基于机器学习的搜索语义理解技术

陈翔

百度资深研发工程师

CONTENT

目录

1 人工智能：搜索的二次进化

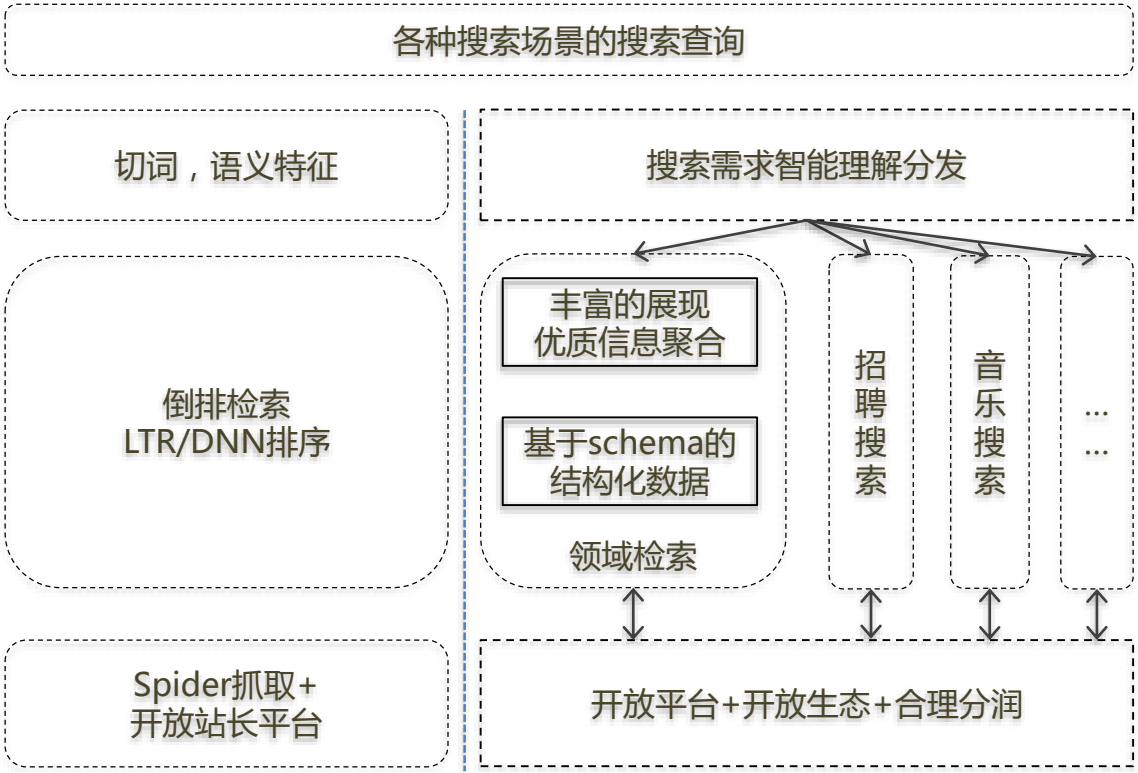
搜索场景智能化

>> 语音搜索

>> 图像搜索

>> 物联网等其他载体

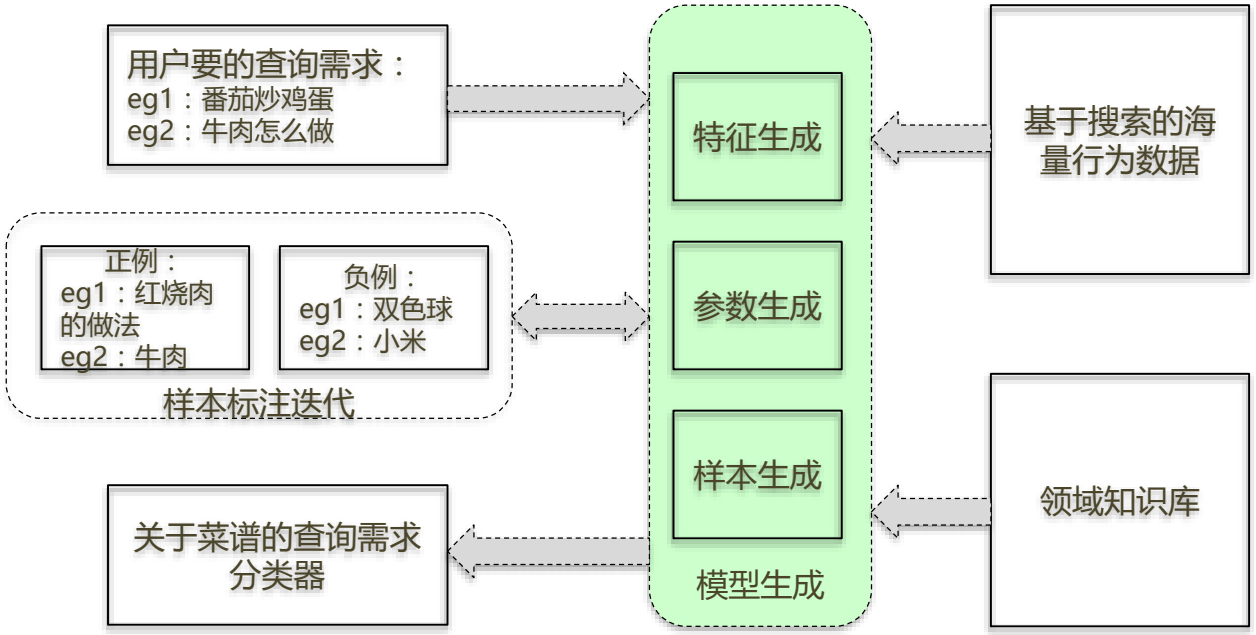
搜索生态进化



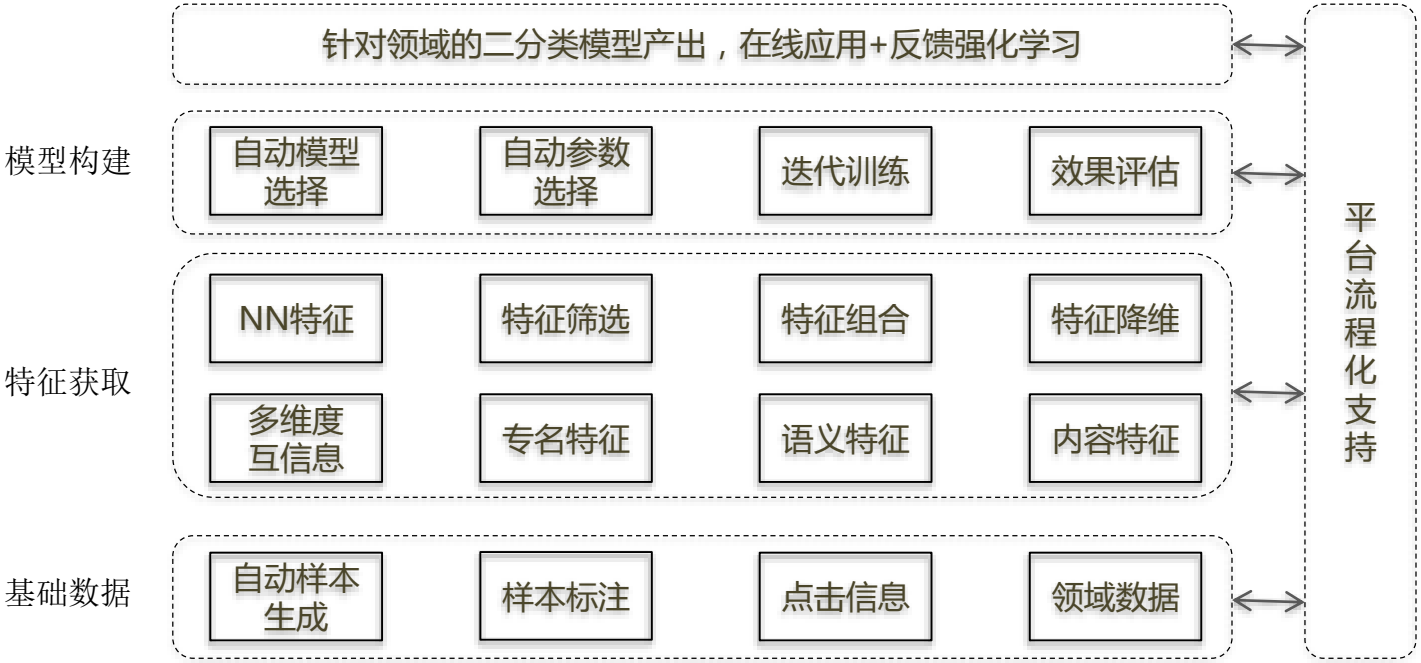
- 通用+领域检索智能聚合
- 基于大数据+机器学习的算法内核
- 基于开放生态的更丰富数据

2 需求理解：通用语义需求识别模型

需求理解


用户要的查询需求：
eg1：番茄炒鸡蛋
eg2：牛肉怎么做
正例：
eg1：红烧肉的做法
eg2：牛肉
负例：
eg1：双色球
eg2：小米
样本标注迭代
关于菜谱的查询需求分类器
特征生成
参数生成
样本生成
模型生成
基于搜索的海量行为数据
领域知识库

策略概览


针对领域的二分类模型产出，在线应用+反馈强化学习
模型构建
自动模型选择
自动参数选择
迭代训练
效果评估
特征获取
NN特征
特征筛选
特征组合
特征降维
多维度互信息
专名特征
语义特征
内容特征
基础数据
自动样本生成
样本标注
点击信息
领域数据
平台流程化支持

