

# 互联网NLP技术及应用

吴华 2013/8/17

# 搜索引擎的用户需求

令人心旷神怡的图片

听起来快乐的歌曲

从中关村到香山饭店怎么走

附近好吃的餐馆

王菲老公代言的汽车品牌

北京哪里能买到漂亮衣服



## 互联网应用趋势

- 基于知识网络的全面需求满足
  - 一知识、语义搜索
    - ●精准需求理解
    - ●直接答案满足
  - 一对话式搜索
    - •智能交互
- 主动推荐,不搜即得
- 精准的个性化服务



# 搜索给NLP带来的挑战

输入复杂

• 需求多样

建议 扩展

交互

来源:暗网、隐含 形式:结构化、半结构 B 化、非结构化 需求识别 知识挖掘 结构:各种层次的知识 框计算 直接答案 聚类 用户引导 结果组织与展现 文摘 关系图 推送 多媒体



## 整合词典、例句、翻译



🚖 收藏 🛮 💹 手机版 🖛 英

请输入要翻译的文字内容或者网页地址 清空

这里所有的衣服都有折扣吗

翻译结果(中>英)

复制结果

Here are all the clothes discount

自动检测语言▼

自动检测语言

中一英

英一中

中一日

日一中

英一日

日 → 英

英一 秦

泰一英

百度翻译

新! 为您提供类似表述,查看示例用法:

- 1. 这里所有的衣服都有折扣吗? Are all the clothes here on sale? www.tingroom.com
- 2. 这里这种衣服所有的尺码都有。 The dress is available in all sizes here. 119.75.215.166

## 移动APP -- 旅游





## 百度翻译免费API -- 电子商务



Bermansa 2012 elderly women's autumn new jacket XL fashion mother decoration body Hooded Coat



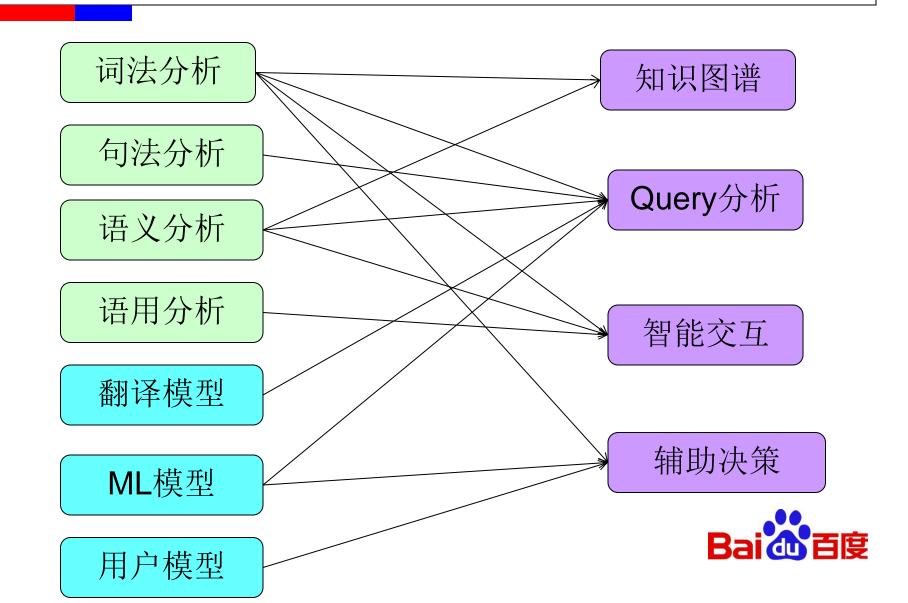
- Total price is final price
- No extra cost available
- 以上翻译结果由Bai → 翻译提供

# 提纲

- ・文本理解
  - **—** Query理解
  - 一知识语义计算
- ●用户理解
  - 一用户建模



# 基础技术与应用的关系示例



# 词法和关系分析应用

实体挖掘

Ontology建设

知识图谱



# 实体挖掘

Bai 面 百度 新闻 网页

- 从Query Log挖掘
- URL pattern 挖掘
- 从title中挖掘

- 中国合伙人
- 中国合伙人电影
- 中国合伙人在线观看
- 中国合伙人百度影音
- 中国合伙人下载
- 中国合伙人原型
- 中国合伙人完整版
- 中国合伙人经典台词
- 中国合伙人影评
- 中国合伙人插曲
- 中国合伙人在线观看完整版

