DataFunSummit

自然语言处理

在线峰会

信息抽取与检索 论坛

2021.07.10（周六） 09：00~17：30

2021

贝壳 | DataFunSummit

贝壳找房NLP技术在搜索推荐中的应用

袁彬-贝壳找房资深算法工程师

自我介绍

2014 - 2016　　北京工业大学

计算机与科学技术学院 硕士 机器学习方向

2016 - 2017　　新浪微博

短视频推荐优化

2017 - 至今　　贝壳找房

推荐平台从0->1搭建（核心成员）
搜索排序效果优化
首页Feed推荐效果优化

目录

CONTENTS

01 介绍

贝壳业务简介

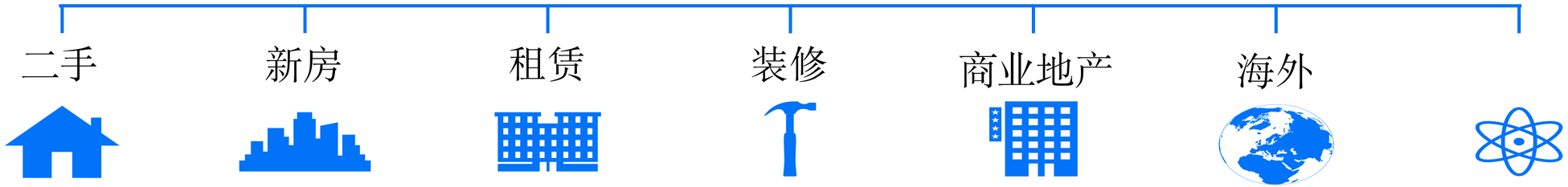
贝壳找房的业务和场景

贝壳介绍

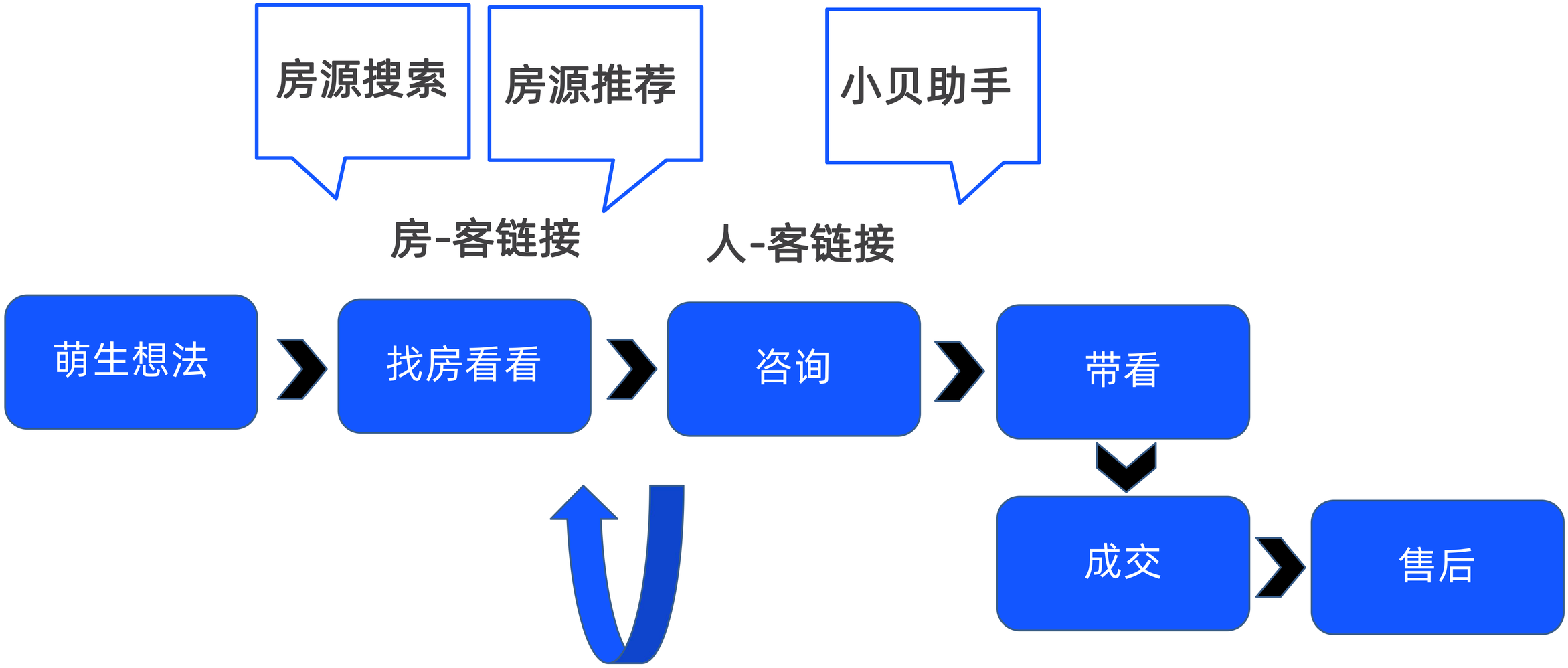
贝壳
找房大平台

科技驱动的新居住服务提供商

多元化的居住服务



用户购房流程



找房场景

结构化需求表达

推荐

搜索

对话式

02
NLU

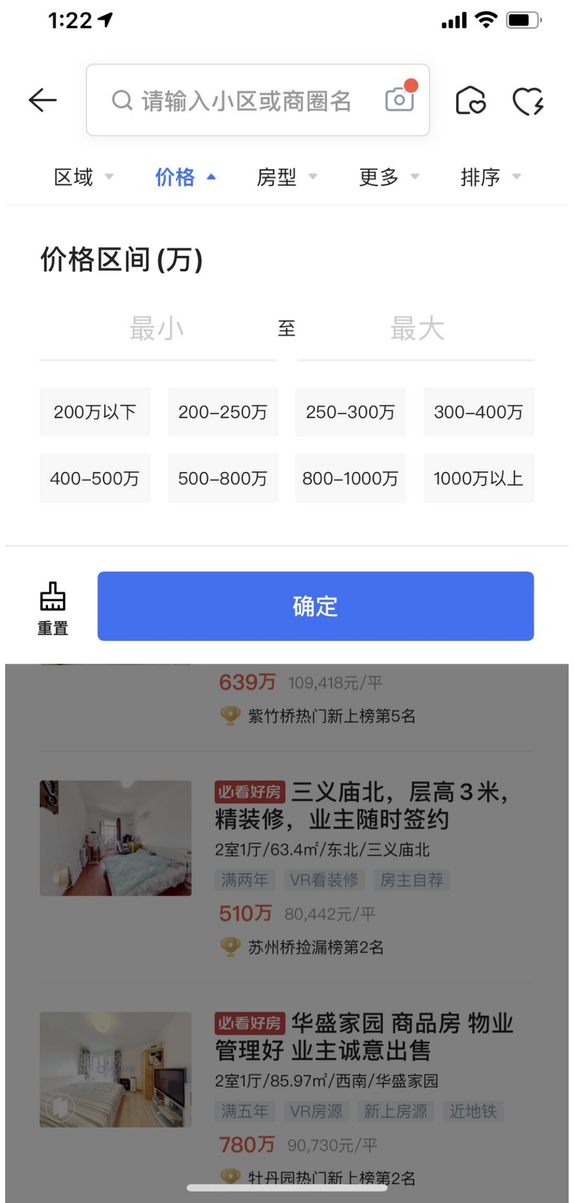
找房中的自然语言理解

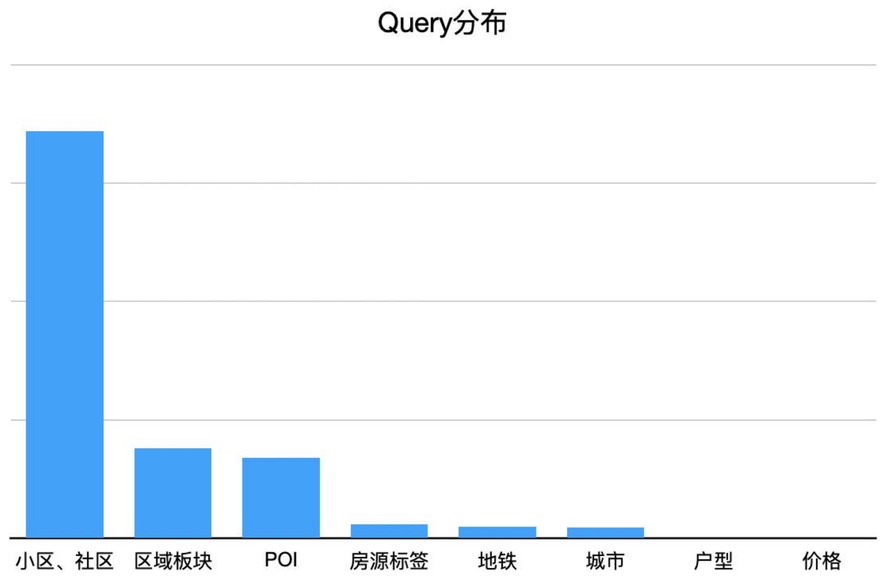
找房业务下NLU的特点

NLU框架介绍

各模块问题拆解与解决方案

特点分析-C端搜索