通过随机游走，产生更丰富的特征信息

- Input：手机百度下载
- Features：
  - Template：[Appname]下载/
    手机[Appname]下载
  - Term：手机百度/下载/手机/百度
  - Related Url-pattern：
    - mobile.baidu.com/
    - dl.pconline.com.cn/download/
    - www.wandoujia.com/apps/
    - …
  - Sim-query features：
    - Features of “手百下载”
    - Features of “手机百度app下载”
    - …
- 基于策略控制的随机游走+标注
- 百万维基础特征

Queries Q

q_1: jobs in chicago
q_2: jobs in boston
q_3: jobs in microsoft
q_4: jobs in motorola
q_5: marketing jobs in motorola
q_6: 401k plans
q_7: illinois employment statistics

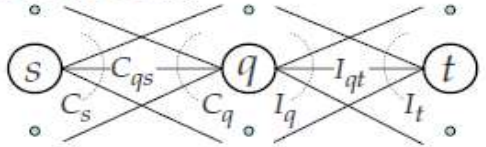
Sites S

s_1: monster.com
s_2: motorola.com
s_3: us401k.com

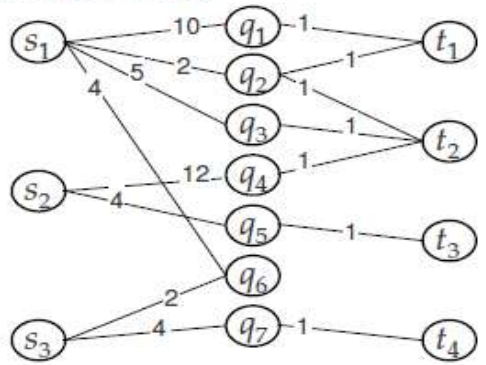
Templates T

t_1: jobs in #location
t_2: jobs in #company
t_3: #category jobs in #company
t_4: #location employment statistics

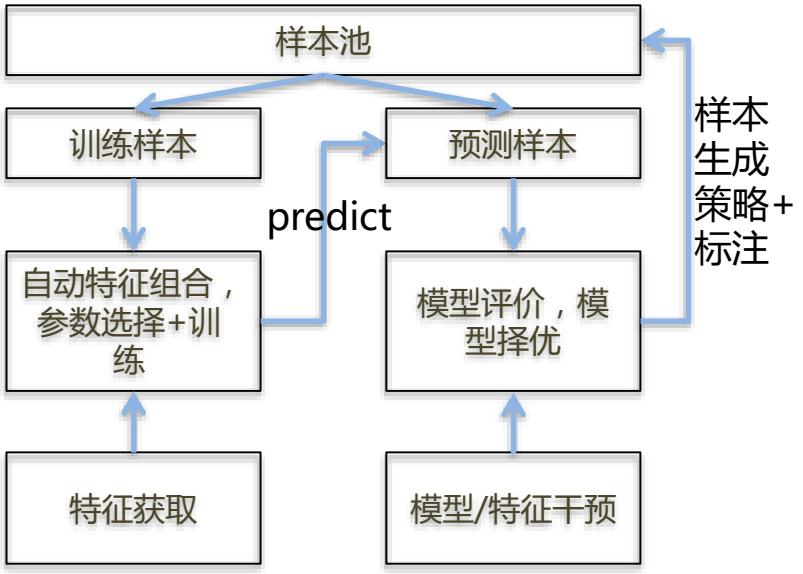
General Form:



Example Graph:



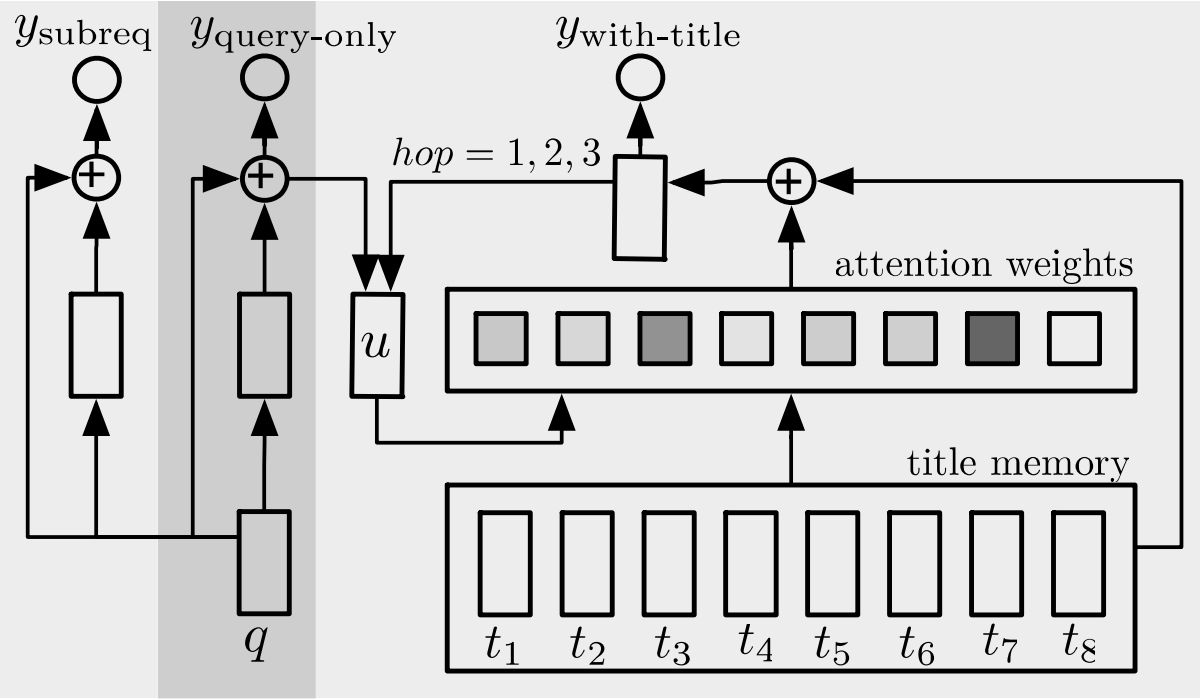
模型生成+强化学习
贴近最终应用的准召优化



Tips

- ✧ 样本挑选—
  边界样本+随机样本
  +关键特征相关样本
  +半监督自动样本挖掘
- ✧ 多种特征组合方案。
- ✧ 多种模型/相关参数
- ✧ 小成本干预手段

