

# 搜狗图片搜索系统智能化演进之路

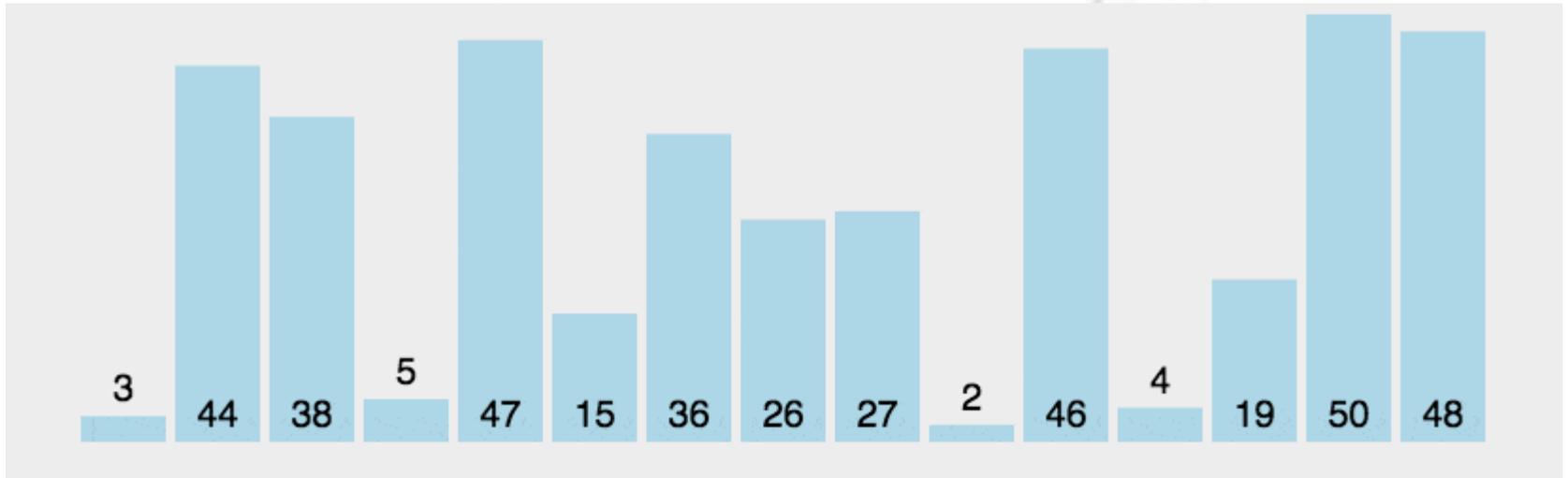
周泽南 搜狗专家研究员

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



The screenshot shows a search interface for '刘德华' (Liu Dehua) on the Baidu image search platform. At the top, there is a search bar with the text '刘德华' and a '搜狗搜索' (Sogou Search) button. Below the search bar, there are navigation tabs for '网页', '新闻', '微信', '知乎', '图片', '视频', '明医', '英文', '学术', '问问', and '更多'. A secondary navigation bar includes '首页', '图说新闻', '美女', '搞笑', '壁纸', '明星', '家居', '汽车', '艺术云图', and 'LOFTER'. Below this, there are filters for '相关搜索' (Related Search) with terms like '刘德华的老婆', '刘德华老婆朱丽倩照片', '刘德华图片大全', '拆弹专家刘德华', and '刘德华女儿'. There are also dropdown menus for '全部尺寸', '全部颜色', and '全部类型'. A third navigation bar lists categories: '全部', '桌面', '剧照', '写真', '发型', '壁纸', '头像', '书法', '年轻时', '颁奖典礼', '结婚照', and '替身'. The main content area displays a grid of 14 image thumbnails of Liu Dehua in various poses and outfits, including formal suits, casual wear, and a sketch.

# 图片搜索排序

图片搜索：满足用户查找图片的需求



网页 新闻 微信 知乎 图片 视频 明医 英文 学术 问问 更多 ▾

首页 图说新闻 美女 搞笑 壁纸 明星 家居 汽车 艺术云图 LOFTER

相关搜索：向日葵手绘图片简笔画 手绘向日葵图片唯美 向日葵黑白手绘图片 手绘向日葵图片素材 向日葵花语

全部尺寸 ▾ 全部颜色 ▾ 全部类型 ▾



# 图片搜索排序

Query:

 **搜狗图片** [新闻](#) [网页](#) [微信](#) [知乎](#) [图片](#) [视频](#) [明医](#) [英文](#) [地图](#) [更多>>](#)

刘德华



搜狗搜索

# 图片搜索排序

Doc:

华仔默认朱丽倩：乖乖地你们不要再问啦！（图）

日期：2008-05-16 08:59:57 来源：中国娱乐网 进入评论0条

导读：华仔默认朱丽倩：乖乖地你们不要再问啦！



刘德华

“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。

数据  
积  
累

```

DESC_TITLE_(0):
(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

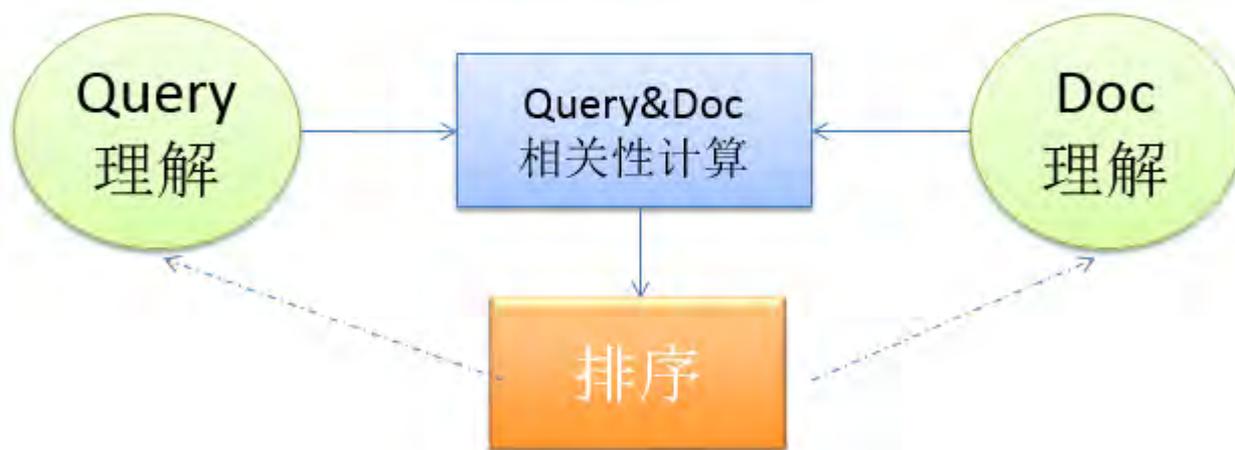
Query-Doc相关性排序:



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](#) [新闻](#) [网页](#) [微信](#) [知乎](#) [图片](#) [视频](#) [明医](#) [英文](#) [地图](#) [更多>>](#)

刘德华



搜狗搜索

