搜狗图片搜索系统智能化演进之路

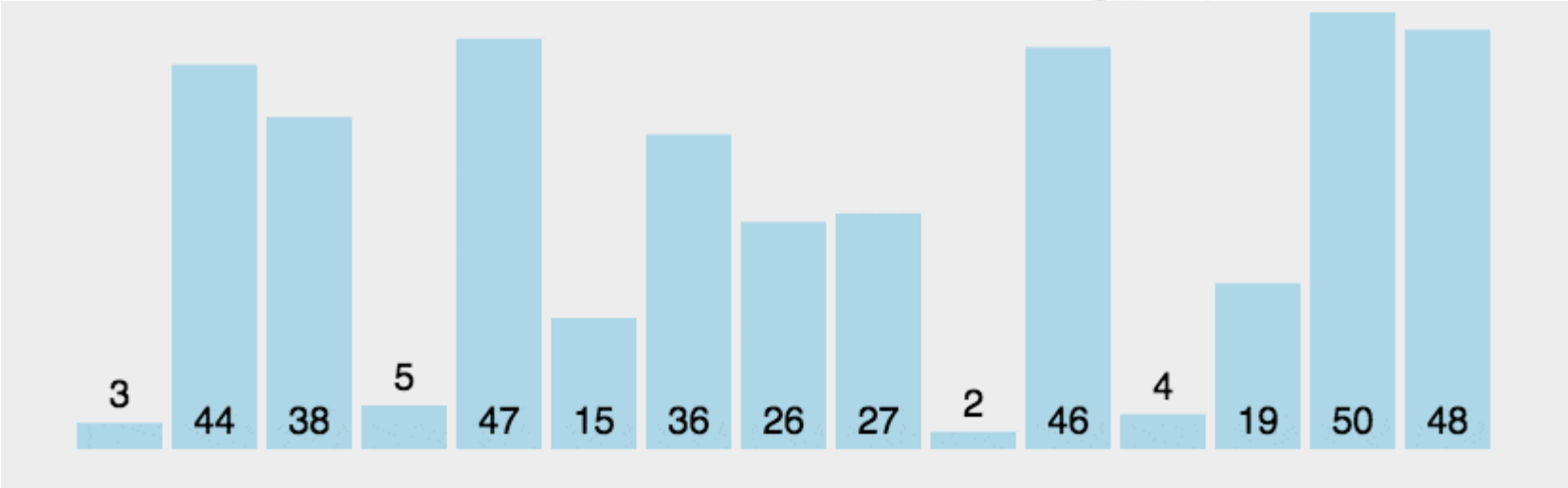
周泽南 搜狗专家研究员

2017年 8月5日

大纲

- 排序
- 图片搜索中的排序
- 搜狗图片搜索中的机器学习技术

排序



排序目标： 请从小(大)到大(小)排序
排序对象： 数值
对排序目标和排序对象都有非常清晰的理解

图片搜索排序

图片搜索：满足用户查找图片的需求

图片搜索排序

图片搜索：满足用户查找图片的需求

图片搜索排序

Query：

图片搜索排序

Doc：

数据积累

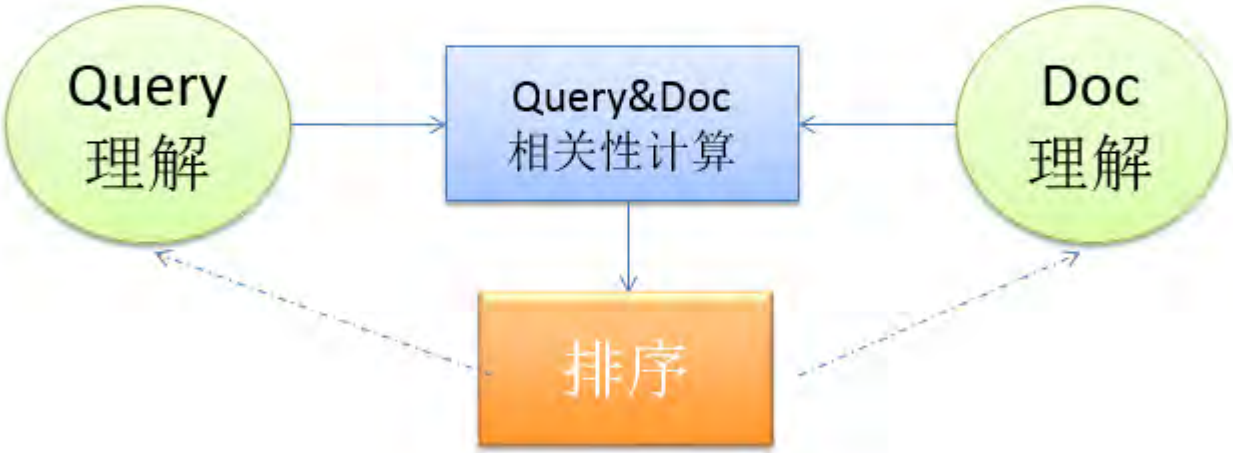
```
(T_P1_TITLE_BODY_)TITLE_(40):华仔默认朱丽倩　乖乖地你们不要再问啦
(T_CLICK_)AUTHOR_(0):
(T_P1_TITLE_DESC_)ANCHOR1_(8):刘德华
(T_P1_ALT_)ANCHOR2_(0):
ANCHOR_EXTEND_(0):
STRIP_URL_(0):
(T_ENTITY_)KEYWORD_(0):
(T_QUERY_GG_)METAINFO_(0):
(T_P1_TITLE_HTML)CONTENTTITLE_(38):华仔默认朱丽倩　乖乖地你们不要再问啦
(T_CLUSTER_TERM_H_)TOPIC_(0):
DESC_CONTENT_(0):
(T_P1_SURR_)CONTENT_(234):“华仔”刘德华上周六以“姐夫”身份专程赴吉隆坡
出席朱丽倩胞妹朱丽华的婚宴，他牵着新娘的照片曝光，让他难以狡辩，两人特殊
关系趋于明朗化，昨天在港出席一项活动时被媒体大逼供，他以微笑见招拆招：“
乖乖地不要再问啦！”一切尽在不言中。
(T_CLUSTER_TERM_L_)CONTENT_RANK_(0):
(T_P1_CRUMB_)BREAD_CRUMB_(32):中国娱乐网；明星；桃色；正文；
```