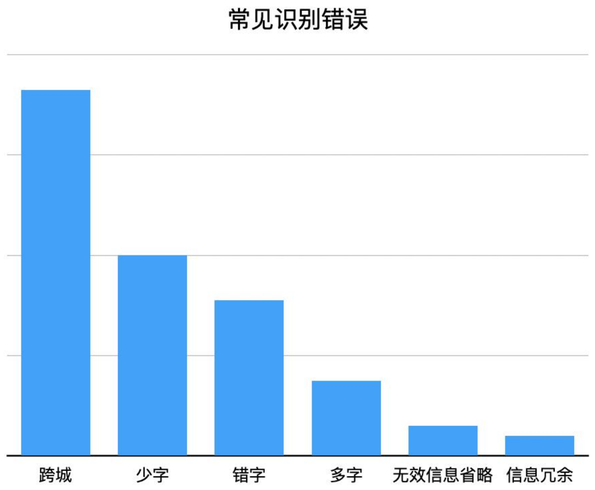


贝壳 | DataFunSummit

特点分析

Query示例

- ✓ 天通苑东一区两室一厅低楼层
- ✓ 十号线牡丹园站附近一居室的房子
- ✓ 北京顺义
- ✓ 亚运新新家园



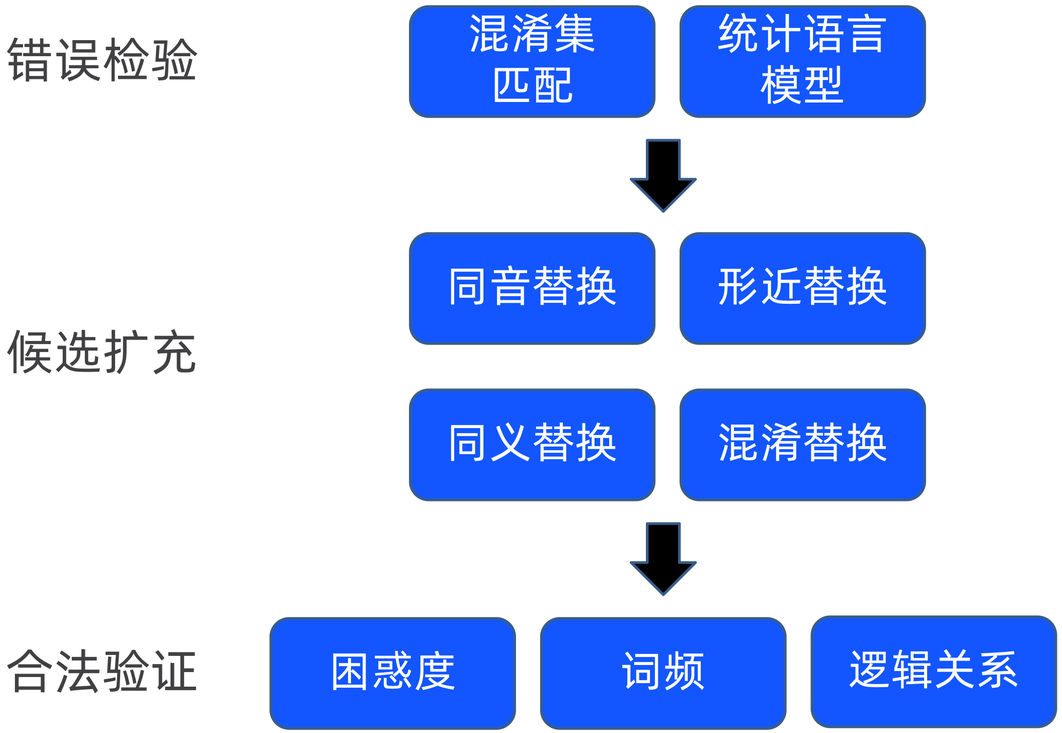
非地址类型（特征明显）

- ✓ 两室一厅、 60平米、朝向
- ✓ 低楼层、满五唯一、带电梯
- ✓ 300万左右
- ✓ Loft、跃层

地理位置（容易混淆）

- ✓ 城市、城区、商圈、板块、小区
- ✓ 地铁站、地铁线
- ✓ 学校、医院、地标、道路

