QCon[上海站]
全球软件开发大会2016

深度学习技术在图片搜索与图像搜索中的实践

SPEAKER
搜狗资深研究员
周泽南

International Software
Development Conference

主办方 Geekbang. 极客邦科技 InfoQ

关于分享内容

《深度学习技术在图片搜索与图像搜索中的实践》

《机器学习技术在图片搜索中的实践》包括深度学习技术

大纲

- 排序
- 图片搜索中的排序
- 对图片搜索新的思考
- Multimodal Learning
- 结合图像特征的关键词抽取

排序

Wiki:在计算机科学与数学中，一个**排序算法**（Sorting algorithm）是一种能将一串资料依照特定方式排序的一种算法。

排序：[5,1,7,9,10] -> (**排序算法**) ->1,5,7,9,10

---**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：通过输出与用户输入query相关的图片来满足用户寻找图片的需求

图片搜索排序

研究对象：

- 用户输入文本->Query
- 输出结果->Doc
- Query-Doc相关性

图片搜索排序

Query：

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

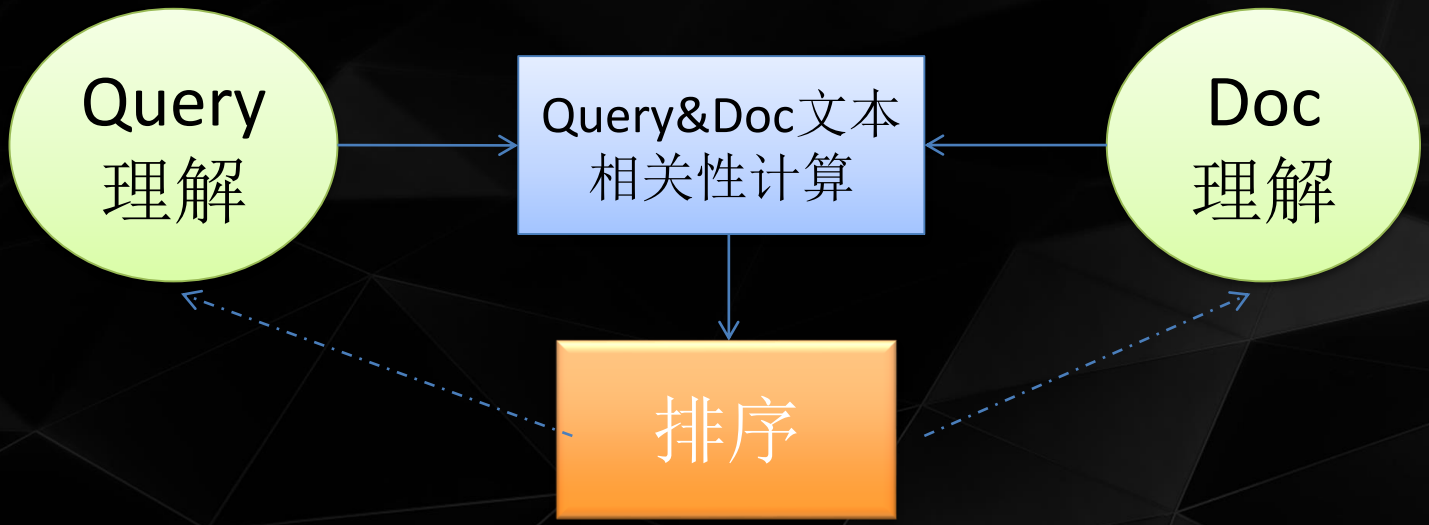
```
(T_P1_TITLE_BODY_)TITLE_(40):华仔默认朱丽倩　乖乖地你们不要再问啦
(T_CLICK_)AUTHOR_(0):
(T_P1_TITLE_DESC_)ANCHOR1_(8):刘德华
(T_P1_ALT_)ANCHOR2_(0):
ANCHOR_EXTEND_(0):
STRIP_URL_(0):
(T_ENTITY_)KEYWORD_(0):
(T_QUERY_GG_)METAINFO_(0):
(T_P1_TITLE_HTML)CONTENTTITLE_(38):华仔默认朱丽倩　乖乖地你们不要再问啦
(T_CLUSTER_TERM_H_)TOPIC_(0):
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。。。

(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；

对图片搜索新的思考与探索

除了计算Query文本与Doc文本的相关性，还有其它计算相关性的维度吗？

相关性两要素：Query，Doc文本(doc)

→

相关性四要素：Query，Doc文本(Doc)，Doc图片(Pic)，Site

单重相关性：
Query-Doc相关性

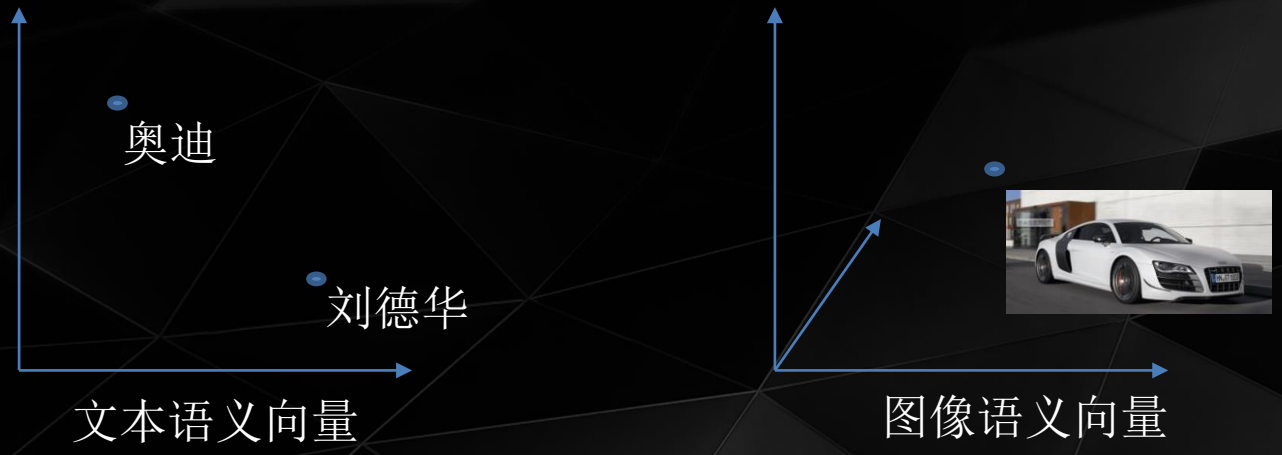