```
DESC_TITLE_(0):  
(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;
```

## 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



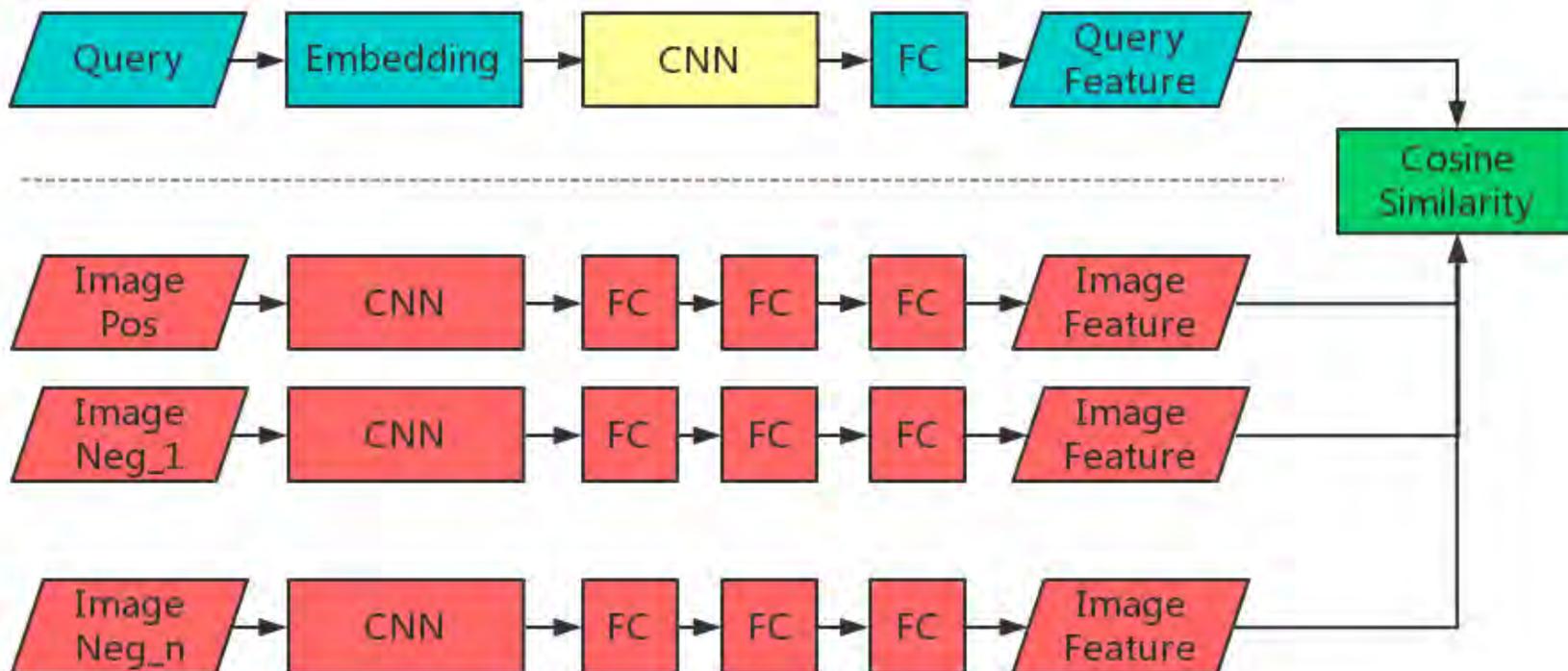
## 查询词-图像相关性特征



$$\text{Cosin}(\text{func}(v(\text{奥迪})), v(\text{img})) \gg \text{Cosin}(\text{func}(v(\text{刘德华})), v(\text{img}))$$

- ❑ 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



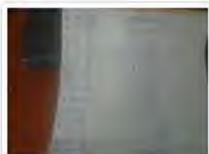
# 查询词-图像相关性特征

query

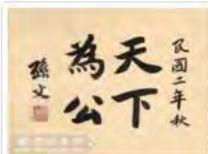
query: 三毛经典语录

三毛(1) 经典(1) 语录(1)

三毛|157 经典|84 语录|125



0.975288



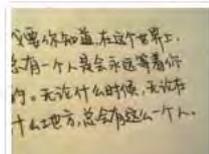
0.968754



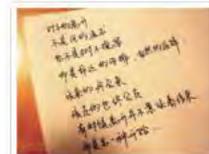
0.96215



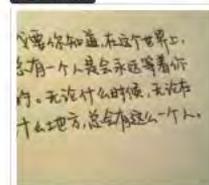
0.962094



0.958758



0.958646



0.958625

# 查询词-图像相关性特征

query 刘德华汽车

query: 刘德华汽车

刘德华(1) 汽车(1)

刘德华|177

汽车|152



0.973907



0.972736



0.970532



0.966198



0.965475



0.959137



0.957821



0.956112



0.95517



0.954138



0.952664