图片搜索排序

Query-Doc相关性排序：

Rank

图片搜索排序



图片搜索排序

Query理解：

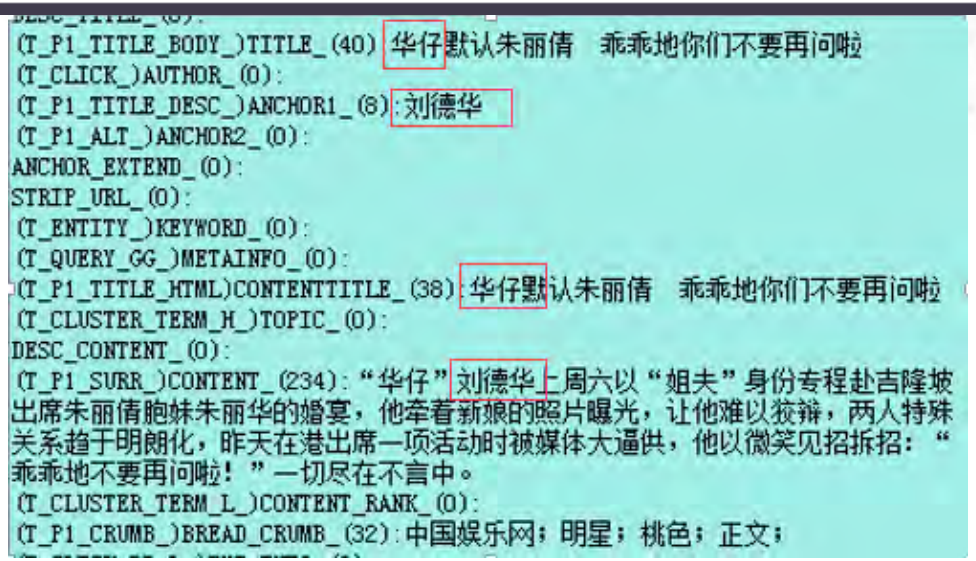
分词、去词、同义词、词重要度、二次查询。。。。

Doc理解：

页面解析、关键词提取、topic、分类。。。

相关性计算：计算Query和Doc相关性

Query与Doc各个域的文本相关性。。。



搜狗图片搜索中的机器学习技术

- 相关性特征
  - 数据积累
  - 排序模型

相关性特征

人工特征

BM25、MatchRank、TitleRank、ContentRank等

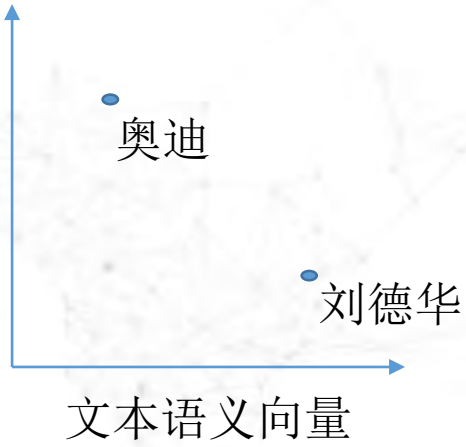
学习特征

- 查询词-图像相关性特征
- 查询词-站点相关性特征
- 先验知识融入相关性计算
- 文本匹配相关性特征
- 文图对应特征
- 结合图像特征的关键词提取

查询词-图像相关性特征

奥迪

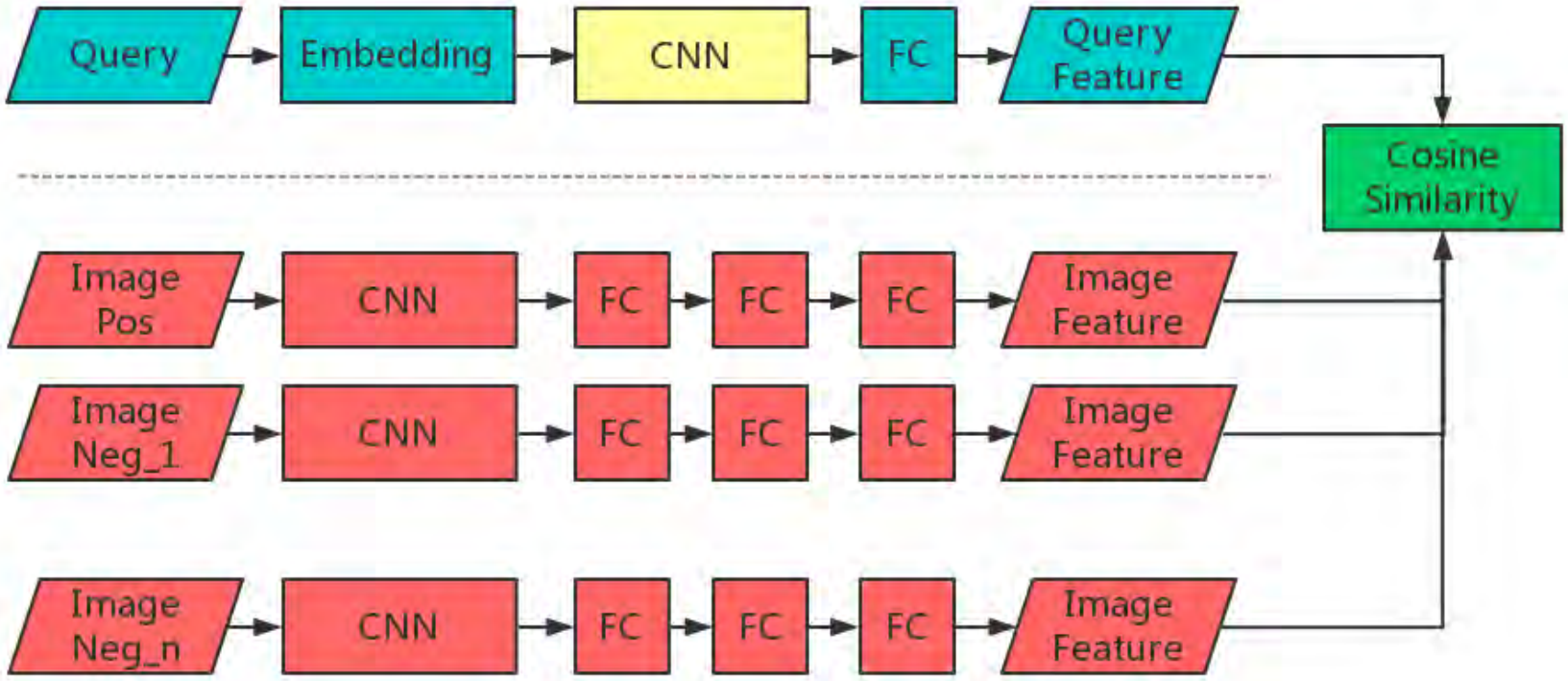
查询词-图像相关性特征



Cosin(func(v(奥迪)，v()) >> Cosin(func(v(刘德华)) ，v())

- 如何得到query文本以及pic的语义向量？
- 如何得到func()?

查询词-图像相关性特征


Query
Embedding
CNN
FC
Query Feature
Cosine Similarity
Image Pos
CNN
FC
FC
FC
Image Feature
Image Neg_1
CNN
FC
FC
FC
Image Feature
Image Neg_n
CNN
FC
FC
FC
Image Feature

查询词-图像相关性特征

query 三毛经典语录 Search

query: 三毛经典语录

三毛(1) 经典(1) 语录(1)

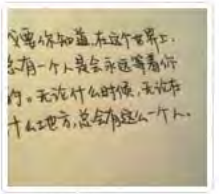
三毛|157 经典|84 语录|125


0.975288