纠错模块



混淆集

- ✓ 词对：平台-阳台
- ✓ 示例如下：带平台的房子->带阳台的房子

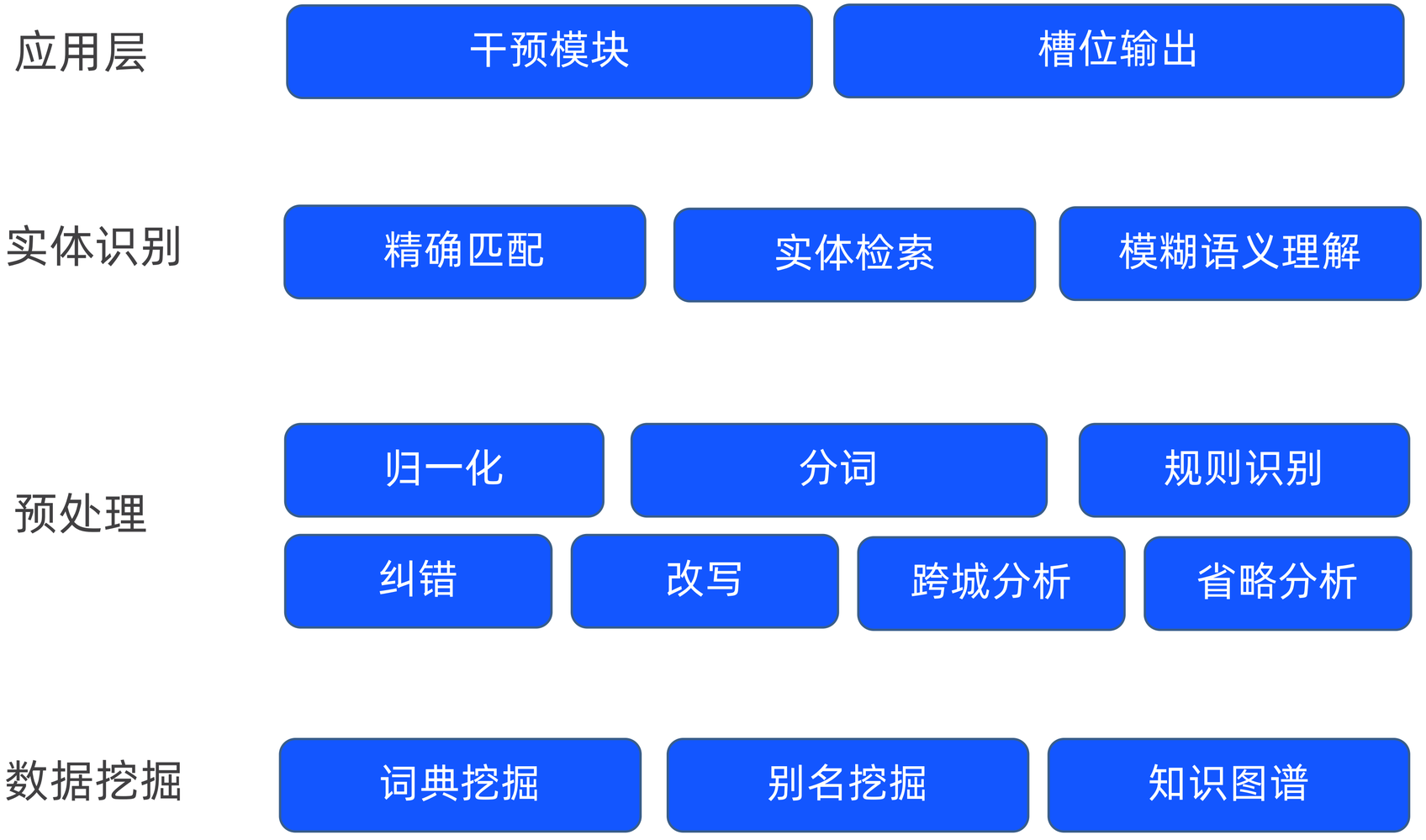
统计语言模型

- ✓ 基于字的 N-gram 模型
- ✓ 检错规则：波动阈值、平均离差

候选扩充

- ✓ 同音替换：远羊山水，错误处“羊”-> 扬阳洋鸯养样痒秧殃氧仰央漾
- ✓ 近形替换：远羊山水，错误处“羊”-> 详样洋鲜群

NLU框架


应用层
干预模块
槽位输出
实体识别
精确匹配
实体检索
模糊语义理解
预处理
归一化
分词
规则识别
纠错
改写
跨城分析
省略分析
数据挖掘
词典挖掘
别名挖掘
知识图谱