### 相关上下文:

电影 | 在线观看 | 百度影音 | 下载 | 完整版 | 经典台词 | 影评 | 插曲

**校词挖掘** 

鲁求标签

新词/专名分类

可面向自定义类别进行灵活分类

步步惊心 下载 莫言作品集 北京大学录取线 山楂树之恋在线观看

.....

新词/专名挖掘

基于多类资源挖掘新词/专名知识,可做到天级别更新

原始数据

**Query log** 

海量网页库

垂直站点资源

# Ontology建设

上下位知识挖掘

三元组知识挖掘

相关实体挖掘

语义标签挖掘

步步惊心



上位词:小说/电视剧/穿越剧……

导演:李国立编剧:王莉芝

主演:刘诗诗/吴奇隆/郑嘉颖

类型: 历史/魔幻/情感/青春/剧情

年份: 2011 集数: 35

制片地区:中国大陆

.....

相关实体: 01 宫 / 02 宫锁珠帘 / 03 新白发魔女传 / 04 轩辕剑之天之痕……

**语义标签**: 穿越 / 清宫 / **好看** / 清穿 / 刘诗诗 / **养眼** / 雍正 / 虐心······

#### Baidu 百隻 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

类似北京遇上西雅图的电影

百度一下

#### 类似北京遇上西雅图的电影推荐 百度视频





向左走。向右走



海洋天堂 剧情/爱情/伦理 \*\*\*\*\*\*\*\*\* 7.8



月满轩尼诗

### Bai 面 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

凶猛的狗

#### 쏙 您要找的是不是以下结果:



#### 比特犬

#### 中国藏獒大战凶猛比特犬-57445712-我的搜狐

是世界三大斗犬。比特犬见狗就会攻击天生攻击性强,还曾有咬不到狗扭过 头来...即将出战的比特,体重38斤主人正在鼓励<mark>凶猛</mark>好斗有点神...

比特犬 百度百科 | 更多相关搜索结果>>



#### 藏獒

#### 比藏獒凶猛的犬-比藏獒凶猛的狗-比藏獒还凶的狗

藏獒是最<mark>凶猛的</mark>犬种,世界上还有比藏獒更凶的犬吗?目前争议比较大。世界上有高加索犬、中亚牧羊犬、美国比特犬等诸如此类的猛犬,...



## 语义计算

### 三元组知识融合

- •属性名归一化
- 属性值去噪
- •属性值规范化
- 同义实体识别
- 多源数据合并

### 于谦







### 三元组分类

• 判定实体在给 定三元组中的 分类





### 于谦





历史人物

### 三元组相似度计算

• 计算任意两组 三元组之间的 相似度







20%

### 实体消歧

•对于多义实体, 判定其在特定 语境下对应的 三元组





相关人物

于谦

郭德纲

### 主需求识别

•识别用户对给 定实体的主需 求所对应的三 元组









### 知识推理

• 基于已有三元 组推理出未知 三元组



### 于谦



出生日期 年龄 性别(男)+配偶

### Bai 面 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

干谦

百度一下



#### 干谦 百度百科

于谦,著名相声演员,德云社成员,常与著名相声演员郭德纲搭档表演相声。 1982年考入北京市戏曲学校相声班学艺,在校期间曾跟随相声名家王世臣、罗荣 寿、高凤山、赵世忠学习。1985年拜师石富宽... 查看百度百科>>

个人档案 - 人物简介 - 主要作品 - 首登春晚 - 经典语录 - 相关新闻

主需求识别

三元组分类

实体消歧

相关人物

#### 干谦的图片











#### 干谦

代表作品: 《九九归一》;《天下第一丑》;《编辑部的故事》;《天...

