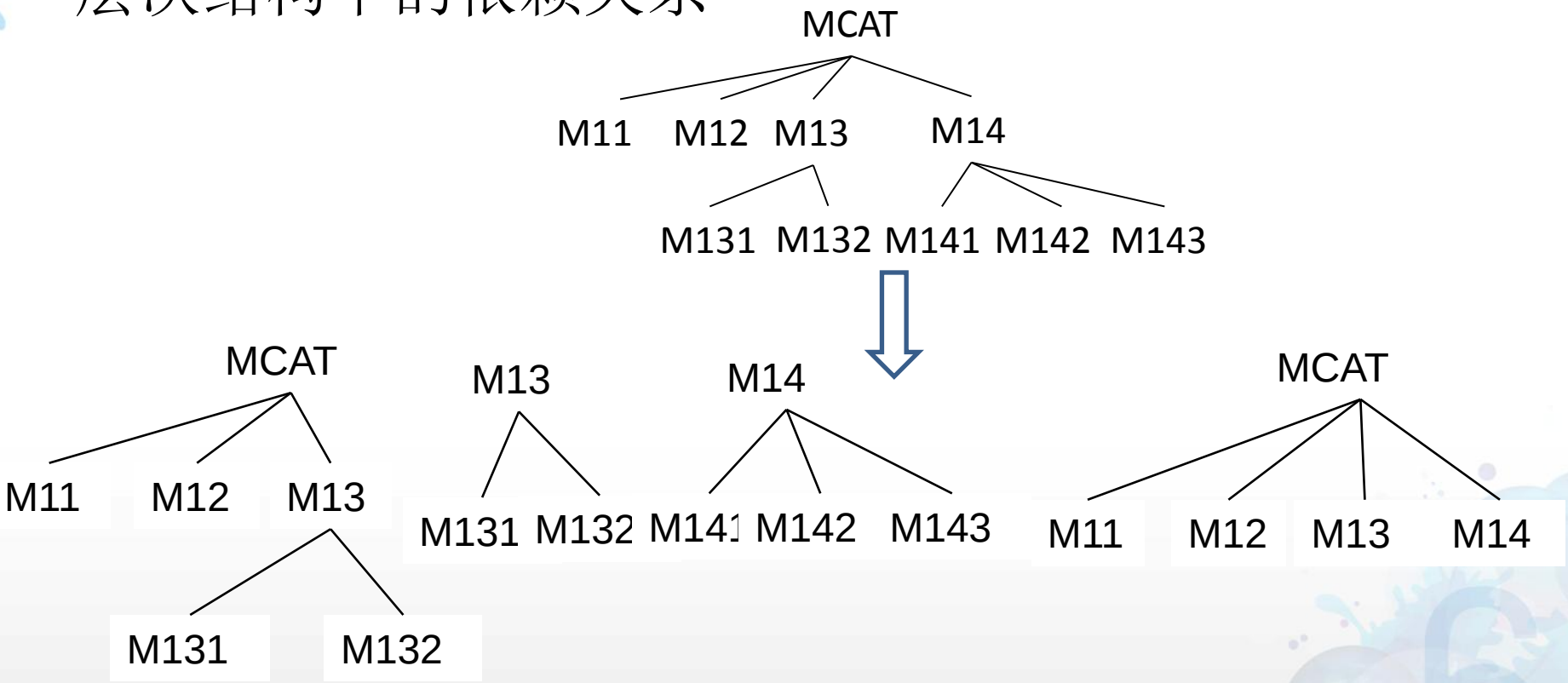
- 最优性能提供了一个我们可优化的空间

Reranker--训练



依赖关系

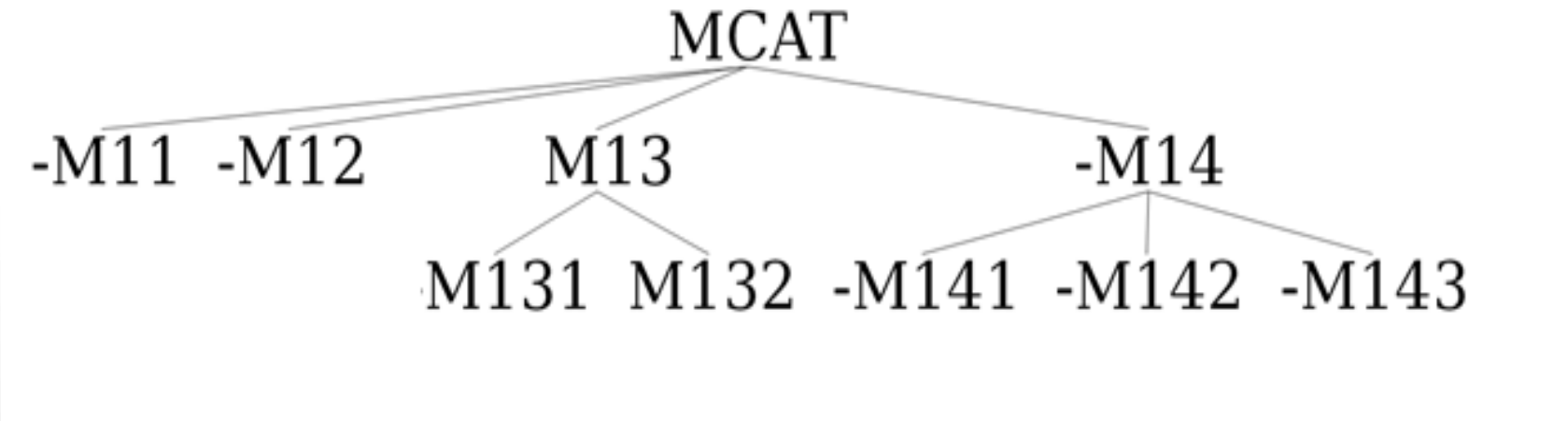
- 层次结构中的依赖关系



边相连的类都具有依赖关系

依赖关系表示

- 层次结构:
  - 样本 x：M131，M132