改写模块

解决核心问题

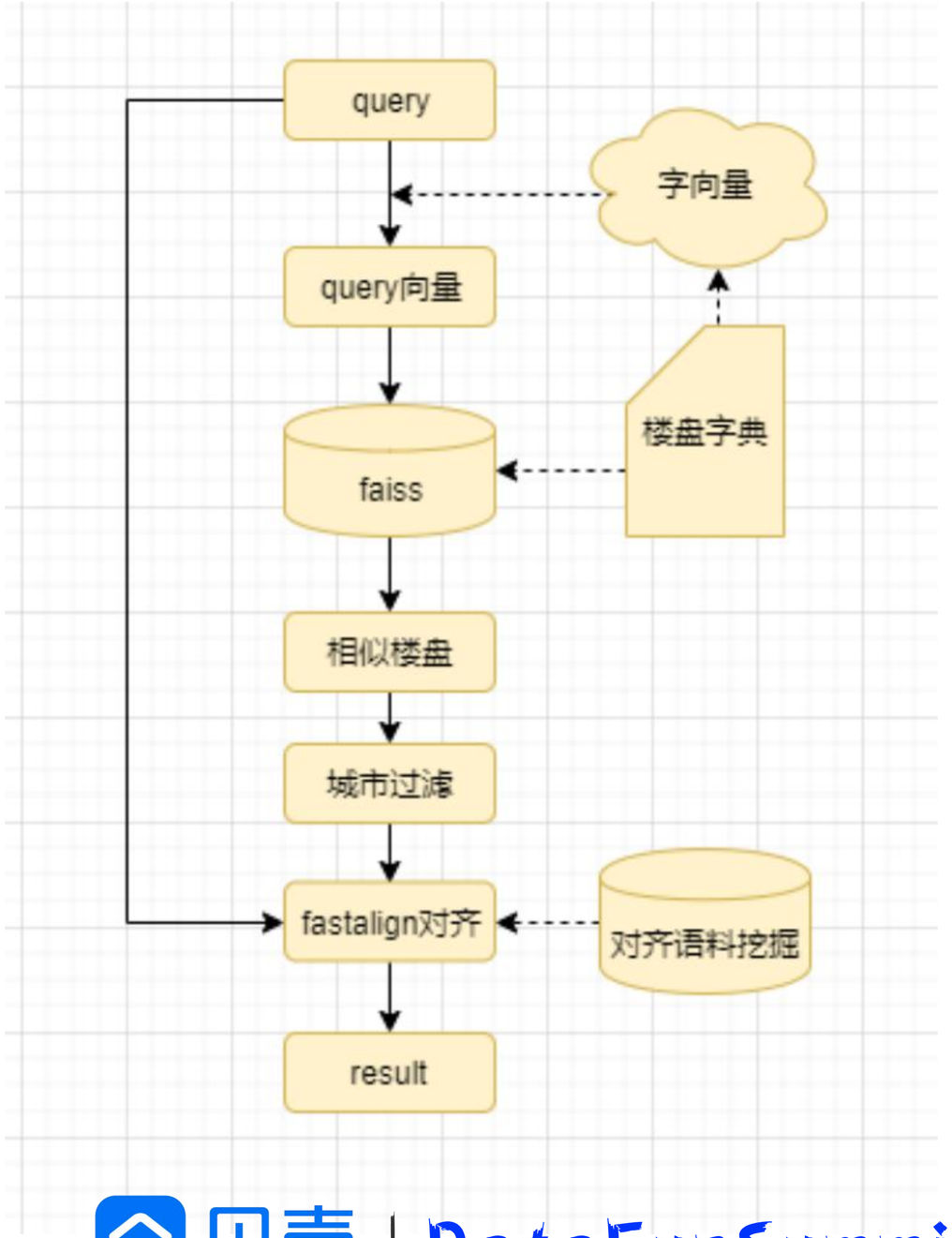
✓ 对少字、多字情况解决方案

少字问题

错误类型	错误query	正确query
少后缀	合正	合正观澜汇
少前缀	海珠湾	星汇海珠湾
少中缀	龙华365花园	龙华美丽365花园
少后字	翰玲珑	翰林珑城
少前字	力洲悦	新力洲悦
少中字	小北小学	小北路小学

多字问题

错误类型	错误query	正确query
多后缀	孔雀城大堡	孔雀城
多前缀	中州锦城湖二期	锦城湖二期
多后字	和平路55号院	和平路55号
多前字	租开元公馆	开元公馆大学毕业生租赁房
多括号	官悦欣园(B区)	官悦欣园B区