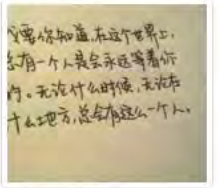
0.968754


0.96215


0.962094


0.958758


0.958646


0.958625

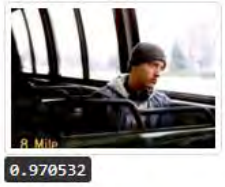
查询词-图像相关性特征

query 刘德华汽车 Search

query: 刘德华汽车

刘德华(1) 汽车(1)

刘德华|177 汽车|152



上线碰到的问题

- 在线计算：
  - 前向计算：线上对个doc进行计算，每个Doc都需要1次前向计算，计算量巨大，需要GPU集群
  - 相似度计算：2个200维的向量计算余弦距离
- 做进数据：
  - 一个doc有200维float，需要200 * 4 = 800(byte)，而DocInfo一共只有338byte，存储开销太大

解决方法：hash(PCA-ITQ)

	存储	计算
Hashing前	200*4*8=6400bit	200d的float vector计算余弦相似度
Hashing后	**64bit**	**64bit的01-vector计算海明距离**

查询词-图像相关性特征

效果：单特征NDCG@10:0.7
NDCG@10提升3%

查询词-站点相关性特征

简笔画视频教程


蝙蝠简笔画视频教程


长颈鹿简笔画视频教程

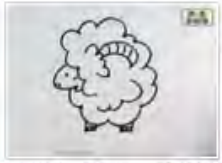

小刺猬简笔画视频教程


海豹简笔画视频教程


鳄鱼简笔画视频教程


绘制的简笔画工具和重


荡秋千的女孩和漂亮的


卡通小绵羊简笔画视频


卡通女孩简笔画视频教


卡通人物画家简笔画视

简笔画图片教程

- 动物简笔画画法教程
- 动力滑翔伞简笔画图片教程
- 乒乓球拍简笔画图片教程
- 蝴蝶风筝简笔画图片教程
- 降落伞简笔画画法图解
- 双翼老式飞机简笔画图片教程
- 螺旋桨飞机简笔画图片教程
- 客机简笔画图片教程
- 热气球简笔画图片教程
- 小车简笔画图片教程
- 旱冰鞋简笔画图片教程
- 滑冰刀鞋简笔画图片教程