## 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	$200 * 4 * 8 = 6400$ bit	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

<h3>简笔画视频教程</h3> <div style="display: flex; justify-content: space-around;"> <div style="text-align: center;">  <p>蝙蝠简笔画视频教程</p> </div> <div style="text-align: center;">  <p>长颈鹿简笔画视频教程</p> </div> <div style="text-align: center;">  <p>小刺猬简笔画视频教程</p> </div> <div style="text-align: center;">  <p>海豹简笔画视频教程</p> </div> <div style="text-align: center;">  <p>鳄鱼简笔画视频教程</p> </div> </div> <div style="display: flex; justify-content: space-around; margin-top: 10px;"> <div style="text-align: center;">  <p>绘制的简笔画工具和重</p> </div> <div style="text-align: center;">  <p>荡秋千的女孩和漂亮的</p> </div> <div style="text-align: center;">  <p>卡通小绵羊简笔画视频</p> </div> <div style="text-align: center;">  <p>卡通女孩简笔画视频教</p> </div> <div style="text-align: center;">  <p>卡通人物画家简笔画视</p> </div> </div>					<h3>简笔画图片教程</h3> <ul style="list-style-type: none"> <li>▶ 动物简笔画画法教程</li> <li>▶ 动力滑翔伞简笔画图片教程</li> <li>▶ 乒乓球拍简笔画图片教程</li> <li>▶ 蝴蝶风筝简笔画图片教程</li> <li>▶ 降落伞简笔画画法图解</li> <li>▶ 双翼老式飞机简笔画图片教程</li> <li>▶ 螺旋桨飞机简笔画图片教程</li> <li>▶ 客机简笔画图片教程</li> <li>▶ 热气球简笔画图片教程</li> <li>▶ 小车简笔画图片教程</li> <li>▶ 旱冰鞋简笔画图片教程</li> <li>▶ 滑冰刀鞋简笔画图片教程</li> </ul>				