用深度学习构建记忆网络
丰富语义信息，提升泛化能力



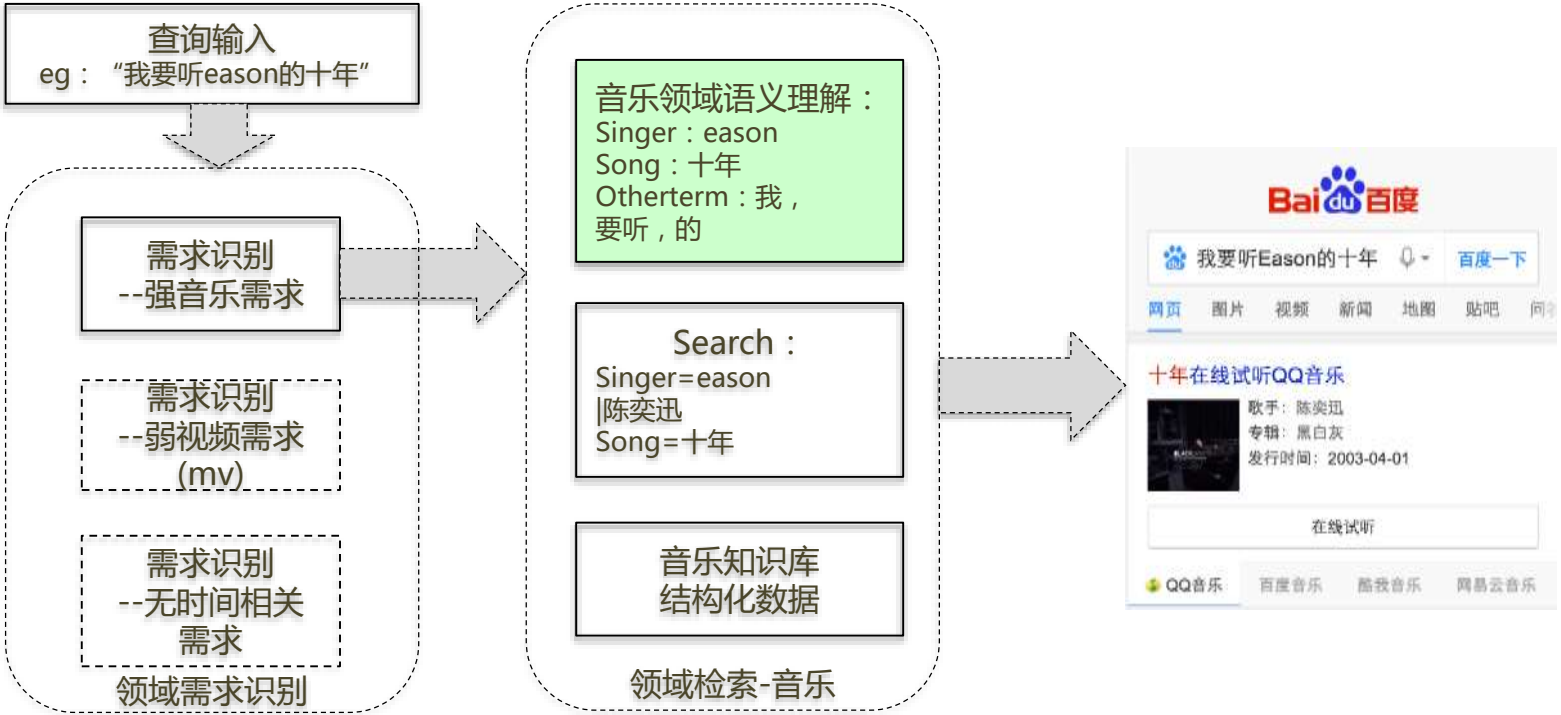
Tips

- End-To-End,训练记忆网络，进一步提升泛化能力
  - Eg：局气->餐厅
  - Eg：桥西外国语小学vs桥东
- 复用已有的算法和特征，产生大量高准召样本
- 深度学习的产出，作为最终模型的特征输入，重新参与最终模型生成