重排序模型

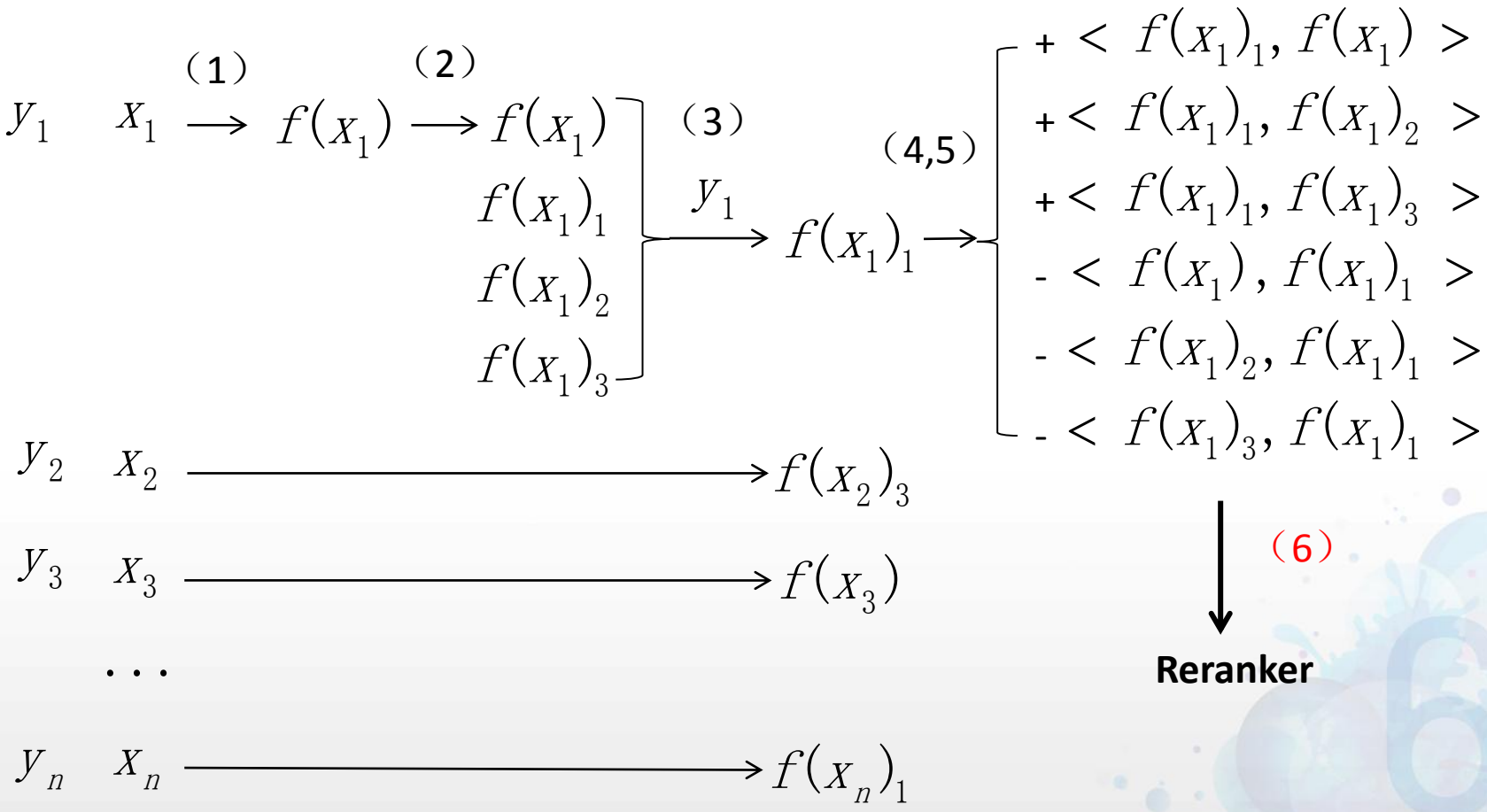
- 基于对样本x 产生的结果，我们构造正负样例：
  - 正样本：$\langle f(\mathrm{x})_i, f(\mathrm{x})_j \rangle$
  - 负样本：$\langle f(\mathrm{x})_j, f(\mathrm{x})_i \rangle$