部小: 演艺, 相声演员

民族: 满族

生日: 1969年01月24日

星座: 水瓶座

来自百度百科>>

#### 其他人还搜



郭德纲



郭麒麟





李菁



# Query 理解

Query改写- --机器翻译模型

意图归一 --- 机器学习模型

复杂Query 理解 --- Parser



# Query意图理解



# Query理解

• 纠错、同义、省略

天龙八步 —> 天龙八部

成都**的哥**罢工 —> 成都**出租车**罢工

康柏**笔记本**vista**系统一**键恢复—>康柏vista一键恢复

• 一个问题用多种方式表达

姚明身高;姚明几米;

上地在哪里;上地在什么地方

· 相对复杂的query不能理解

身高180以上的自由泳运动员 王菲老公代言的汽车品牌是什么 Query改写

意图归一

句法分析

# 机器翻译模型用于 Query改写



# Query改写模型

### 成都的哥罢工 —> 成都出租车罢工

给定query f, 求其最佳的改写query e\*

$$e^* = \arg \max p(e \mid f)$$
  
=  $\arg \max p(f \mid e)p(e)$ 

$$\hat{e}_{1}^{I} = \arg \max_{e_{1}^{I}} \left\{ \frac{\exp[\sum_{m=1}^{M} \lambda_{m} h_{m}(e_{1}^{I}, f_{1}^{J})]}{\sum_{e_{1}^{I}} \exp[\sum_{m=1}^{M} \lambda_{m} h_{m}(e_{1}^{I}, f_{1}^{J})]} \right\}$$

$$= \arg \max_{e_{1}^{I}} \left\{ \exp[\sum_{m=1}^{M} \lambda_{m} h_{m}(e_{1}^{I}, f_{1}^{J})] \right\}$$

