

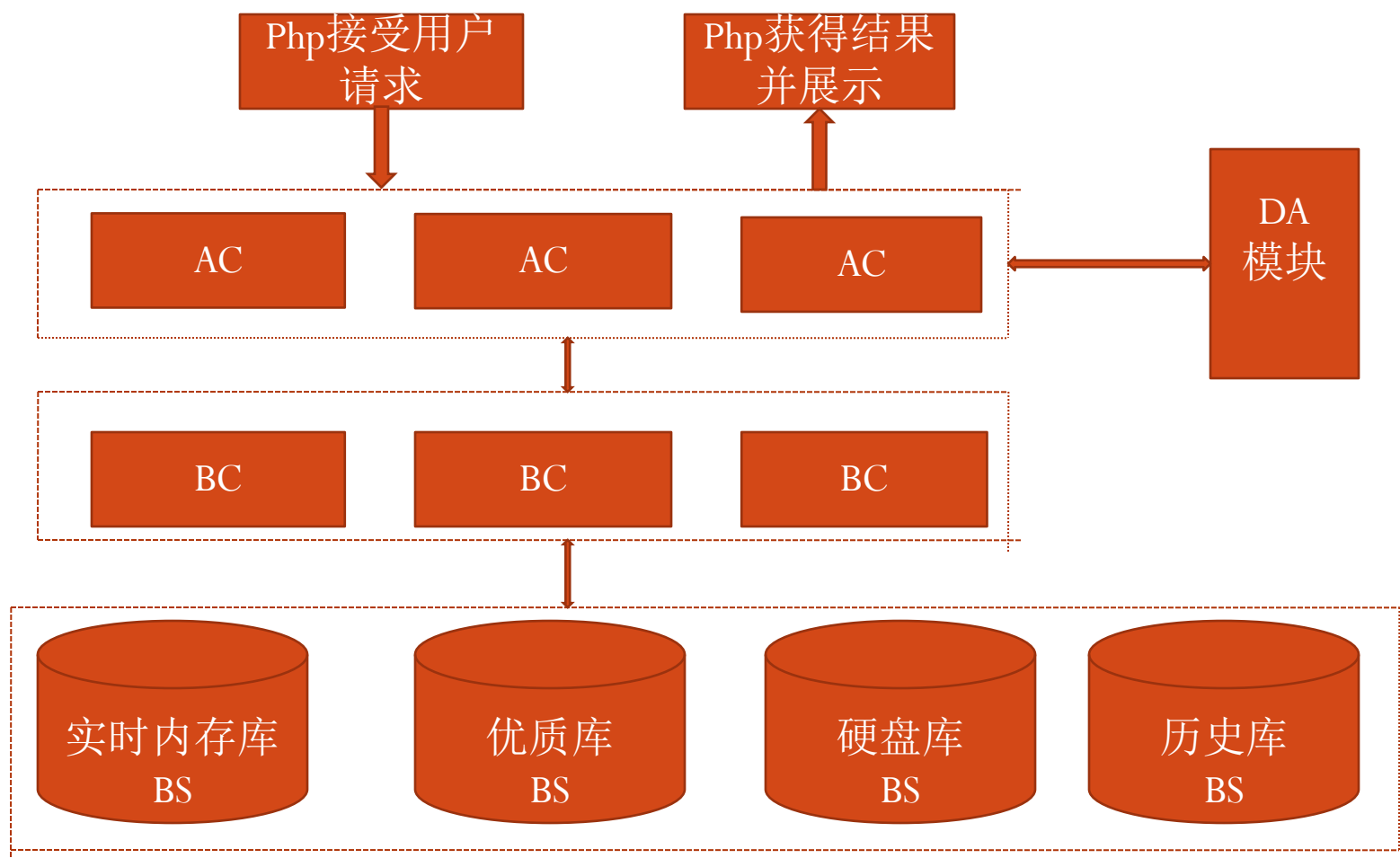
微博搜索相关性分享

杨旭

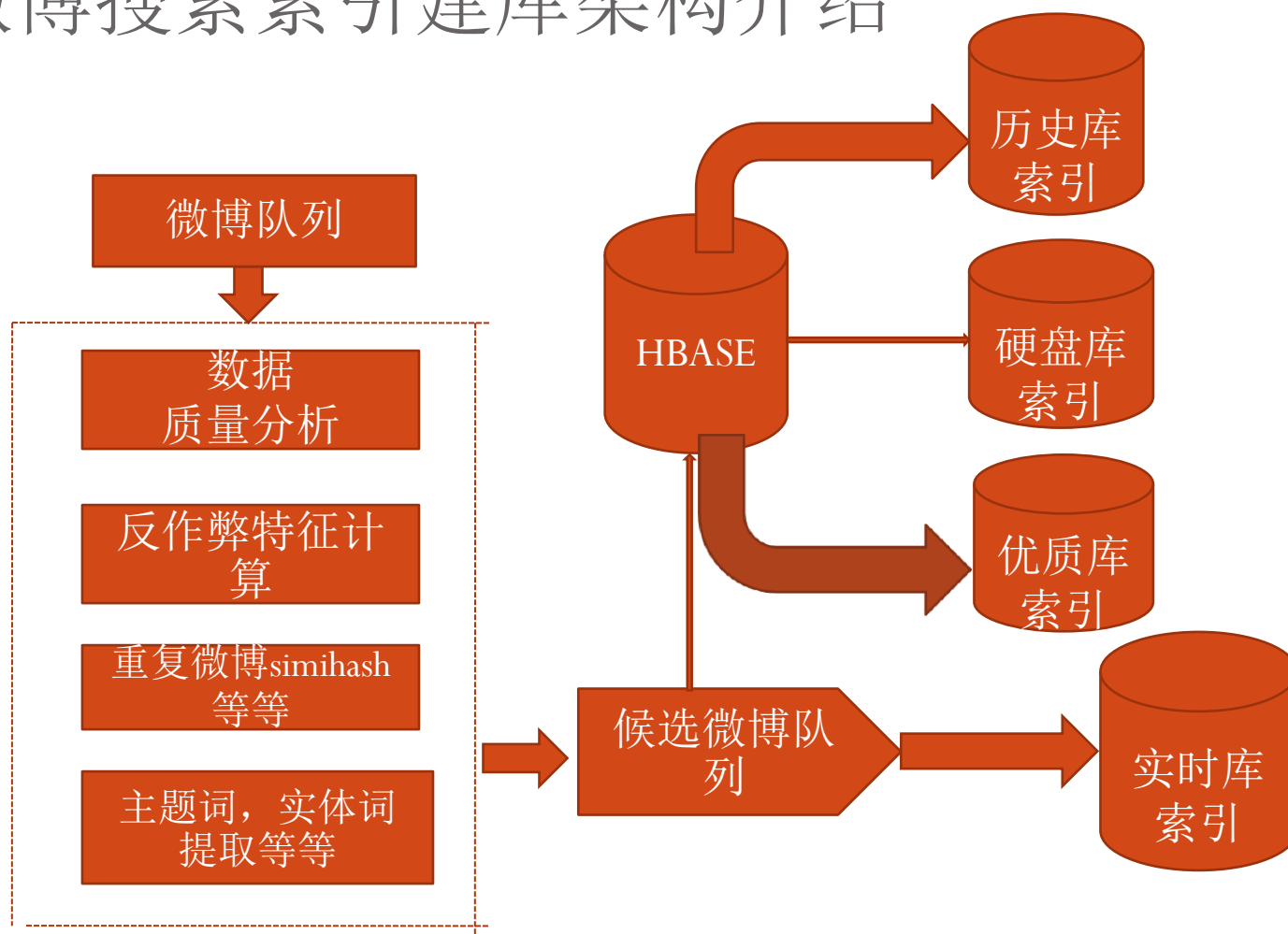
自我介绍

- ◆ 杨旭
- ◆ 毕业时间：2009
- ◆ 院校：东北大学
- ◆ 研究方向：自然语言处理
- ◆ 主要工作经历
 - 阿里商品搜索和查询词推荐
 - 百度图片搜索
 - 微博综合搜索
- ◆ 兴趣领域
 - NLP, 搜索, ranking, 相关性, 广告

微博搜索基本架构介绍



微博搜索索引建库架构介绍



微博搜索相关性

■ 功能

相关性，本质上要解决一个微博是不是符合query搜索意图。

Query: 疯狂动物城



杨洋icon48天团经纪人杨阿毛



★★★★★，既然兔子能当警察狐狸是商人狮子能做总统，所以为什么要纠结狐狸和兔子能不能生孩子呢？😊😊😊其实故事放在真人之中挺狗血的，但是一些观念挺好，比如每个体型不同的动物有专属的火车通道等。总体挺好，喜欢狐狸的死鱼眼哈哈哈哈。

query : 中国好声音第二季



木心花坊

+关注

2014-9-2 14:32 来自 360安全浏览器

某天意外看了一段中国好声音，然后突然想到我是歌手，然后去看了第二季的几段，然后发现还是第一季对我的胃口，然后回去把所有林志炫的演唱重新拿来循

■ 难点

- ◆ 微博一般都是短文本，上下文交叉验证的信息比较少
- ◆ 微博的更新速度快，所以点击反馈信息比较稀疏

■ 效果评价

人工标注

相关性工作

- query分析
- Doc分析
 - ◆ 非核心词
 - ◆ 实体词提取
 - ◆ 主题词
- query, doc相关性匹配

Query分析

- ◆ 分词
- ◆ 词性标注
- ◆ Term重要性
- ◆ 意图识别
- ◆ 紧密度计算
- ◆ 色情敏感词过滤
- ◆ 查询树的生成

便宜的iphone5s手机壳超薄

便宜的手机壳超薄

便宜的iphone5s手机壳

iphone5s手机壳超薄

手机壳 & ((iphone5s & (便宜 | 超薄) | (便宜 & 超薄))