→

多重相关性：
Query-Pic相关性，
Query-Site相关性，
Doc-Pic相关性，
pic-site相关性。。。

对图片搜索新的思考与探索

如何计算Query-Pic相关性、Query-Site相关性以及Doc-Pic相关性呢？

- ✓ Multimodal Learning
- ✓ 直观解释：寻找不同的语义空间之间的映射函数。



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(也就是embedding)？
- ☐ 如何得到func()?

Multimodal Learning的具体实现：

- Query-Pic相关性
- Query-Site相关性
- Doc-Pic相关性

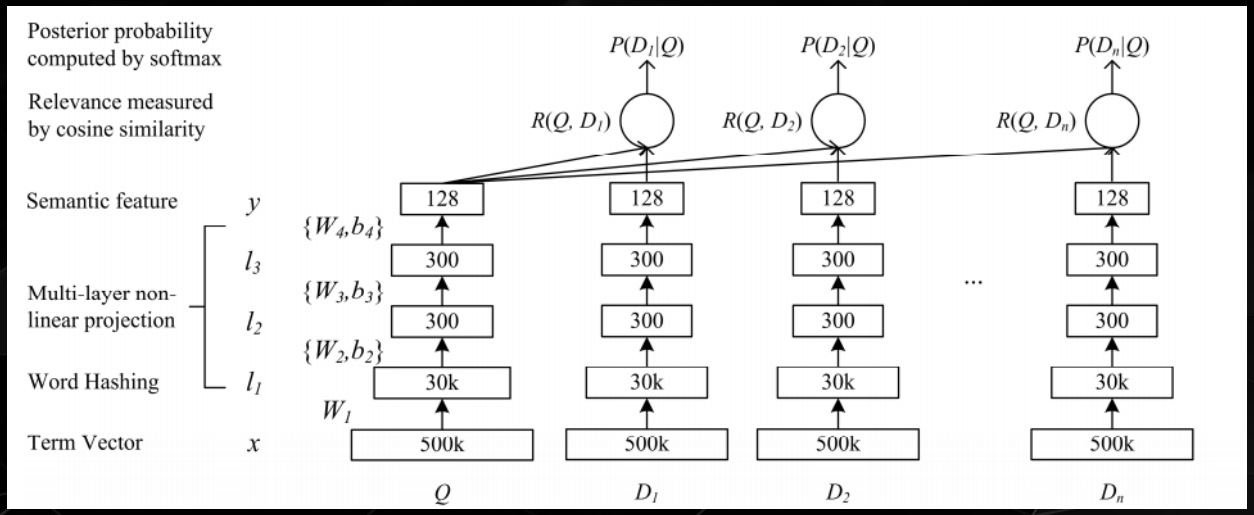
Query-Pic相关性

直接计算Query与Pic的相关性

- 如何获得Query的语义向量，即V(Query)？
  - ✓ Word2vector
- 如何获得Pic的语义向量，即V(Pic)？
  - ✓ CNN
- 如何计算V(Query)与V(Pic)相关性？
  - ✓ DSSM->DISSM

Query-pic相关性

DISSM(Doc Image Semantic Similarity Model)