人物简笔画推荐

中国人物简笔画　雷锋简笔画


小朋友植树简笔画图片


小朋友洗澡简笔画图片


小朋友喂鸡简笔画图片


小女孩跳绳简笔画图片


小朋友踢足球简笔画图


小朋友上课简笔画图片


上班族简笔画图片、教


小朋友扫落叶简笔画图

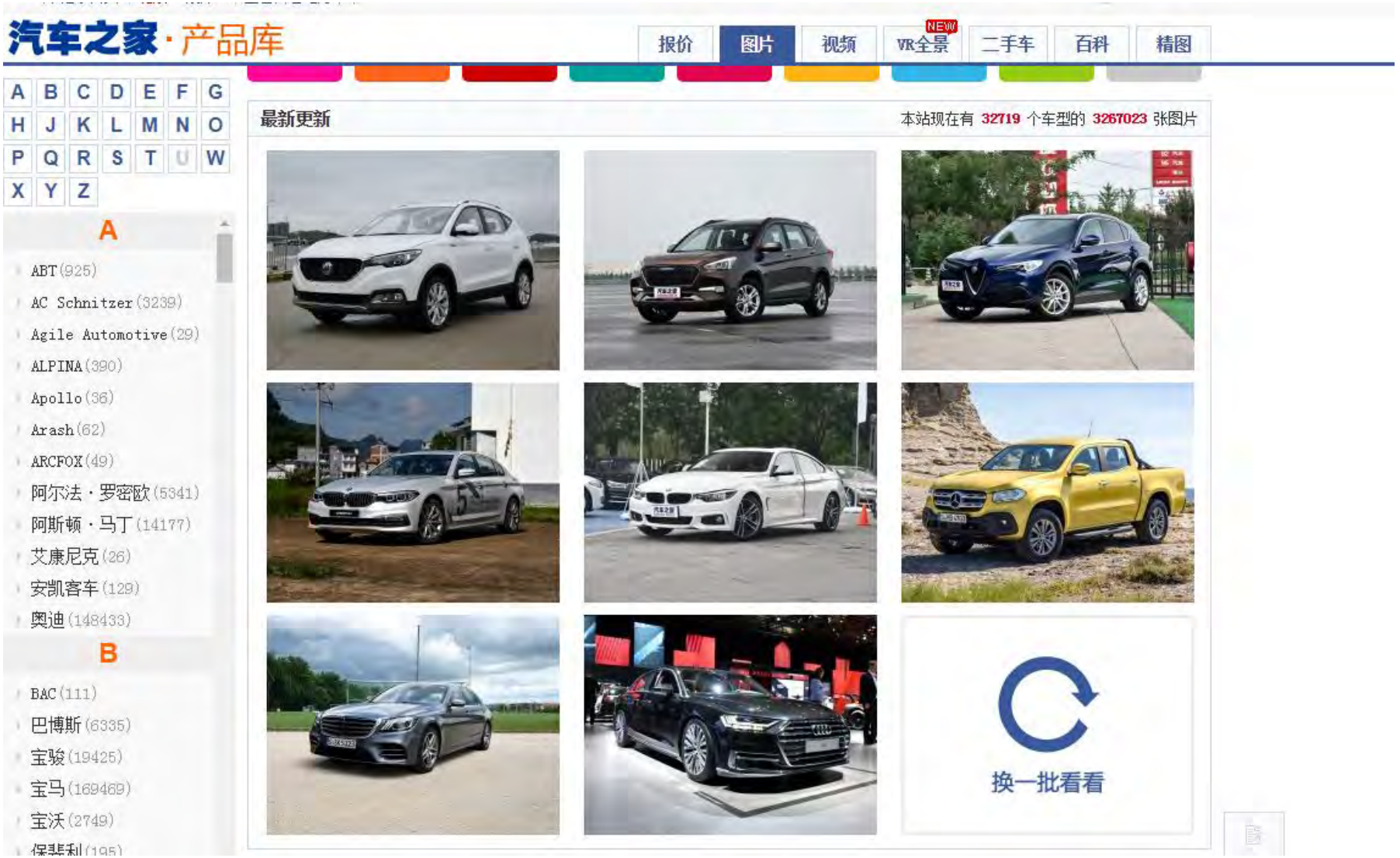

小朋友去上学简笔画图


小朋友骑童车简笔画图

风景简笔画

- 棕榈树简笔画图片、教程
- 紫藤花简笔画图片、教程
- 紫荆花简笔画图片、教程
- 好看竹子简笔画图片大全
- 栀子花简笔画图片大全
- 樟树简笔画图片、教程
- 月季花简笔画分步骤教程
- 郁金香花简笔画图片大全
- 玉兰花简笔画图片大全
- 樱花简笔画图片大全
- 罂粟花简笔画图片、教程
- 椰子树简笔画怎么画图解

查询词-站点相关性特征


汽车之家·产品库
报价
图片
视频
VR全景
NEW
二手车
百科
精图
A B C D E F G
H J K L M N O
P Q R S T U W
X Y Z
A
ABT(925)
AC Schnitzer(3239)
Agile Automotive(29)
ALPINA(390)
Apollo(36)
Arash(62)
ARCFOX(49)
阿尔法·罗密欧(5341)
阿斯顿·马丁(14177)
艾康尼克(26)
安凯客车(129)
奥迪(148433)
B
BAC(111)
巴博斯(6335)
宝骏(19425)
宝马(169469)
宝沃(2749)
最新更新
本站现在有 32719 个车型的 3267023 张图片
换一批看看

查询词-站点相关性特征

如何获得查询词的语义向量、站点的语义向量？

查询词-站点相关性特征

如何获得查询词的语义向量，以及如何获得站点的语义向量？

- ✓ 数据获取：随机挑选一批查询词，获取点击数据
- ✓ 数据处理：通过点击获得查询词和doc站点的链接关系，生成二部图，如下：

  猫简笔画---www.jianbihua.cc-兔子简笔画---www.jbhua.com---…

  奥迪---photo.auto.sina.com.cn---宝马---news.bitauto.com---。。。
- ✓ Deepwalk：将如上二部图给deepwalk，生成query和site的语义向量。基于这语义向量，即可计算query与site的相关关系。

  Deepwalk包括两部分逻辑，randomwalk和word2vector，randomwalk基于二部图生成随机的序列，将这些序列类比成句子，输给word2vector，生成语义向量。