紧密度

◆ 解决的问题

query:joy高跟鞋

#Red Velvet##WENDY# 151026UFO回复 【饭】呃...我知道Joy欧尼是最高的但是呃..所以..你比Yeri矮吗?? 【Wendy】她们穿了高跟鞋!!!!但是我觉得Yeri长高了点...啊..无力的辩解..我们一样高

◆ 紧密度计算

一般都用互信息，结合短语，实体机构词典做补充

互信息是联合分布与乘积分布的相对熵，即

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

◆ 紧密度,字距模型

抓外网结果统计互信息点分布,均值和方差

根据分布和query在content打散命中综合计算一个level

根据结果的分布再做综合调序

◆ 评价

diff数据加上人工评测

实际diff面40%， g:s:b=78:40:32

相关性工作

- query分析

- Doc分析

- ◆ 非核心词

- ◆ 实体词

- ◆ 主题词

- query, doc相关性匹配

非核心词

◆ 为什么要做非核心词

直接判断微博与query的相关程度较难，但是有些微博明显不相关则比较好判断

◆ 非核心词

实体词堆砌

影视，明星，话题词热搜词非核心词

普通昵称中影视，明星非核心词

无关@

◆ 算法

通过搜索结果挖掘相关词，再结合上下文特征再加上一些合理的规则做反向判断
在bs命中非核心词直接丢弃结果

◆ 评价

准确基本都能达到90%，召回80%

实体词策略

◆ 解决什么问题？

query:道士下山

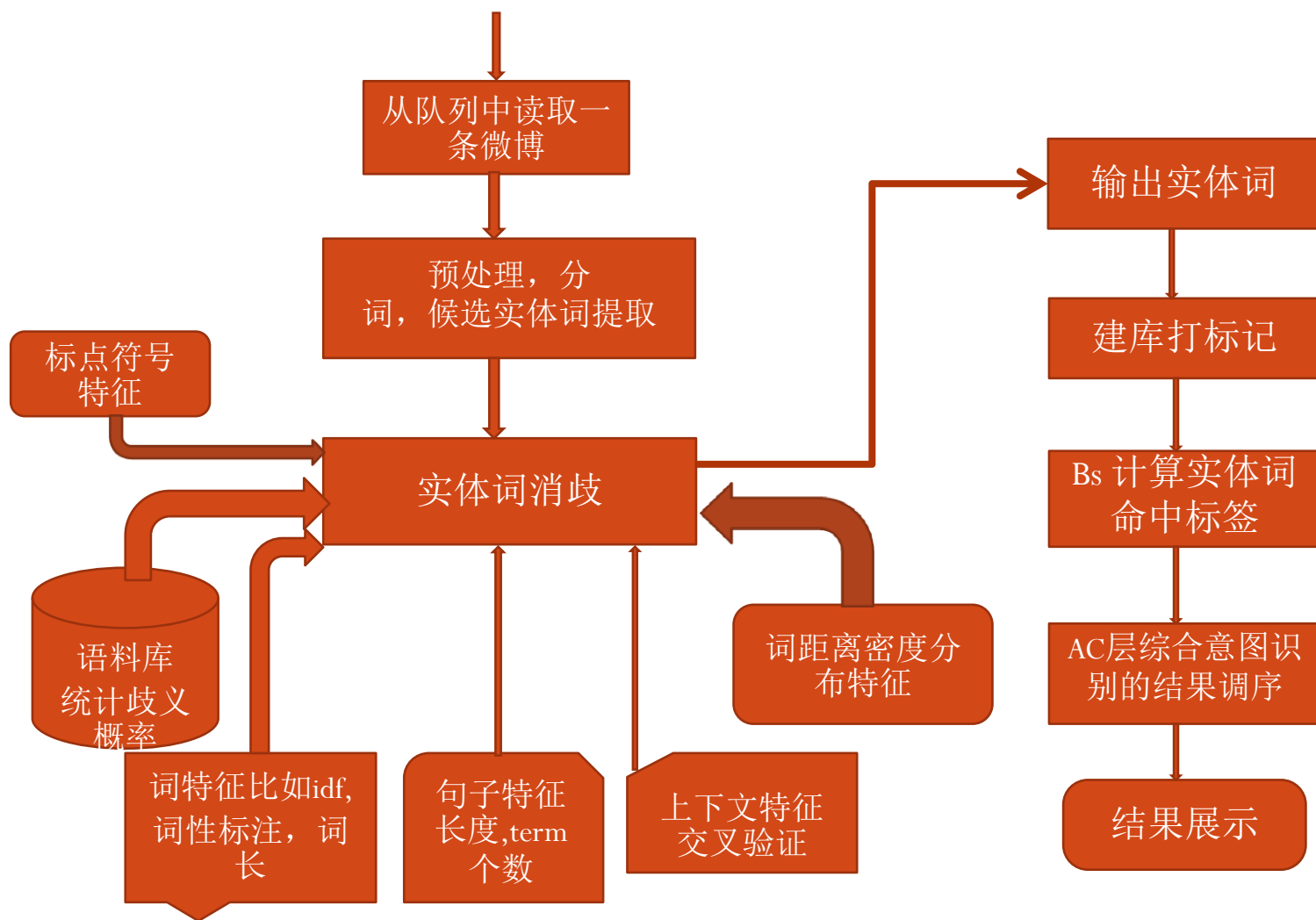
道士下山 ★★★★★, 简单来说, 电影讲述了范伟和吴建豪, 郭富城和张震, 王学圻和王宝强这三对基佬的故事。阴阳交合, 大地吹箫, 好一派生基勃勃。而志玲的主要作用是出演声优! 陈导就是要让大家知道: 今年的基友戏, 被我承包了! <http://t.cn/RLzcNjN>