DSSM(Deep Structured Semantic Model， MSRA， 2013)



$$P(D|Q) = \frac{\exp(\gamma R(Q,D))}{\sum_{D' \in \mathbf{D}} \exp(\gamma R(Q,D'))}$$

$$L(\Lambda) = -\log \prod_{(Q,D^+)} P(D^+|Q)$$

Q:表示query， 也就是我们的V(Query)
D:表示doc， 也就相当于我们的V(Pic)

REFERENCE:"Learning Deep Structured Semantic Models for Web Search using Clickthrough Data"， Huan， CIKM13

Query-Pic相关性

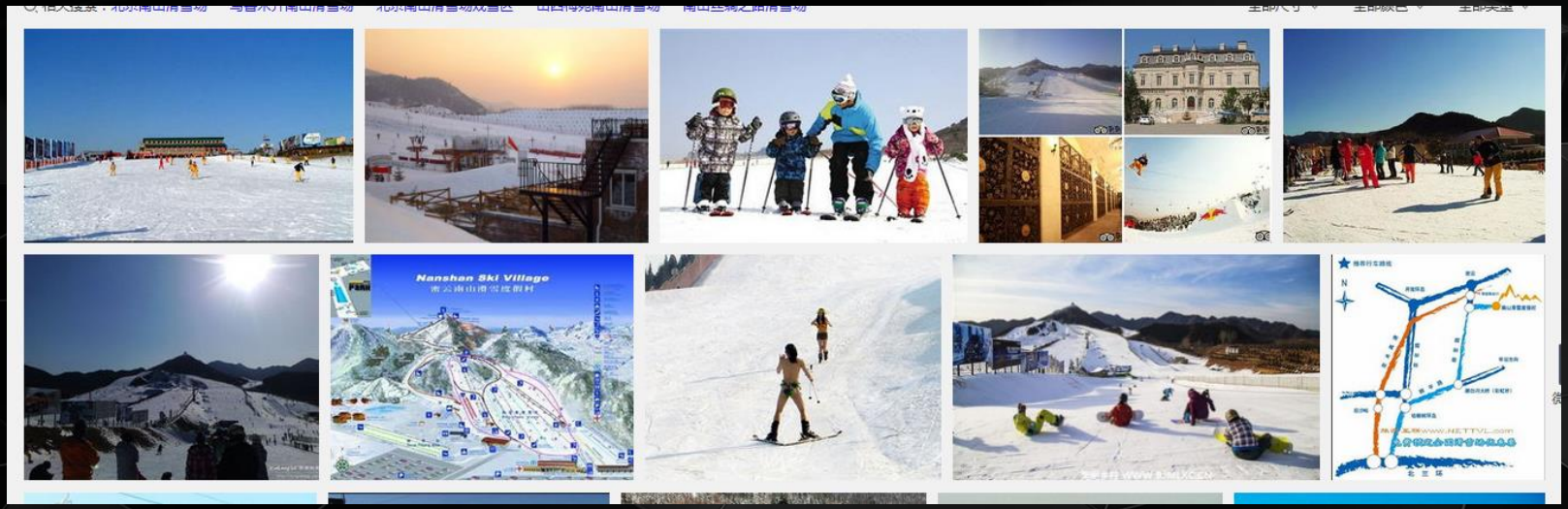
Qeury:南山滑雪场

加入Query-Pic相关性前：

Query-Pic相关性

Qeury:南山滑雪场

加入Query-Pic相关性后：



Query-Site相关性

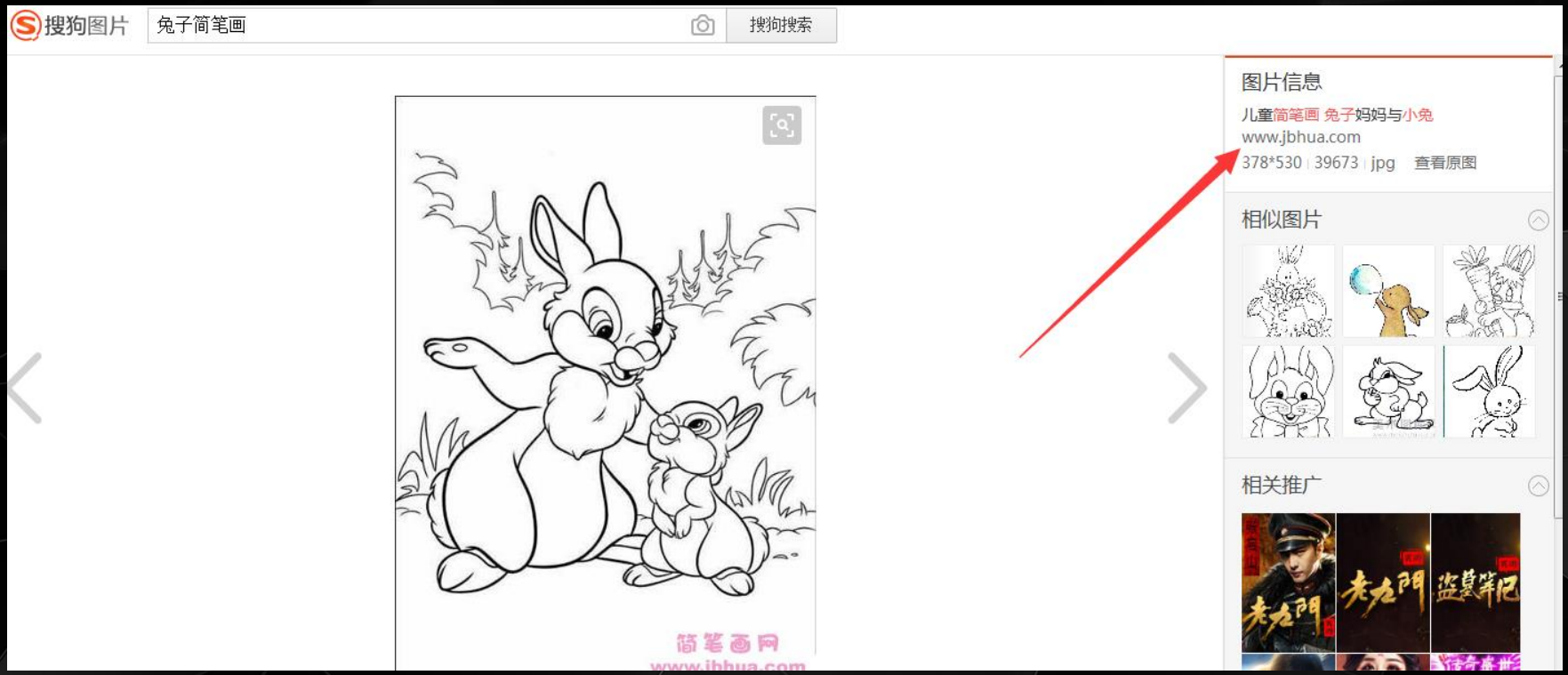
垂直站点的图片资源是能够很好满足对应的图片查询需求的，如汽车垂直网站基本能够满足汽车类query查询需求，如果能够挖掘出优质的汽车类的垂直网站，将对效果会有很大帮助。

直接计算Query与Site的相关性