<h3>人物简笔画推荐</h3> <p style="text-align: center;">中国人物简笔画 雷锋简笔画</p> <div style="display: grid; grid-template-columns: repeat(5, 1fr); gap: 5px;"> <div style="text-align: center;">  <p>小朋友植树简笔画图片</p> </div> <div style="text-align: center;">  <p>小朋友洗澡简笔画图片</p> </div> <div style="text-align: center;">  <p>小朋友喂鸡简笔画图片</p> </div> <div style="text-align: center;">  <p>小女孩跳绳简笔画图片</p> </div> <div style="text-align: center;">  <p>小朋友踢足球简笔画图</p> </div> <div style="text-align: center;">  <p>小朋友上课简笔画图片</p> </div> <div style="text-align: center;">  <p>上班族简笔画图片、教</p> </div> <div style="text-align: center;">  <p>小朋友扫落叶简笔画图</p> </div> <div style="text-align: center;">  <p>小朋友去上学简笔画图</p> </div> <div style="text-align: center;">  <p>小朋友骑童车简笔画图</p> </div> </div>					<h3>风景简笔画</h3> <ul style="list-style-type: none"> <li>▶ 棕榈树简笔画图片、教程</li> <li>▶ 紫藤花简笔画图片、教程</li> <li>▶ 紫荆花简笔画图片、教程</li> <li>▶ 好看竹子简笔画图片大全</li> <li>▶ 椰子树简笔画图片大全</li> <li>▶ 樟树简笔画图片、教程</li> <li>▶ 月季花简笔画分步骤教程</li> <li>▶ 郁金香花简笔画图片大全</li> <li>▶ 玉兰花简笔画图片大全</li> <li>▶ 樱花简笔画图片大全</li> <li>▶ 罂粟花简笔画图片、教程</li> <li>▶ 椰子树简笔画怎么画图解</li> </ul>				