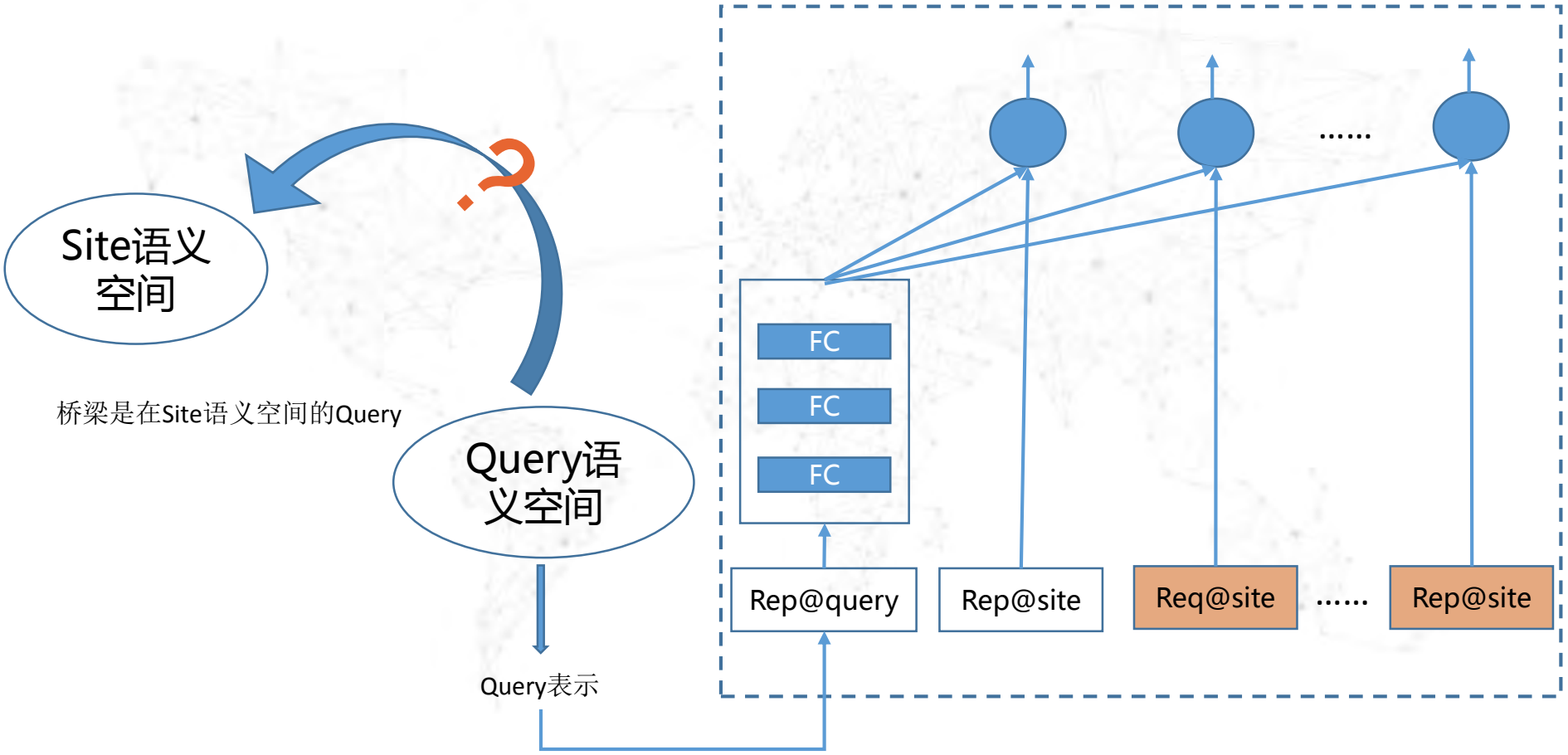
查询词-站点相关性特征

如何解决未登录词问题？

我们只得到了 ‘兔子简笔画’ 的语义向量，如何得到 ‘猫简笔画’ 的语义向量呢？

---和查询词-图像相关性特征计算同理，把站点看成是图像，采用相同架构学习既可解决。

查询词-站点相关性特征


Site语义空间
桥梁是在Site语义空间的Query
Query语义空间
Query表示
FC
FC
FC
Rep@query
Rep@site
Req@site
Rep@site

查询词-站点相关性特征

与[两只老虎简笔画]相近的站点

#	Site	Sim
1	www.zgyzweb.com	0.665068
2	www.shuobaobao.com	0.658024
3	www.baby-edu.com	0.655322
4	www.haomum.com	0.655112
5	www.henanart.com	0.654309

与[jianbihua.org]相近的词语

#	Word	Sim
1	小班画画小火车简笔画	0.561953963238
2	儿童简笔画图片大全	0.508199520869
3	圆圈简笔画	0.503759834183
4	托马斯火车简笔画大全	0.494985946527
5	远离陌生人简笔画	0.490371766657