  其中，$f(\mathrm{x})_i$ 是最好的，$f(\mathrm{x})_j \in \{f(\mathrm{x})_1, \ldots, f(\mathrm{x})_k\} - f(\mathrm{x})_i$

- 应用著名的偏好性核函数方法 训练二分类模型

- 该二分类能够区分$f(\mathrm{x})_i$是否比$f(\mathrm{x})_j$更好。

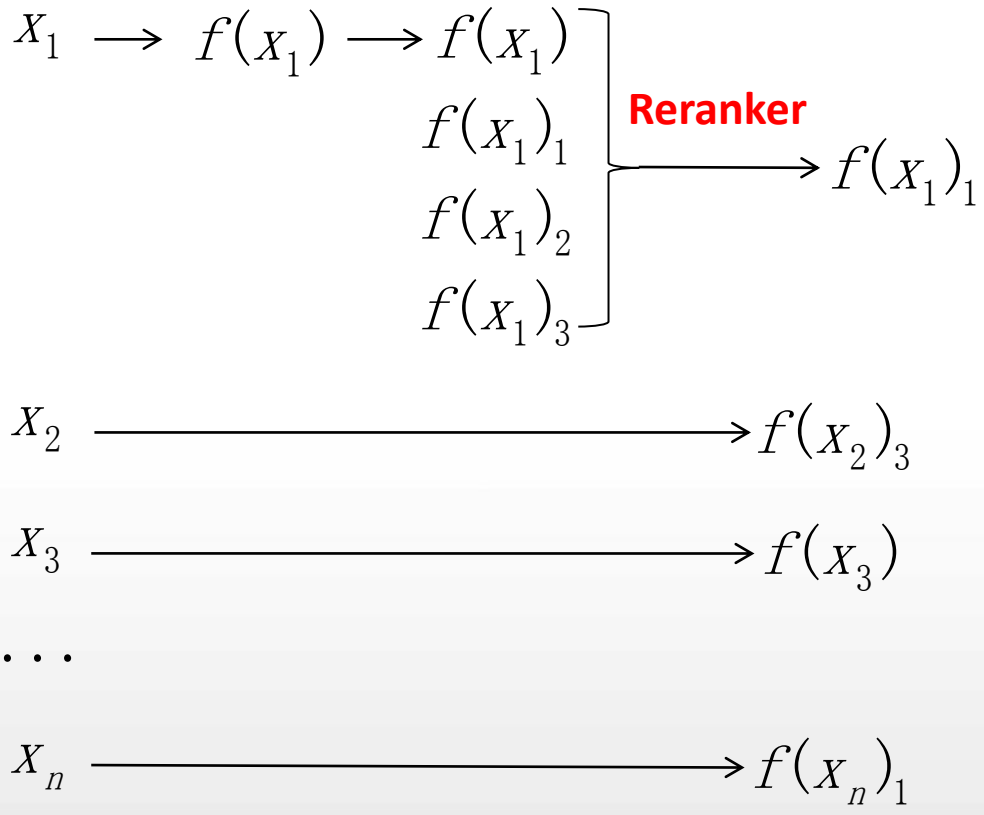
Reranker--训练


（1）
（2）
（3）
（4,5）
（6）
y_1 x_1 $f(x_1)$ $f(x_1)$ $f(x_1)_1$ $f(x_1)_2$ $f(x_1)_3$
y_1 $f(x_1)_1$
+ < $f(x_1)_1$, $f(x_1)$ >
+ < $f(x_1)_1$, $f(x_1)_2$ >
+ < $f(x_1)_1$, $f(x_1)_3$ >
- < $f(x_1)$, $f(x_1)_1$ >
- < $f(x_1)_2$, $f(x_1)_1$ >
- < $f(x_1)_3$, $f(x_1)_1$ >
y_2 x_2 $f(x_2)_3$
y_3 x_3 $f(x_3)$
. . .
y_n x_n $f(x_n)_1$
Reranker