# 查询词-站点相关性特征

汽车之家·产品库

报价

图片

视频

NEW  
VR全景

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

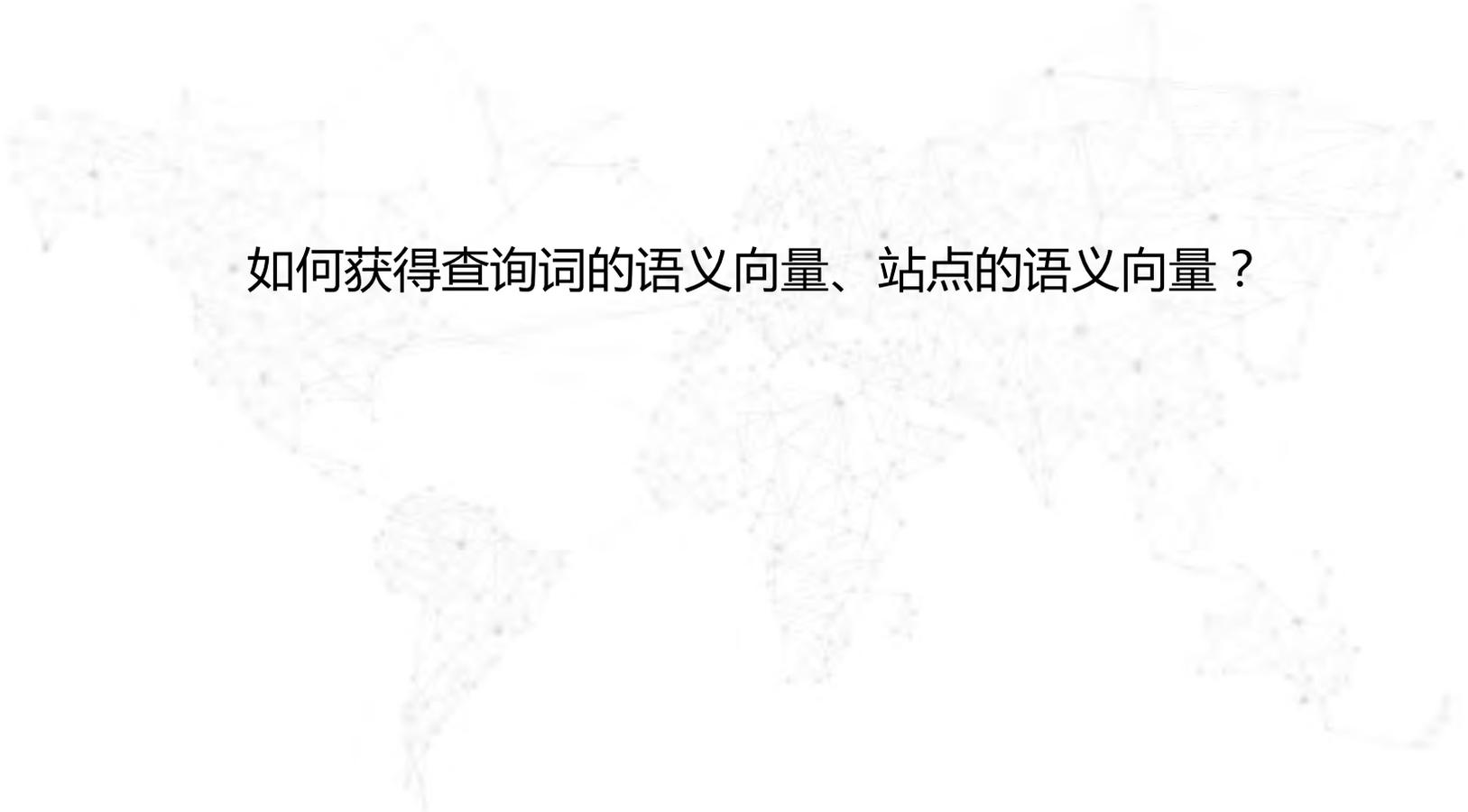
- BAC (111)
- 巴博斯 (6335)
- 宝骏 (19425)
- 宝马 (169469)
- 宝沃 (2749)
- 保斐利 (195)

最新更新

本站现在有 32719 个车型的 3267023 张图片



## 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？

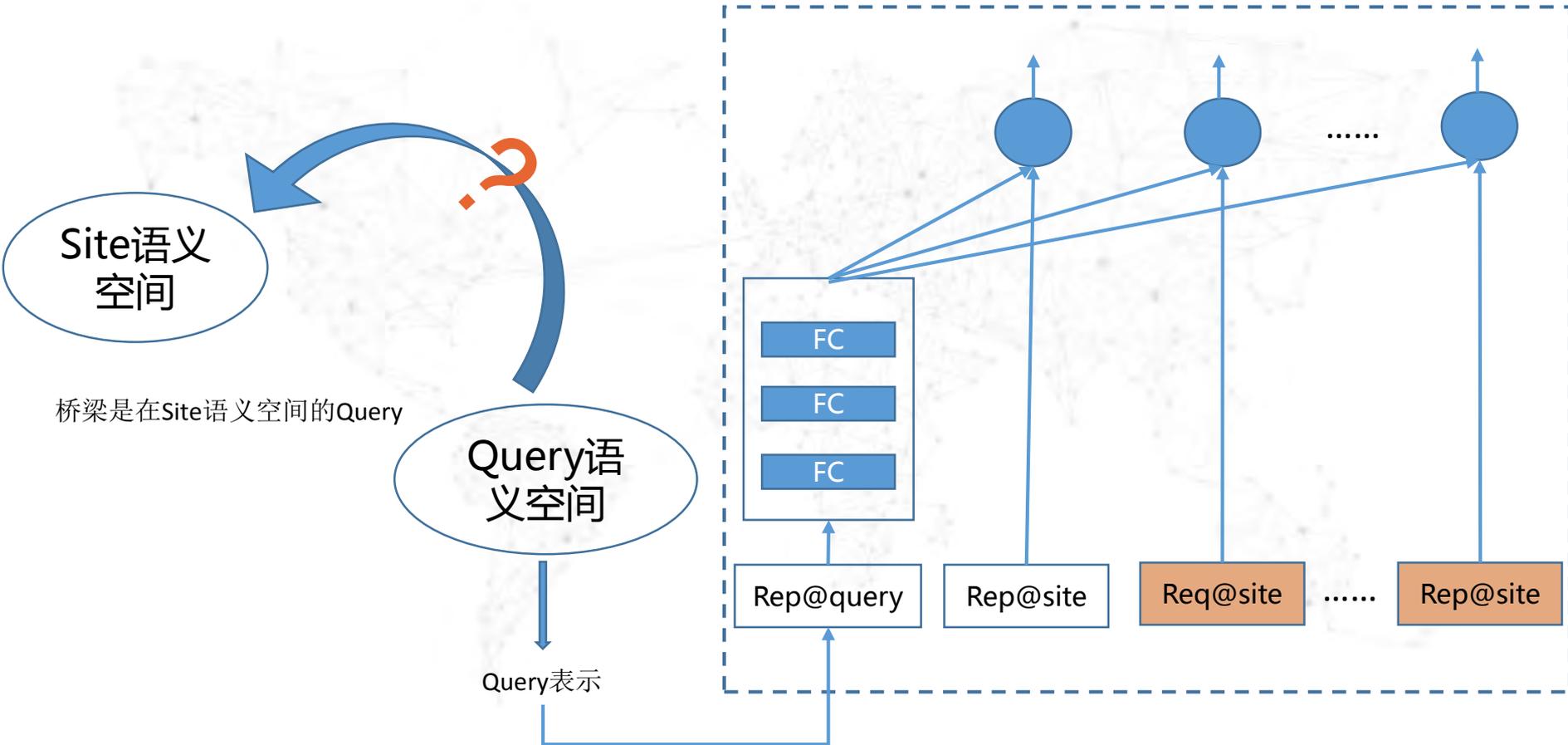
- ✓ 数据获取：随机挑选一批查询词，获取点击数据
- ✓ 数据处理：通过点击获得查询词和doc站点的链接关系，生成二部图，如下：  
猫简笔画---www.jianbihua.cc-兔子简笔画---www.jbhua.com----...  
奥迪---photo.auto.sina.com.cn---宝马---news.bitauto.com---。。。
- ✓ Deepwalk：将如上二部图给deepwalk，生成query和site的语义向量。基于这语义向量，即可计算query与site的相关关系。  
Deepwalk包括两部分逻辑，randomwalk和word2vector，randomwalk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



# 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	<a href="http://www.zgyzweb.com">www.zgyzweb.com</a>	0.665068
2	<a href="http://www.shuobaobao.com">www.shuobaobao.com</a>	0.658024