查询词-站点相关性特征

效果：单特征NDCG@10：0.63
NDCG@10提升1%

先验知识融入相关性计算

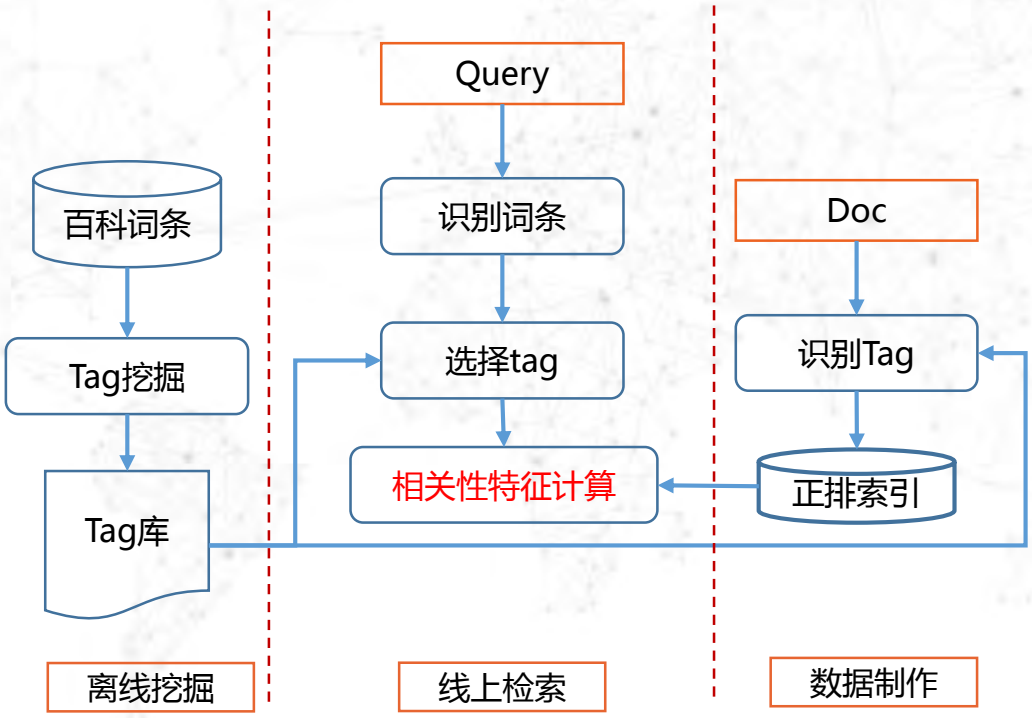
先验知识融入相关性计算

知识？

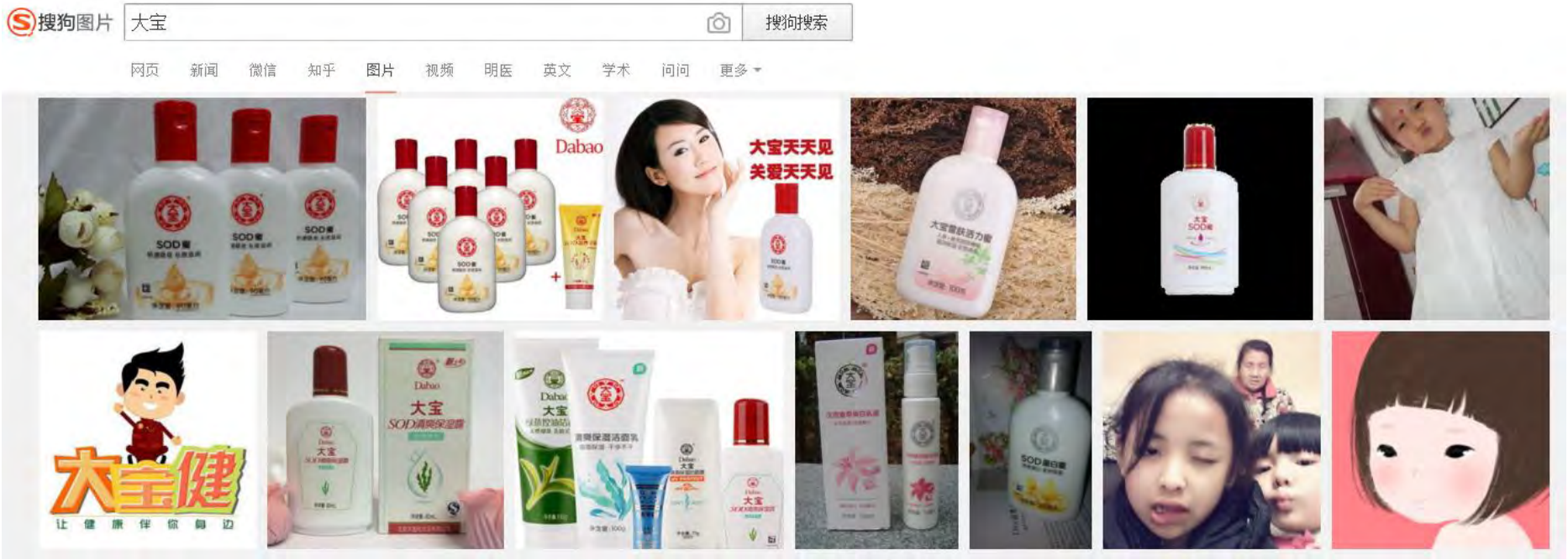
先验知识融入相关性计算

先验知识融入相关性计算

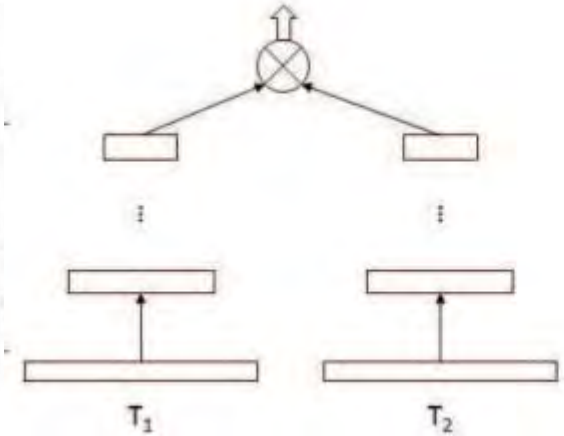
大宝（护肤品）
化妆品:0.7384， 护肤品:0.2822， 公司:0.2150， 企业:0.2101， 美容:0.1975， 公益:0.1834， 皮肤:0.1775， 北京|Beijing|帝都|Peking|顺天府|北平|燕京|京城:0.1245， 制药:0.1182， 名牌:0.1075， 技术:0.1019，