- 如何获得Query的语义向量(即V(Query)),以及如何获得Pic的语义向量(即V(Site))？
- 如何解决未登录词问题？

Query-Site相关性

如何获得Query的语义向量(即V(Query)),以及如何获得Pic的语义向量(即V(Site))?



Query-Site相关性

如何获得Query的语义向量(即V(Query)),以及如何获得Pic的语义向量(即V(Site))?

http://www.jbhua.com/

Query-Site相关性

如何获得Query的语义向量(即V(Query)),以及如何获得Pic的语义向量(即V(Site))?

- ✓ 数据获取：随机挑选60W条query，爬取图片搜索结果。
- ✓ 数据处理：通过query和site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。

Query-Site相关性

- 如何解决未登录词问题？

我们只得到了‘兔子简笔画’、‘猫简笔画’的embedding，如何得到‘熊猫简笔画’的embedding呢？

---和query-pic同理，把‘兔子简笔画’、‘猫简笔画’的embedding，也就是语义向量理解成pic的语义向量，采用相同的DISSM架构学习既可解决。

Query-Site相关性

与[奥迪abc]相近的站点

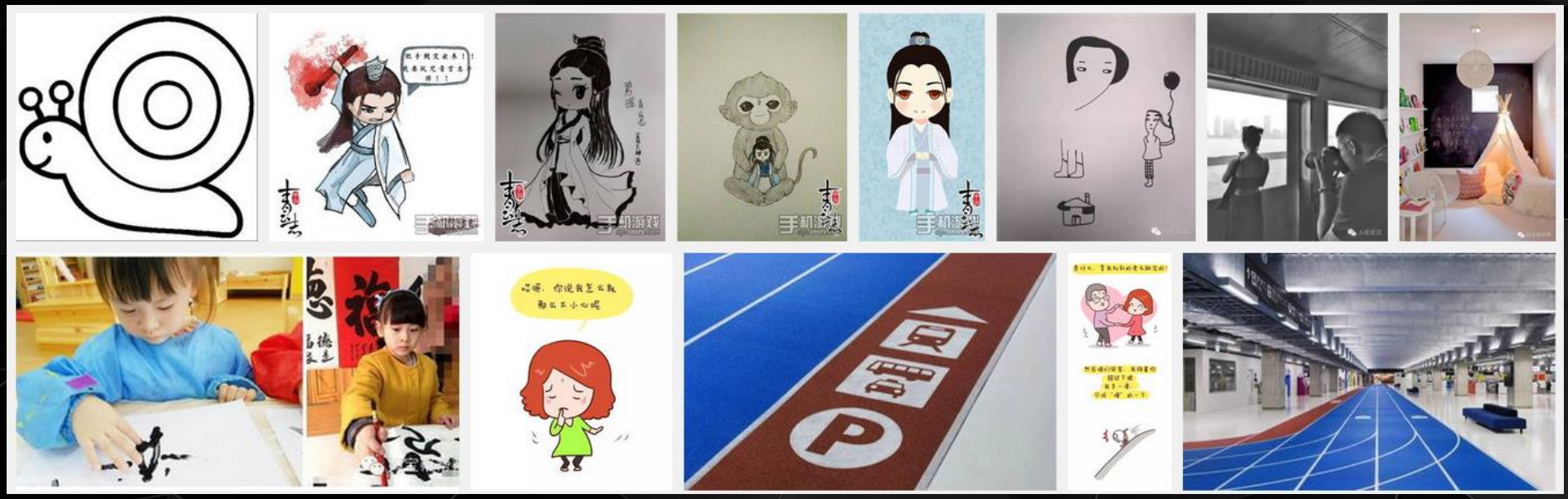