道士下山吃甜点。来吧, 涨个粉 @Juriveno @开在岩石缝里的一朵花儿li @__大er童 @王小花nice花 @Miss肖肖肖 @西依凹小花 @in酒泉

- ◆ 线下提取微博中的实体, 建库端打上标记
- ◆ 在bs相关性模块打上命中标记
- ◆ AC层根据意图识别的结果, 以及结果分布做调序
- ◆ 效果评价

包含实体词的query做随机样本, 跑目标case, 目标case上评估准确和召回。准确92%, 召回91%

实体词提取以及应用流程





主题词提取

◆ 主题词提取

我们更希望用户的查询词能命中微博的主题。

query: 道士下山

 **道士下山** ★★★★★, 简单来说, 电影讲述了范伟和吴建豪, 郭富城和张震, 王学圻和王宝强这三对基佬的故事。阴阳交合, 大地吹箫, 好一派生基勃勃。而志玲的主要作用是出演声优! 陈导就是要让大家知道: 今年的基友戏, 被我承包了! <http://t.cn/RLzcNjN>

东方卫视番茄台 

[#和动感101拉偶像一把#](#) 第23届东方风云榜强势来袭! 有着为电影《**道士下山**》、电视剧《盗墓笔记》、《古剑奇谭》等多部热门影视剧演唱主题曲的经历, 张杰在过去一年几乎成为主题曲“收割机”, 在本次盛典现场, [@张杰](#) 将首次公开演唱电视剧《何以笙箫默》的主题曲《My Sunshine》, 现场版演绎款款深情。

主题词抽取模型

◆ 训练样本获取

选取点击样本, query, 微博, clickNum>10(会过滤一些色情数据)
按照, caseCade原则负样本

◆ 特征选取

词特征: tf, tf*idf及其相对值, 词性, 词长, 是否实体词(0 or 1), 熵

上下文相关系数: 词和上下文算cos相似度, 相对值

相关词特征: 包含相关词的个数以及权重

标点符号特征: 括号, 中括号, #号

介词特征: 前面是否有介词

搜索特征: 词是否被经常单独搜索, 词在queryLog中作为修饰词或者关键词比率

距离分布特征:

◆ 模型选取以及应用方式

svm

BS检索提权

◆ 评价

交叉验证, 准确率大约85%左右

PM标注, 准确78%, 召回70%.

相关性工作

- query分析

- Doc分析

- ◆ 非核心词

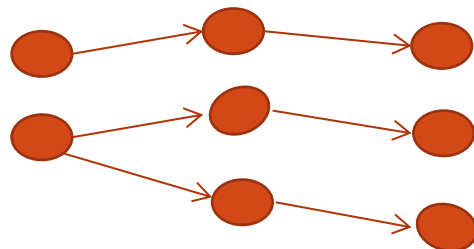
- ◆ 实体词

- ◆ 主题词

- query,doc相关性匹配

query, doc匹配

◆ viterbi算法获得最小匹配距离



◆ 一些中间特征的计算

主题词和非核心词命中

紧密度计算level

query精确命中

实体词精确命中

◆ 检索相关性score的计算

$$\sum_i (q * d) * Score_{tight-punish} * Score_{non-punish}$$

todo

- 文本扩展
- ◆ 相似微博特征交叉利用
- ◆ 转评赞数据的深层次利用
- ◆ 时间, 质量, 相关性的综合平衡

Thanks and QA!