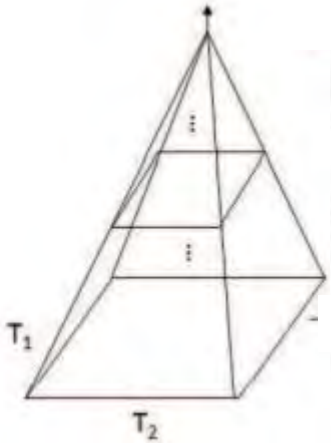
先验知识融入相关性计算



文本匹配相关性特征



• 基于表示模型



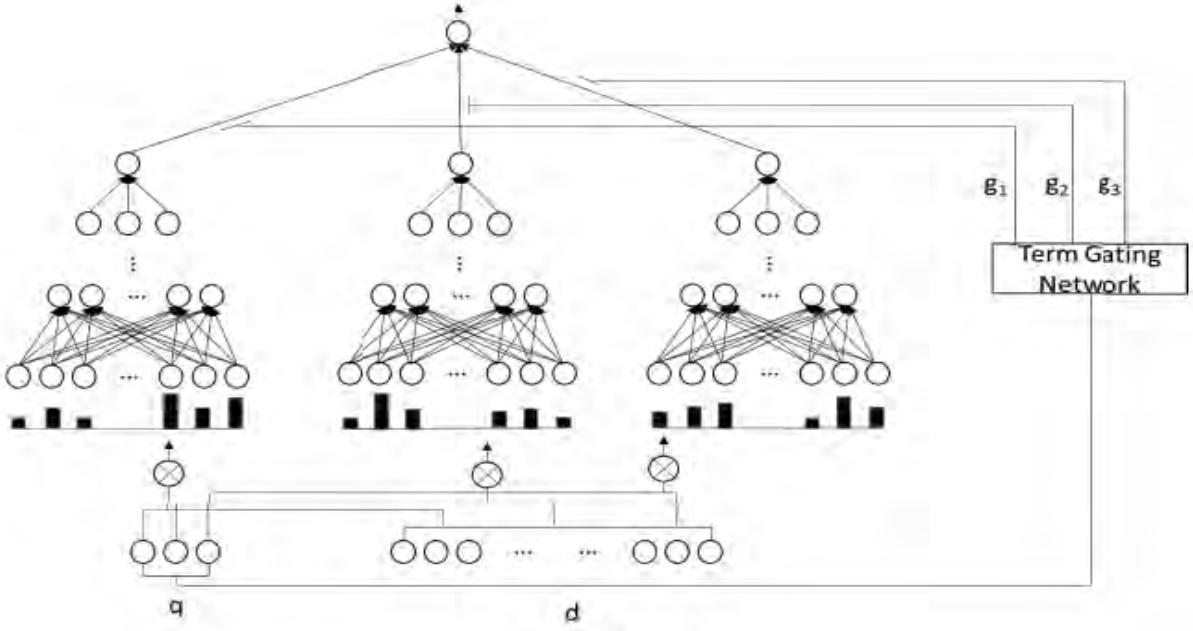
• 基于交互模型

– 表示模型： DSSM/CDSSM，ARC-I

– 交互模型： Deep-Match，ARC-II，MatchPyramid，DRMM

文本匹配相关性特征

- **思路：融合匹配网络与词选择网络(DRMM)**
- **方法：构造直方图，DNN/Term-gating**

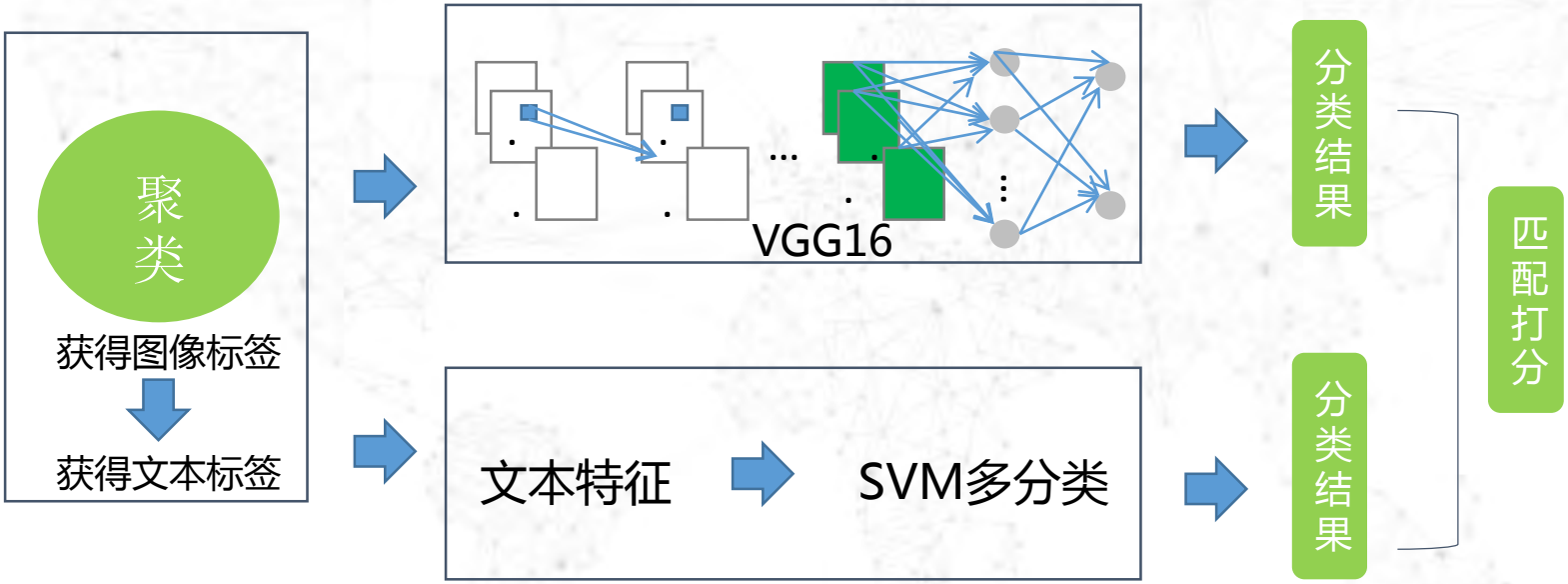


单特征NDCG高于最好的人工设计的匹配相关性特征

文图对应特征

图片来自页面，页面中的文本图片是否相关，从数据层面直接影响到图片搜索的相关性效果。因此直接从数据端出发，分析页面文本与页面图片之间的相关性。

文图对应特征



文图对应特征

```
itle Body :
v i v o  S 6
**********
Image Prediction Result is:
     top 10 label : confidence
7:0.561660 562:0.333887 433:0.055889 753:0.019239
**********


**********
Text Prediction Result is:  17.000000
**********

[dicmap unlock success.]