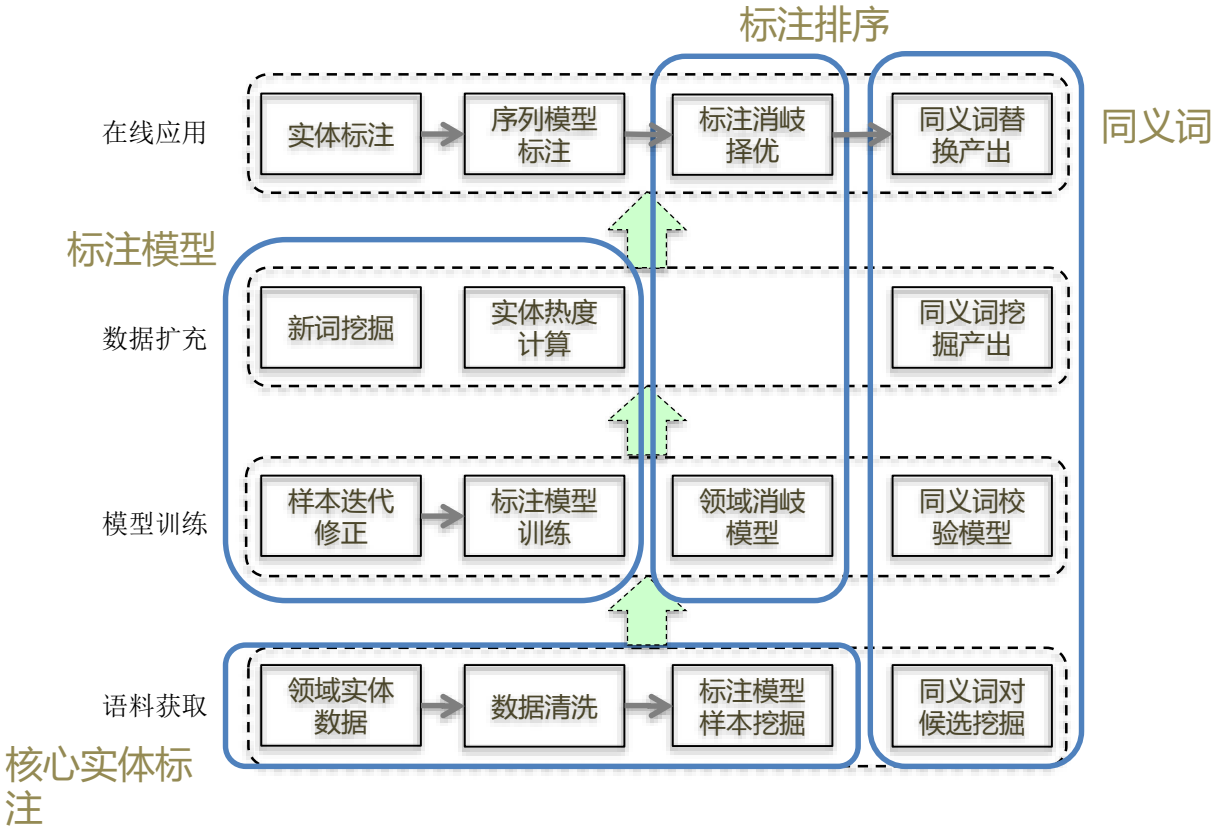
- 效果:
  - 核心评测集F1：0.84 -> 0.90

3 领域语义理解模型

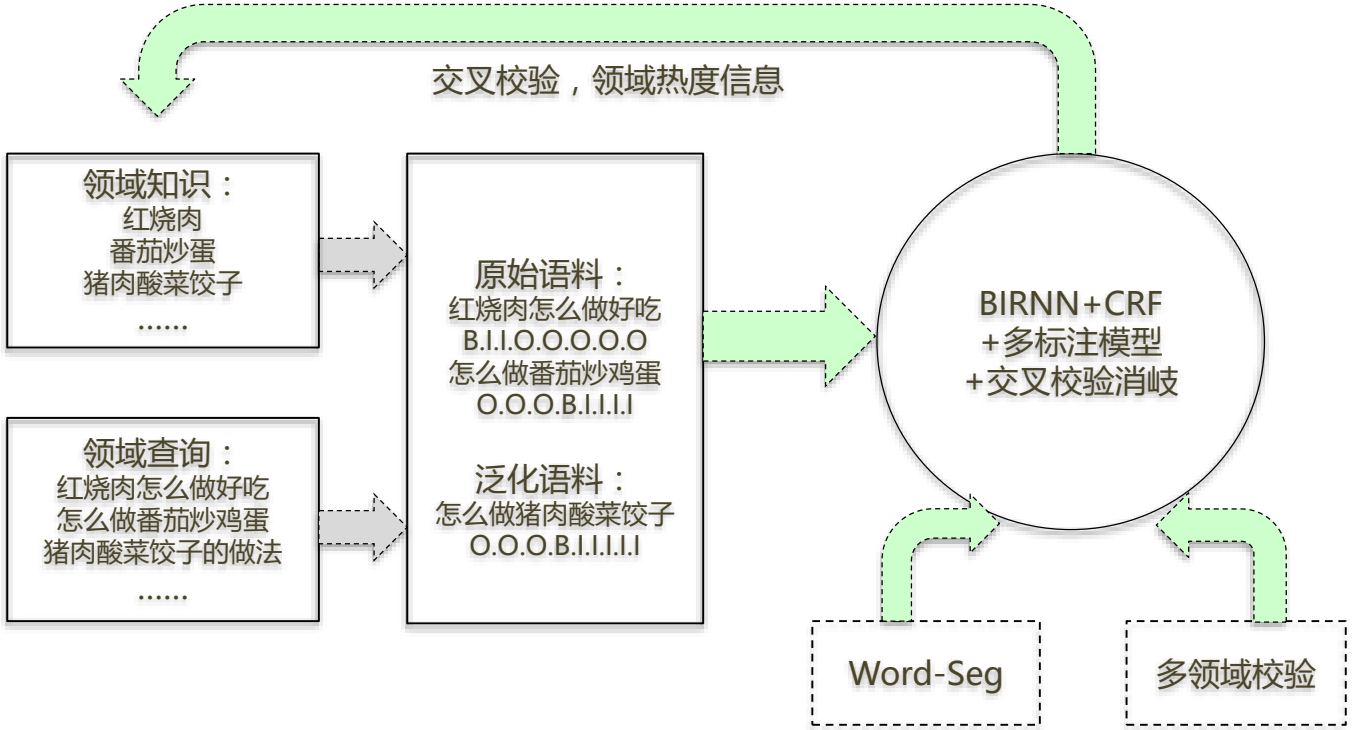
领域语义理解


查询输入
eg： “我要听eason的十年”
需求识别
--强音乐需求
需求识别
--弱视频需求
(mv)
需求识别
--无时间相关
需求
领域需求识别
音乐领域语义理解：
Singer：eason
Song：十年
Otherterm：我，
要听，的
Search：
Singer=eason
|陈奕迅
Song=十年
音乐知识库
结构化数据
领域检索-音乐
我要听Eason的十年
百度一下
网页
图片
视频
新闻
地图
贴吧
十年在线试听QQ音乐
歌手：陈奕迅
专辑：黑白灰
发行时间：2003-04-01
在线试听
QQ音乐
百度音乐
酷我音乐
网易云音乐