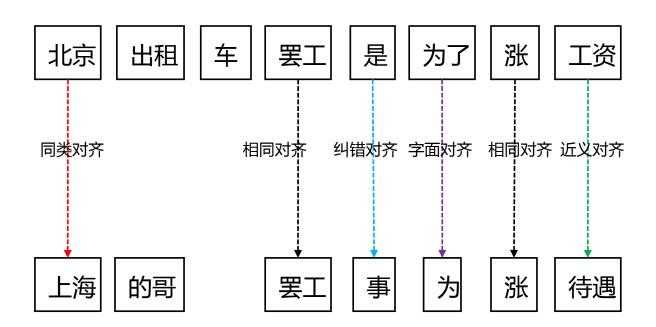
### 特征

- 同义改写概率
- 纠错改写概率
- 上下文改写概率
- 语言模型



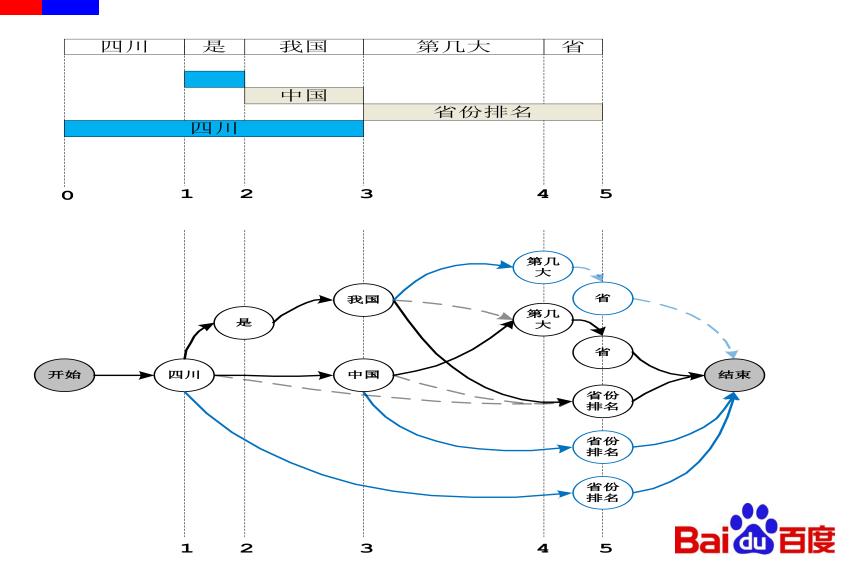
# Query改写

### 词对齐示例





# Query变换算法示例



# 示例

### Baidu 百隻 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

鳗鱼干炒辣椒

### 青椒炒鳗鱼干的做法,青椒炒鳗鱼干怎么做好吃,青椒炒鳗鱼干的家常...

图文 青椒炒鳗鱼干怎么做好吃?青椒炒鳗鱼干的家常菜做法有哪些?好豆网为您提供了青椒炒 鳗鱼干的做法大全,配有非常详细图文步骤,菜鸟也能做出好味道。

www.haodou.com/recipe/196... 2013-7-17 - 百度快照

#### <u>青椒炒鳗鱼干的做法 家常青椒炒鳗鱼干的做法【图】青椒炒鳗鱼干的...</u>

青椒炒鳗鱼干HM.jpg 5.锅里再放油,放青椒煸炒 青椒炒鳗鱼干Zn.jpg 6.直至青椒变软后放盐再炒 青椒炒鳗鱼干kT.jpg 7.放鳗鱼干同炒 青椒炒鳗鱼干HC.jpg...

www.meishij.net/zuofa/qingjiaochaoma... 2013-7-30 - 百度快照

#### 辣炒鳗鱼 矫健 新浪博客

图文 将鳗鱼收拾干净后切成小段<mark>青椒</mark>切同鳗鱼长度备用 鳗鱼段先用滚水烫至变色后立即捞出 控干水分备用 油热后放入葱姜 蒜爆香 烹入老抽白糖加入<mark>青椒翻炒</mark>均匀后放入鳗鱼...

blog.sina.com.cn/s/blog\_4fab6c1b0100... 2013-7-20 - <u>百度快照</u>



# 机器学习模型用于 Query意图归一



# 一个Query有多种表达

哈文的丈夫是谁
谁是哈文的丈夫
哈文的丈夫是哪个
哈文的丈夫叫什么名字
哈文的丈夫的名字
哈文的男人是谁
哈文丈夫叫什么
哈文的丈夫是哪位
哈文丈夫是谁
谁知道哈文的丈夫的姓名
哈文的老公叫什么
哈文丈夫名字
哈文的老公
哈文的丈夫名字
哈文的丈夫
哈文的老公叫什么名字
哈文的老公是

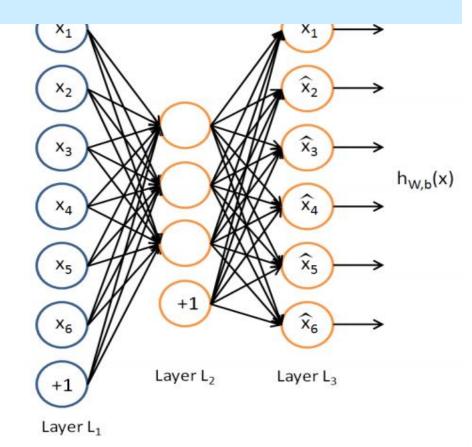
氮气的化学性质		
氮气有哪些化学性质		
氮气化学性质怎么样		
氮气的化学性质是什么		
氮气有什么化学性质		
氮气化学性质活泼吗		
氮气的化学性质活泼么		
氮气化学性质表现为什么		
氮气化学性质如何		
氮气通常化学性质是什么		
谁知道氮气的化学性质怎么样		
氮气化学性质		
氮气具备什么化学性质		
氮气的化学性质活泼吗		
氮气化学性质是什么		
氮气的化学性质怎么样		
氮气化学性质包括什么		
氮气属于什么化学性质		
氮气化学性质是否活泼		
24 TIGT II/A/E II /II //X	Baid	古庙

# Query Pair文本语义匹配计算

# Query Pa

## 确率为90%,比baseline绝对值高10%

- 一 姚明有多高 姚明的身高
- 一 谁写的《平凡的世界》 《平凡的世界》作者是谁
- 方法
  - -DNN算法
- 产出
  - 一词表示
  - Query Pair 相似度