**********
The correlation coefficient is 0.000000
**********
注:文本抽取的是图片Title_boy,当文本类别为-1,代表相
```


```
v i v o手机价格
**********
Image Prediction Result is:
     top 10 label : confidence
627:0.999003 734:0.000361 790:0.000270 898:0.0
**********


**********
Text Prediction Result is:  17.000000
**********

[dicmap unlock success.]

**********
The correlation coefficien is 0.999003
**********
注:文本抽取的是图片Title_boy,
```

结合图像特征的关键词提取

传统的关键词提取处理的对象是纯文本。而图片搜索由于其业务特殊性，不仅有文本，还有图片，因此关键词提取需要将图片考虑进来。

结合图像特征的关键词提取

传统的关键词提取相关特征：词出现位置、出现域、tfidf、embedding等

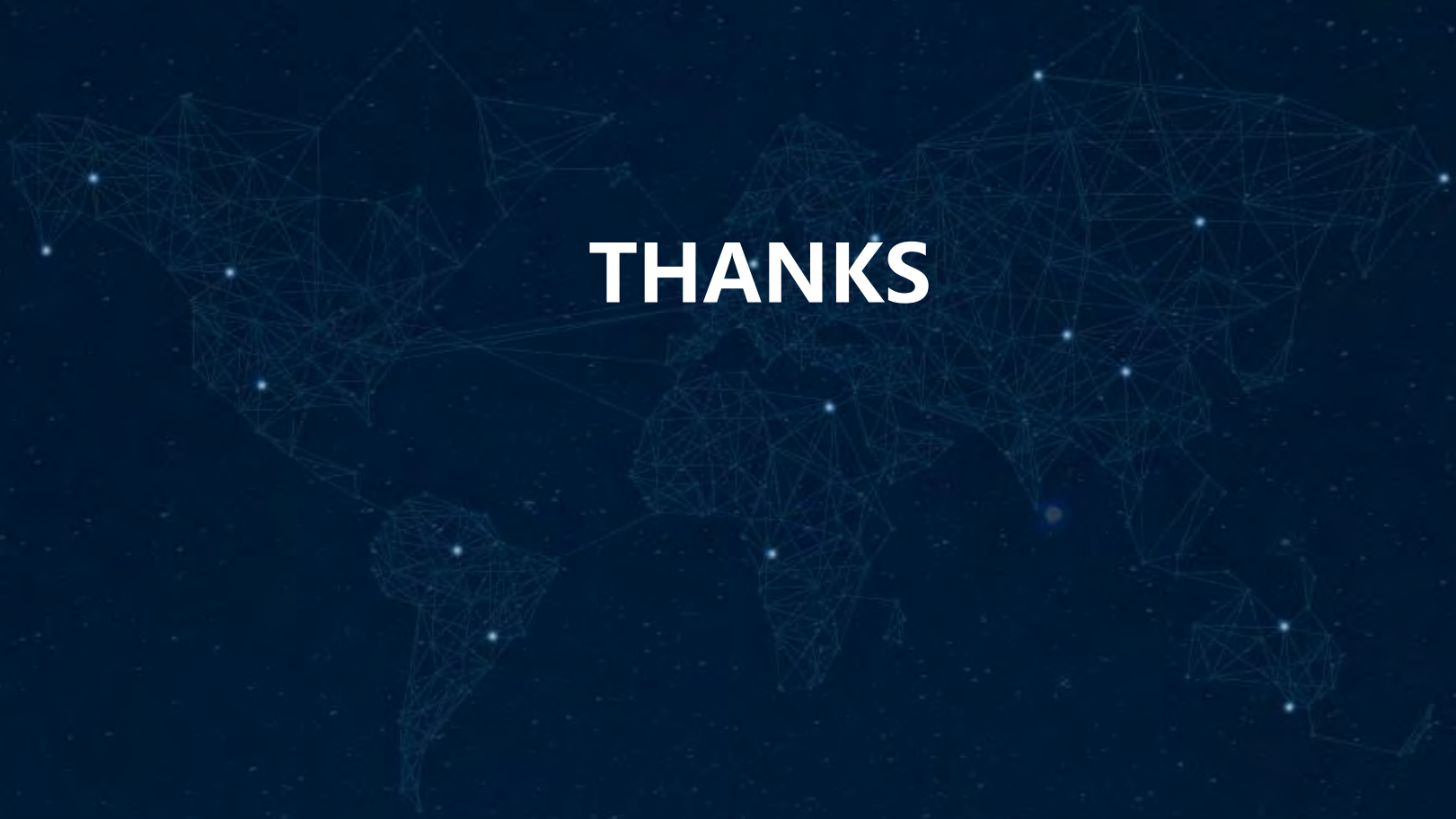
图像相关的特征：图像CNN特征、图像通过DSSM后的特征

模型：LambdaMart

效果：加入图像相关的特征后，NDCG@3提升9%

效果

一年多时间里，从人工设计、调试特征转到特征学习，让数据指导效果优化，相关性提升15%，超过过去三年里提升的总和。



THANKS