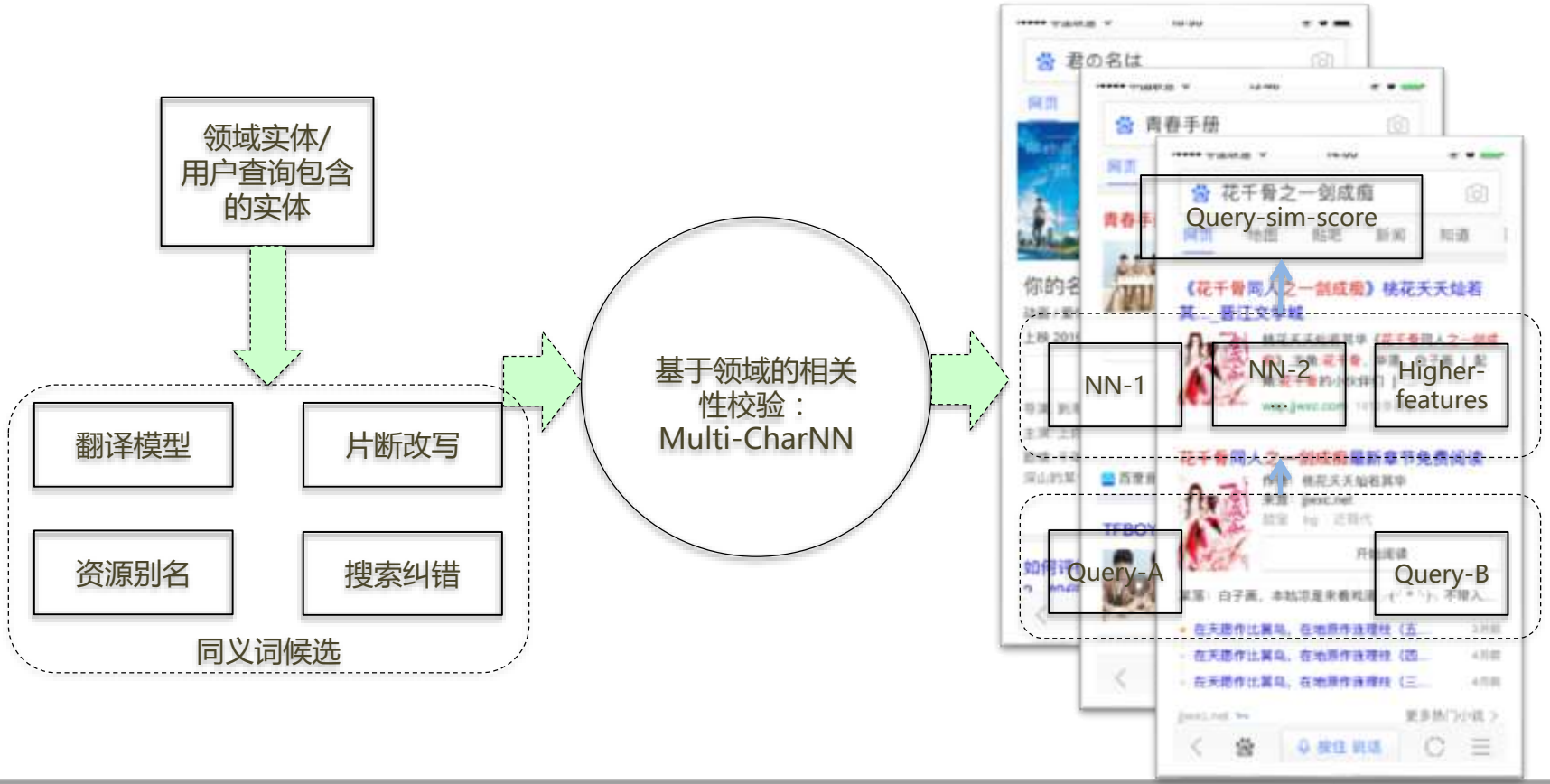
策略概览


标注排序
在线应用
实体标注
序列模型标注
标注消岐择优
同义词替换产出
同义词
标注模型
数据扩充
新词挖掘
实体热度计算
同义词挖掘产出
模型训练
样本迭代修正
标注模型训练
领域消岐模型
同义词校验模型
语料获取
领域实体数据
数据清洗
标注模型样本挖掘
同义词对候选挖掘
核心实体标注