提纲

- 重排序模型
  - 传统层次分类方法（1）
  - 扩展结果产生（2）
  - 最优结果选择（3）
  - 结果的结构化表示（4）
  - 正负样本构建（5）
  - Reranker训练（6）
  - Reranker测试效果及性能（7）

Reranker--预测



实验数据及工具

- RCV1-v2/LYRL2004
  - 103个类，5层。（MCAT来自一个小分支）
  - 训练集：23,149，测试集：781,265
- DMOZ 数据集（来自ECML/PKDD Discovery Challenge）
  - 5层， 35,448 个类，其中27,875个是叶子类
  - 300,000 训练样本， 94,756测试样本
- Liblinear vs reranker

实验结果

- 准确性（RCV1）

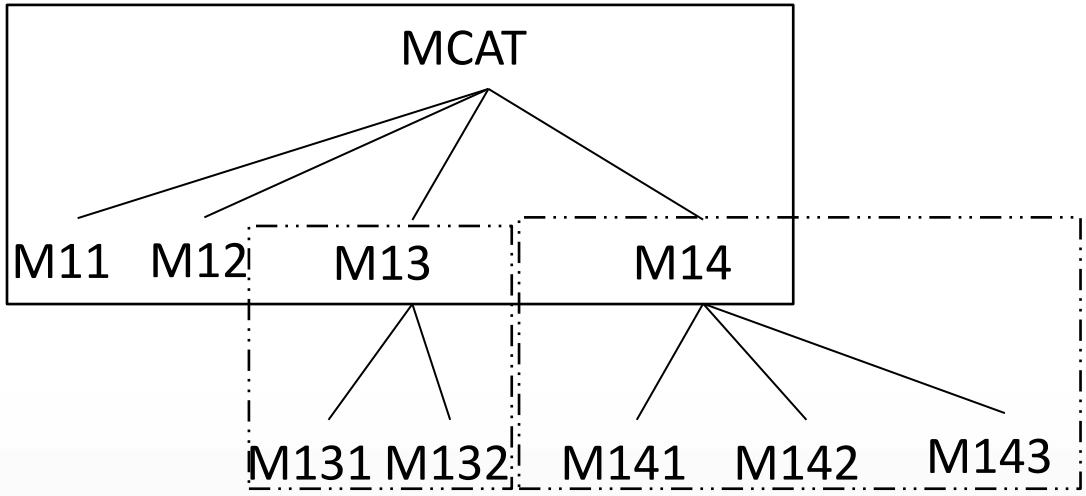
	扁平分类	
	liblinear	reranker
Micro_F1	0.775	0.849
Macro_F1	0.516	0.615

提纲

- 背景及意义
- 传统层次分类方法
- 重排序算法
  - 扩展结果产生
  - 最优结果选择
  - 结果的结构化表示
  - 正负样本构建
  - Reranker训练
  - Reranker测试效果及性能
- 局部渐增式重排序模型
- 总结及下一步工作

局部重排序模型

- reranker：效率稍低；
- 局部rerankers：



- 局部依赖关系
- 提高高层次判定准确性
- 自顶向下保证了性能

实验结果

- 准确性（RCV1）：

F1	reranker	局部渐增reranker
Micro-F1	0.855	0.846
Macro-F1	0.634	0.619

- 效率：

Time cost	reranker	局部渐增reranker
Training (s)	9023.24	508.75
Test (min)	434.08	14.19