实体识别

命名实体识别

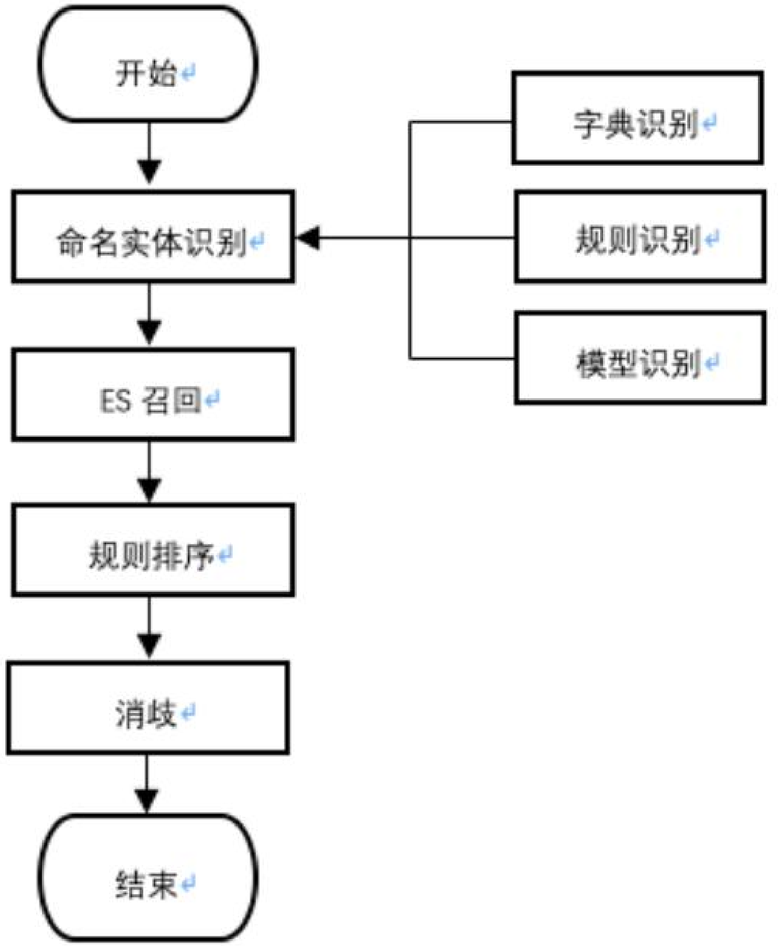
- ✓ BiLSTM+CRF
- ✓ 半监督、模版数据增强

召回

- ✓ 基于bigram拼音的候选实体召回

排序

- ✓ 特征加权：编辑距离、jaccard距离、前缀/后缀、公共子串



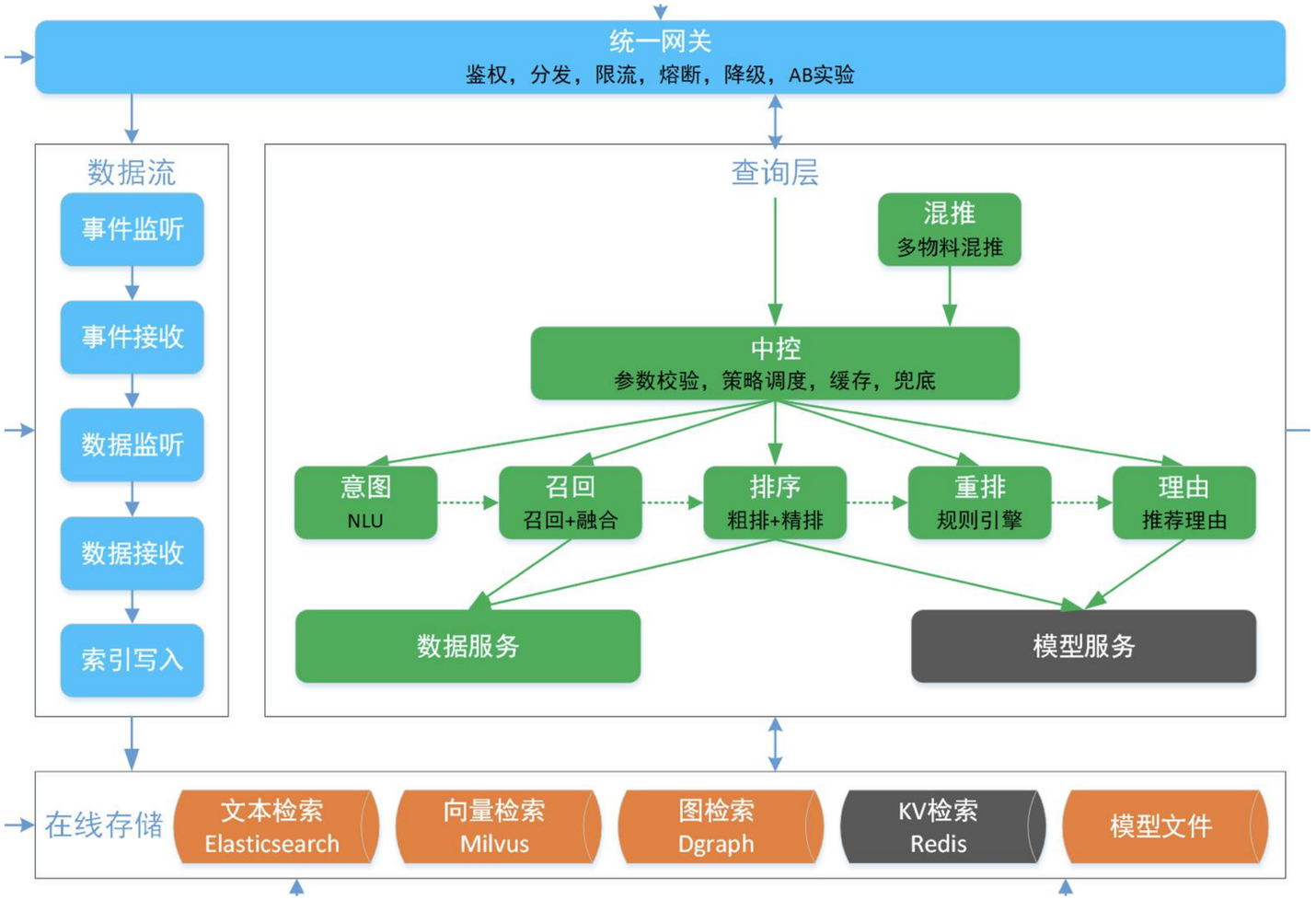
03
应用

NLU在搜索推荐中的应用

C 端搜索中的应用

C 端推荐中的应用

C 端搜索应用



意图理解

- ✓ Query转化为结构化信息
- ✓ 精/泛意图判断

召回

- ✓ 精确意图：相关性约束 + 个性化
- ✓ 泛意图：协同、向量化
- ✓ 扩召回

排序

- ✓ Query理解类特征

C 推荐中的应用

Query1 → Query2 → Query3

Query文本	槽位种类	置信度
天通苑东一区	resblock	0.95
两室	bedroom	1.0
一厅	hall	1.0
低楼层	tag	1.0

Query文本	槽位种类	置信度
天通苑北	resblock	0.95
两室	bedroom	1.0

Query文本	槽位种类	置信度
龙泽家园	resblock	1.0

长期兴趣

✓ 构建用户搜索兴趣

短期兴趣

✓ 实时搜索行为作为trigger

贝壳 | DataFunSummit

THANKS!

今天的分享就到这里...

Ending