基于深度序列标注模型的迭代



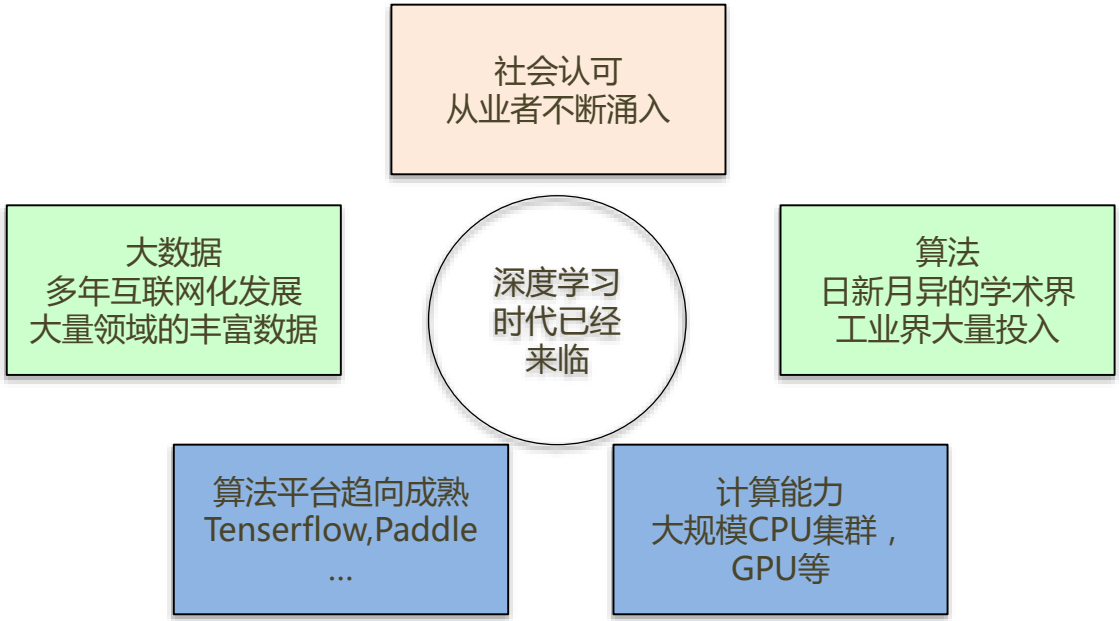
基于领域实体的同义词模型



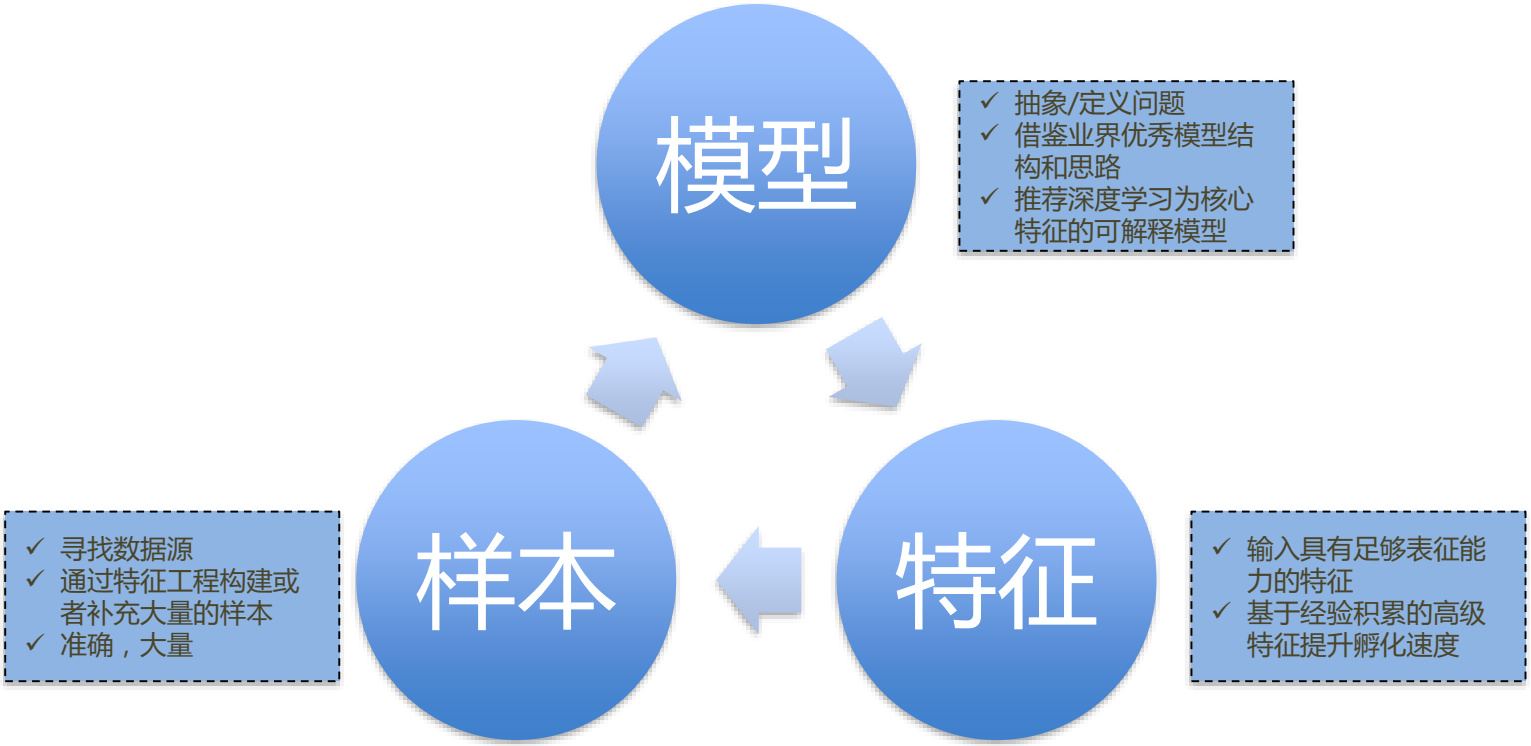
4

如何利用深度学习做更多的策略，做更好的效果

深度学习现状


社会认可
从业者不断涌入
大数据
多年互联网化发展
大量领域的丰富数据
深度学习
时代已经
来临
算法
日新月异的学术界
工业界大量投入
算法平台趋向成熟
Tenserflow,Paddle
...
计算能力
大规模CPU集群，
GPU等

深度学习应用Tips


模型
✓ 抽象/定义问题
✓ 借鉴业界优秀模型结构和思路
✓ 推荐深度学习为核心特征的可解释模型
特征
✓ 输入具有足够表征能力的特征
✓ 基于经验积累的高级特征提升孵化速度
样本
✓ 寻找数据源
✓ 通过特征工程构建或者补充大量的样本
✓ 准确，大量


THANK YOU