Query term vector 1





## 为什么DNN有效

### 简单表示

"dog"=
$$[1,0,0,\ldots,0]$$
,
"cat"= $[0,1,0,\ldots,0]$ 
"the"= $[0,0,0,\ldots,1]$ 

### 聚类

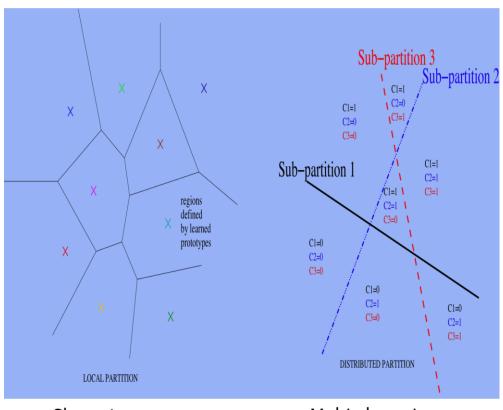
'' dog'' = [1, 0, 0, 0]

"cat" = [1, 0, 0, 0]

"the" = [0, 1, 0, 0]

## **Embedding**

"dog" = [1, 0, 0.9, 0.0]"cat" = [1, 0, 0.5, 0.2]"the" = [0, 1, 0.0, 0.0]



Clustering

Multi-clustering



# Parser技术用于复杂 query解析



# 复杂query解析

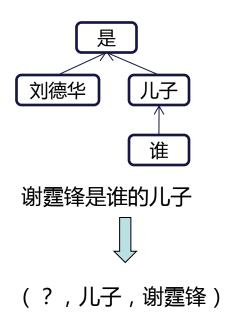
- 谢霆锋是谁的儿子?
- 谢霆锋的儿子是谁?
- 王菲老公代言的汽车品牌是哪个?

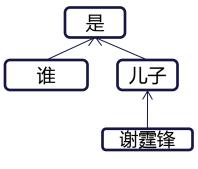




# 复杂query解析

- 谢霆锋是谁的儿子?
- 谢霆锋的儿子是谁?





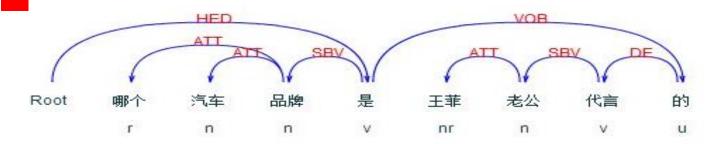
谁是谢霆锋的儿子



(谢霆锋,儿子,?)