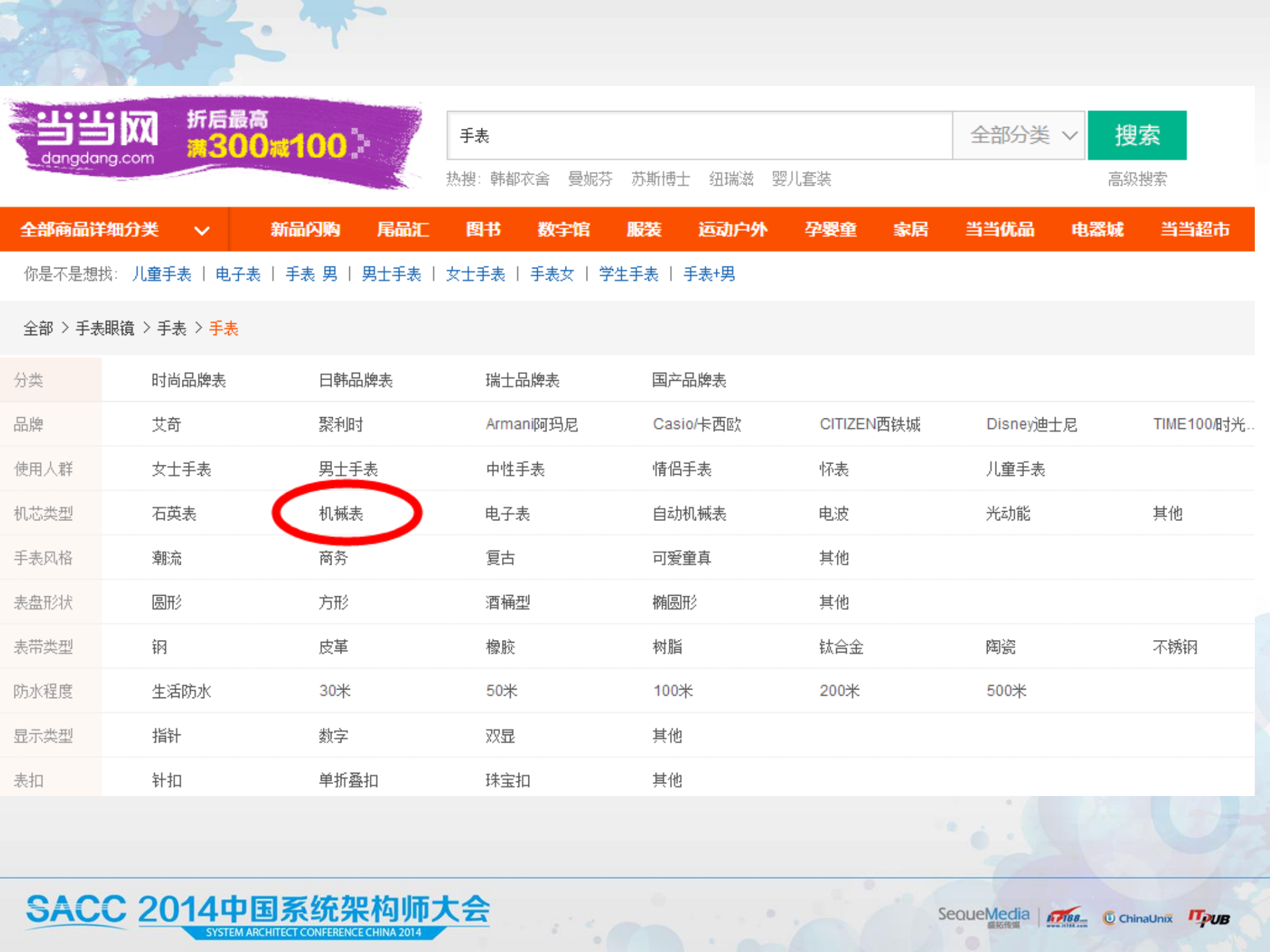
实验结果

- 准确性（DMOZ）：

F1	liblinear	局部渐增reranker
Micro-F1	0.601	0.734
Macro-F1	0.202	0.366

- 效率：

Time cost	liblinear	局部渐增reranker
Training (min)	60.18	81.26
Test (min)	19.74	39.83



分类	时尚品牌表	日韩品牌表	瑞士品牌表	国产品牌表			
品牌	艾奇	聚利时	Armani阿玛尼	Casio/卡西欧	CITIZEN西铁城	Disney迪士尼	TIME100/时光..
使用人群	女士手表	男士手表	中性手表	情侣手表	怀表	儿童手表	
机芯类型	石英表	机械表	电子表	自动机械表	电波	光动能	其他
手表风格	潮流	商务	复古	可爱童真	其他		
表盘形状	圆形	方形	酒桶型	椭圆形	其他		
表带类型	钢	皮革	橡胶	树脂	钛合金	陶瓷	不锈钢
防水程度	生活防水	30米	50米	100米	200米	500米	
显示类型	指针	数字	双显	其他			
表扣	针扣	单折叠扣	珠宝扣	其他			

总结

- 背景及意义
- 传统层次分类方法
- 重排序模型
  - 扩展结果产生
  - 最优结果选择
  - 结果的结构化表示
  - 正负样本构建
  - Reranker训练
  - Reranker测试效果及性能
- 局部渐增式重排序模型
- 总结

Q&A

THANKS

SequeMedia 盛拓传媒 | IT168 www.it168.com | ChinaUnix | ITPUB