3	<a href="http://www.baby-edu.com">www.baby-edu.com</a>	0.655322
4	<a href="http://www.haomum.com">www.haomum.com</a>	0.655112
5	<a href="http://www.henanart.com">www.henanart.com</a>	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



相关搜索: 大宝法王 大宝洗面奶 大宝漆 大宝法王玛巴 大宝法王不可思议图片

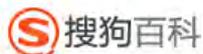


# 先验知识融入相关性计算



知识？

# 先验知识融入相关性计算



大宝

进入词条

搜索词条



首页 精彩百科 任务 用户 公益百科 积分商城

个人中心

大宝是一个多义词，您可以选择查看以下义项：

+ 添加义项

护肤品

演员艺名

年号

词语

葡萄品种

张涵予、黎明演唱的歌曲

护肤品

同义词 收藏 分享

## 大宝

编辑词条



大宝<sup>[1]</sup> 国产知名品牌，“大宝”系列化妆品1985年诞生至今，适应了不同时期、不同层次的消费需求，已陆续形成**护肤、洗发、美容修饰、香水、特殊用途**共五大类100多个品种。其中，1985年—1990年期间推出的**速消眼角皱纹蜜、老年斑霜、眼袋霜、减肥霜、美乳霜、生发灵**等产品在中国外长销不衰，享誉至今。

国产护肤品牌

展开



美加净



郁美净



百雀羚



雅霜



相宜本草



迷奇

快速导航 知乎精选

中文名 大宝

生产单位 北京大宝化妆品有限公司

词条信息

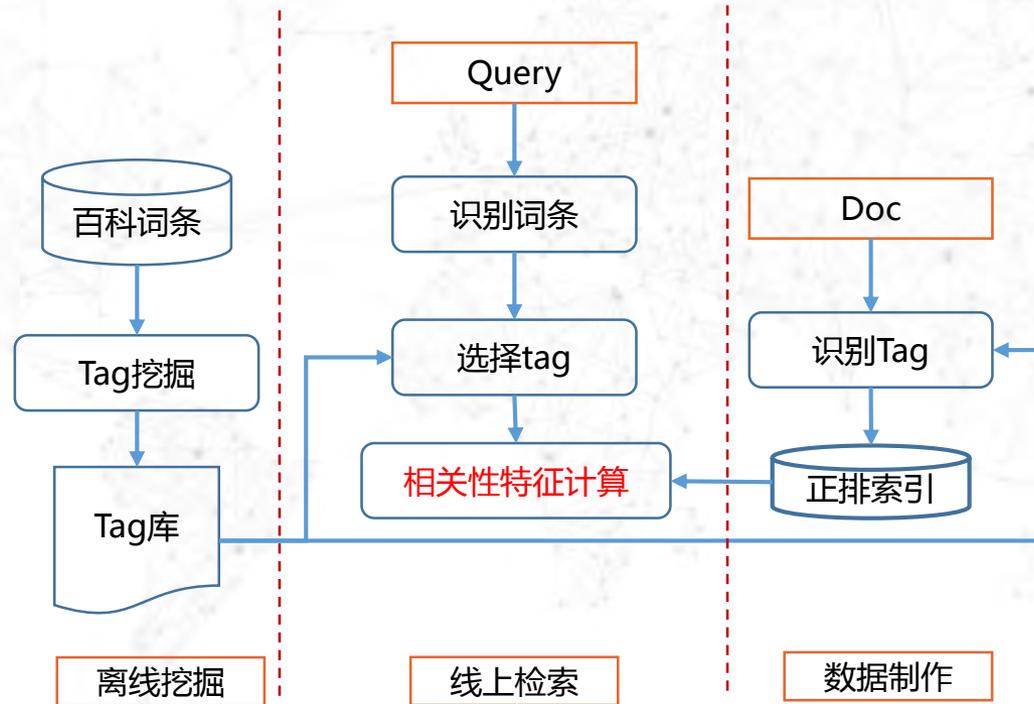
创建者: 会跑的树

编辑次数: 22次

词条浏览: 85277次

# 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,



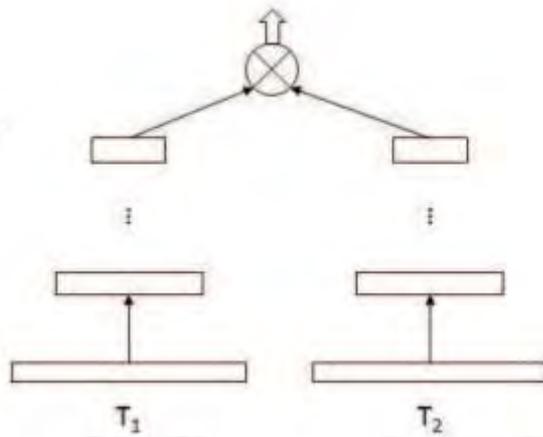
# 先验知识融入相关性计算



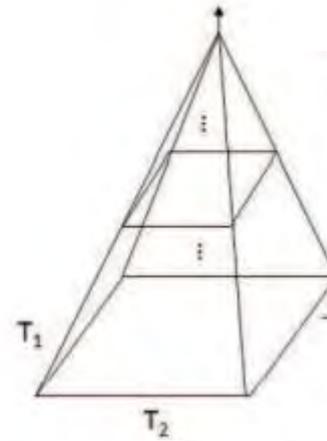
网页 新闻 微信 知乎 图片 视频 明医 英文 学术 问问 更多 ▾



## 文本匹配相关性特征



• 基于表示模型