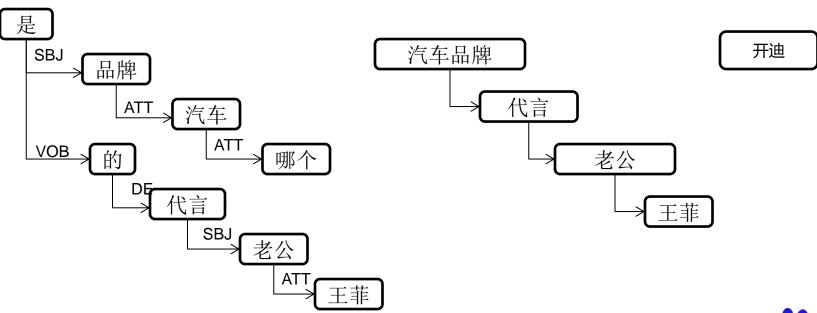
# 复杂query解析



### 依存分析结果

### 查询逻辑表达式

### 知识库推理



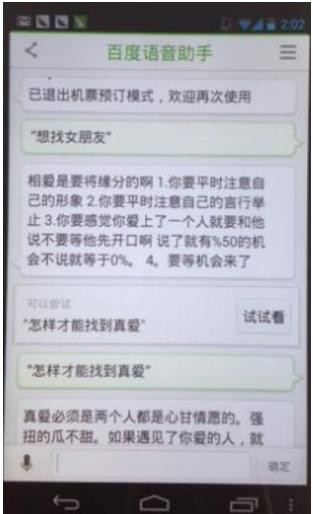


# 语用分析用于智能交互



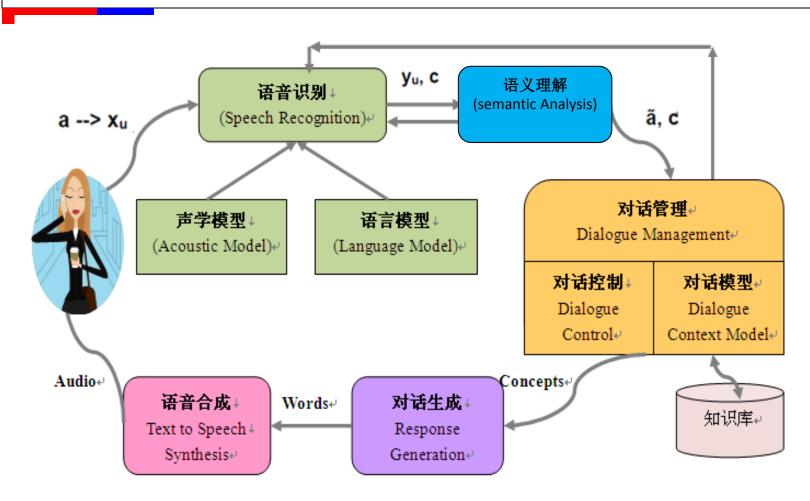
## 语音助手







# 对话管理



a: 用户对话行为; c: 置信度评估↓

Xu: 用户语音输入信号; Yu: 语音识别结果↓

**ā**: 系统估计的用户对话行为↓



# 以订航班为例

### 我想飞伦敦

**对话管理**→ Dialogue Management→

对话控制

Dialogue Control 对话模型

Dialogue

状态集合: S

动作结合: A

状态转移: S×A->S'

R: 回报函数R(i) or R(i,a)

**元**:对话策略 S -> A

### 系统在如下行为中选择:

- 确认: "您想飞伦敦?"
- 问其他问题: "您从哪飞?"
- 查数据库:"去伦敦的飞机"

• .....



## 一些对话管理的任务

- 处理来自于ASR和SLU结果的不确定性
  - 一确认置信度低的信息
  - 根据置信度、信息重要性、对话历史信息等采用不同确认方法
  - 一维护不同可能状态空间
- 确定下一步行动
  - 一要用户提供新的信息
  - 一 放松query的限制条件
- 选择系统信息用语



#### 对话管理模块

- 对话控制
  - 一选择下一步系统行为,比如
    - 要用户提供更多的信息
    - 确认或交互用户之前的输入
    - 给用户提供信息
- 对话环境建模
  - 一维护与对话相关的信息,比如
    - 已经说过的信息
    - 这个信息是否已经交互过?



#### 对话控制策略

- 基于框架的控制策略
  - 一框架定义完成指定任务需要的信息
  - 一根据已知信息,确定之后要问的问题

Slots Values Questions

destination: London

origin: Paris

date: what date?

time of departure: what time do you want to depart?



#### Siri 中的示例

#### 发Email

U: Send an email 用到的框架:

S: who are you sending to?

U: To Mike Reciever

S: Email title?

U: meeting Title

S: What do you want to send to Mike?

U: We have a meeting at seven PM.... Content

S: Here are the contents you want to send to Mike

Should I send out to Mike?

Confirmation

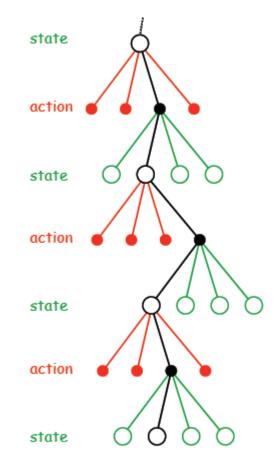
U: Yes

S: It was sent out



#### 更加复杂的对话管理方法

- 基于AI的方法
  - Information State Update Theory
  - Dialogue as Planning
  - Ontology based dialogue
- 统计方法
  - 状态空间S、动作集A、目标G
  - 一方法
    - 马尔科夫决策过程
    - 增强学习(reinforecement learning)



$$Q(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a')$$

# 用户理解

主动推荐,不搜即得

个性化服务



## 什么是用户模型



#### 基于特征选择&参数选择的用户消费判定模型



用户搜索 点击日志



调整消费意图特征空间



训练参数设定



数据抽取模型训练



特征 选择



预测结果内测评估



最优消费 判定模型



获取模型最优参数组



# 消费示范







市场 格局 导向

### 情感分析 - 辅助消费决策

我的収验 | Wu\_h

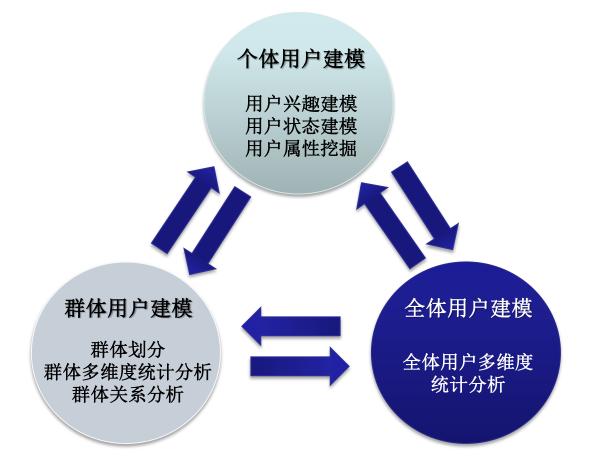


iphone4 <mark>百度一下</mark>



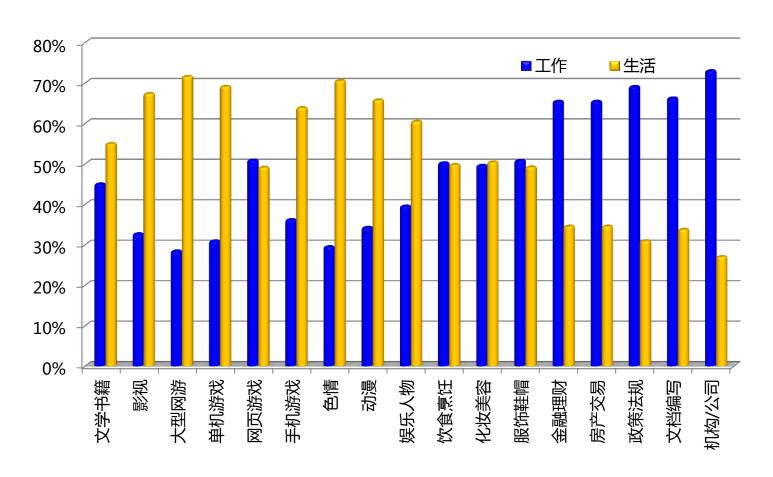


#### 多维度的用户建模



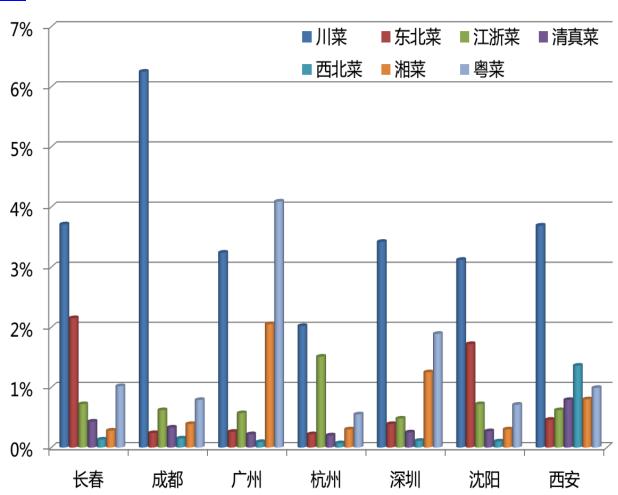


## 多维度用户建模—需求人群vs时间





## 多维度用户建模—菜系vs地域





主题

🛗 时节

3 地域

八 类型

蜜月 情侣 亲子 背包客 文艺 独行 白驾 摄影 美食 购物 潜水 滑雪 徒步 登山 温泉 骑马 常花 漂流 动物 游署 宗教 骑行 历史 民俗 艺术 泡吧 SPA 游轮 家庭 艳遇 毕业旅行。 夕阳红 踏青 采摘