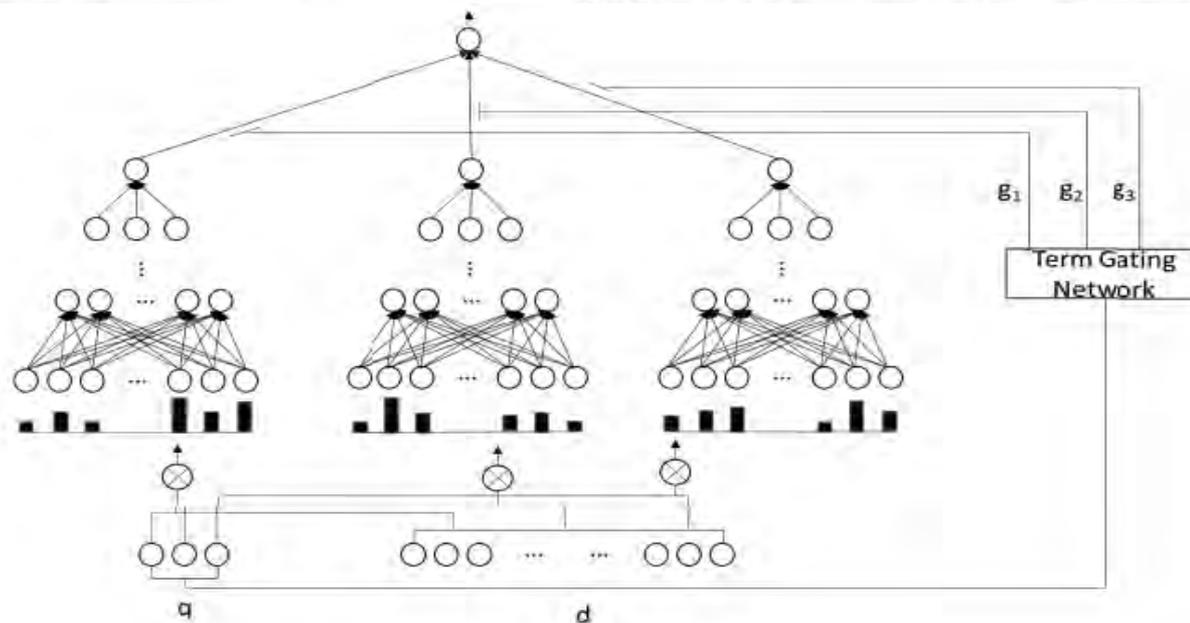


• 基于交互模型

- 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

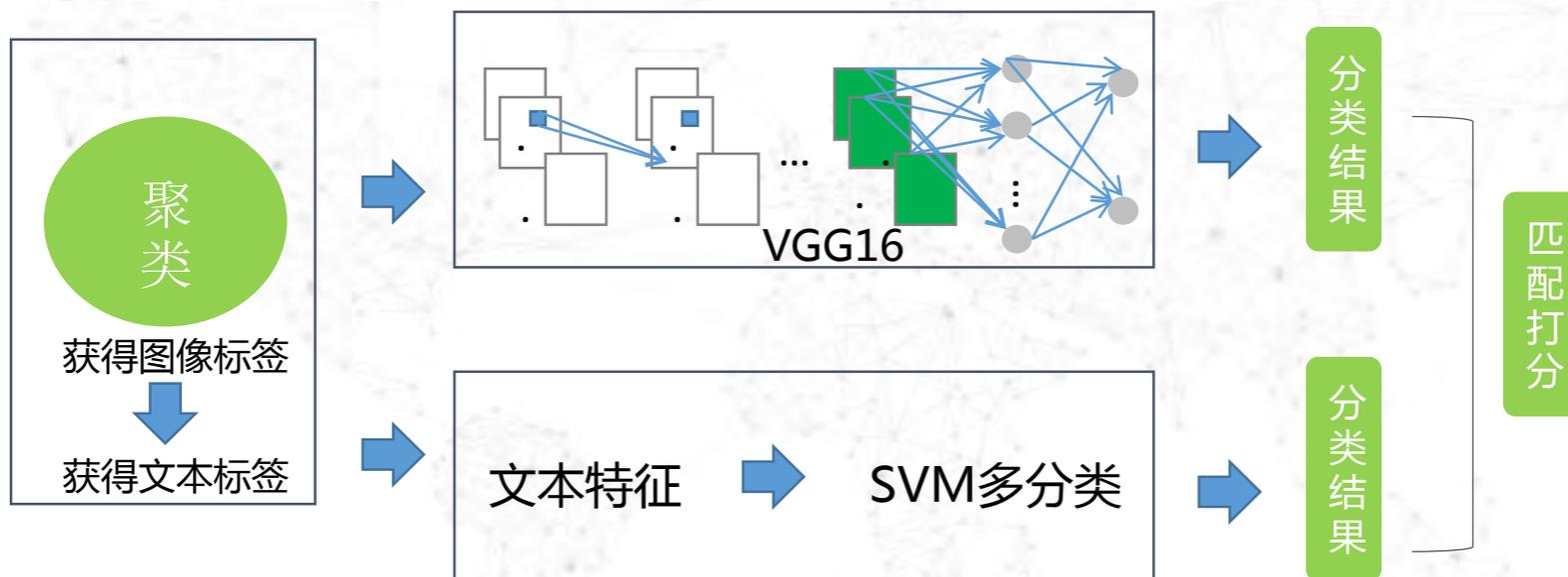
## 文图对应特征

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。

vivo S6 全部(3/37)



# 文图对应特征



# 文图对应特征



```

title Body :
vivo S6
*****
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

```



```

title Body :
vivo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 coefficient is 0.999003
*****
注:文本抽取的是图片Title_boy,当文本类别为1,代

```

## 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。

**[图片] 宝马全系列 PK 哈士奇!(还完后续的)完整篇!** [复制链接]

发表于 2007-8-23 14:35 | 只看该作者 | 1楼

个人欣赏观点: 宝马的前灯设计很像黑白的哈士奇眼睛! 比比谁的目光更“狠”!

[ 本帖最后由 我爱边疆 于 2007-8-23 19:58 编辑 ]

**都让开, 我来了!.jpg** (27.89 KB, 下载次数: 24)

A silver BMW sedan is shown from a front-three-quarter view, parked on a paved area with a pattern of small green plants. In the background, there is a modern building with large glass windows and a covered entrance. The sky is clear and blue.

## 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**