不限 端午 5月 6月 3月 4月 7月 1月 2月 8月 9月 10月 11月 12月 夏天 秋天 春节 冬天 元旦 洁明 中秋 情人节 七夕 光棍节 暑假 圣诞节

国内 东南亚 东亚 南亚 中西亚 欧洲 大洋洲 北美洲 南美洲 非洲

城市 古镇 乡村 海滩 草原 森林 濕布 湖泊 湿地 河流 峡谷 雪山 冰川 沙漠 雪地 山 人文 地质 白然 休闲

#### 旅游推荐

#### Ⅰ 最推荐的 五一 目的地

热门

人少好玩

. 用边游

以下目的地,均由百度高精技术从海里游记、搜索行为中挖掘而出,真实的才是最可信的!

上海 中国



西安 陕西



^

■目的地 ▶景点 🖃 🖺



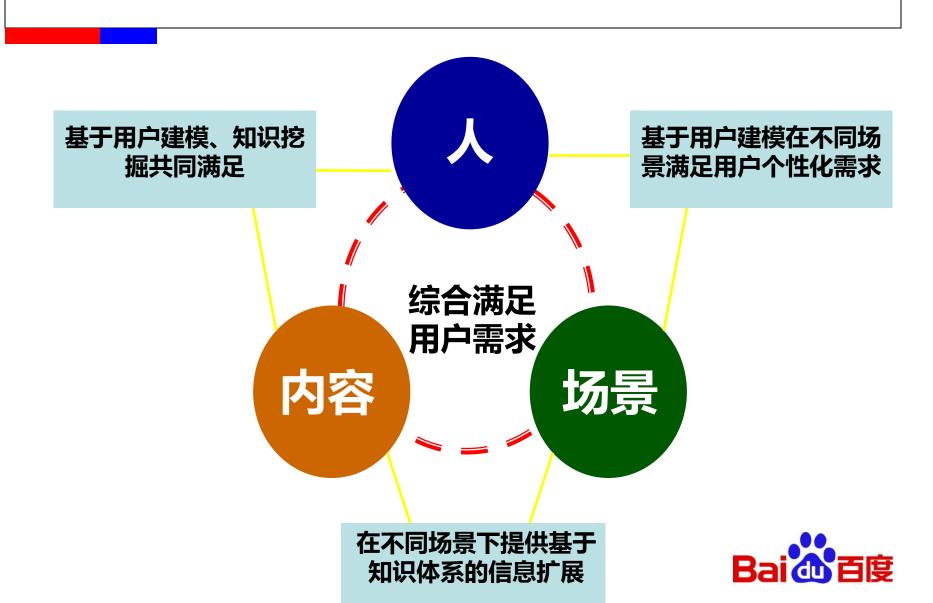
目的地



□ # 游记

\*\*\*\*

#### 人、内容、场景的综合



## 如果互联网是个金矿

## 那么NLP是炼金术



### 百度技术沙龙









关注我们: t.baidu-tech.com

资料下载和详细介绍:<u>infoq.com/cn/zones/baidu-salon</u>

"畅想•交流•争鸣•聚会"是百度技术沙龙的宗旨。 百度技术沙龙是由百度与InfoQ中文站定期组织的线下技术交流活动。目的是让中高端技术人员有一个相对自由的思想交流和交友沟通的的平台。主要分讲师分享和OpenSpace两个关键环节,每期只关注一个焦点话题。

讲师分享和现场Q&A让大家了解百度和其他知名网站技术支持的先进实践经验,OpenSpace环节是百度技术沙龙主题的升华和展开,提供一个自由交流的平台。针对当期主题,参与者人人都可以发起话题,展开讨论。

#### InfoQ 策划·组织·实施

关注我们:weibo.com/infoqchina