#	Site	Sim
1	www.bsaudio.cn	0.565804
2	www.cheyishang.com	0.56521
3	auto.luxtarget.com	0.559637
4	www.wanche168.com	0.559352
5	dl.china2car.com	0.559061
6	bbs.car2100.com	0.556979

与[简笔画]相近的站点

#	Site	Sim
1	www.xxjxsj.cn	0.417012542645
2	www.qlshys.com	0.396987684646
3	www.littleducks.cn	0.381129507785
4	www.jianbihua.cc	0.377285277067
5	www.yzjzxx.com	0.376516830473
6	www.61ertong.com	0.373322454634
7	hnxx.scxcedu.com	0.372100166976
8	www.mypsd.com.cn	0.368969233933
9	www.jianbihua.org	0.368941444404

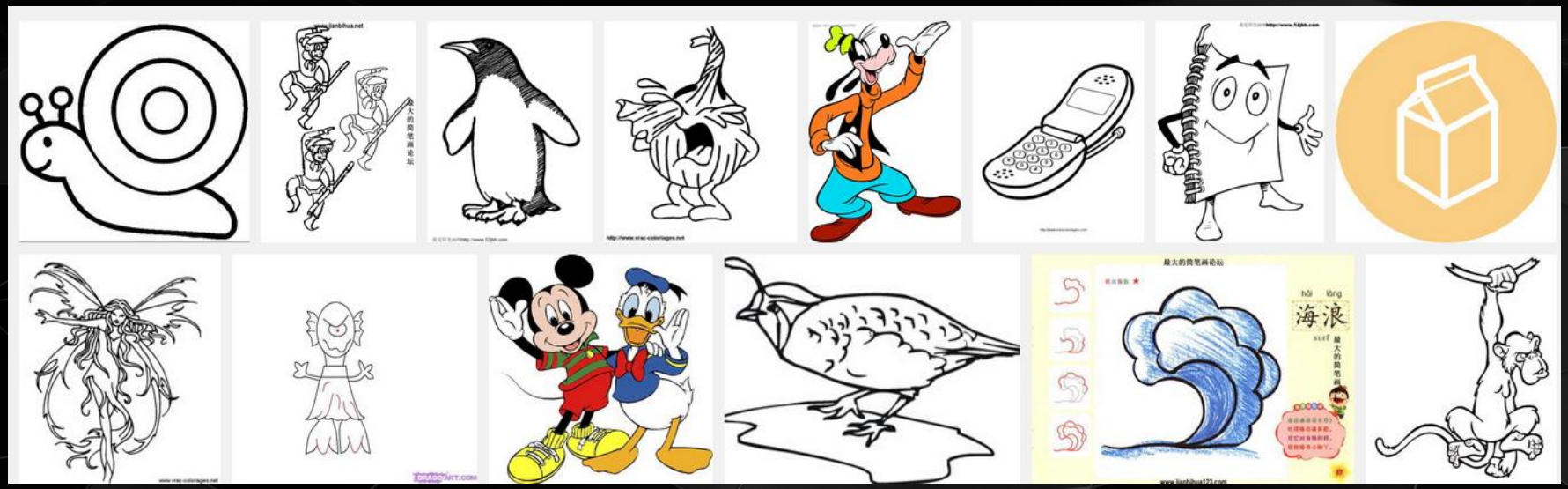
Query-Site相关性

Query：简笔画大全
加入Query-Site相关性前：



Query-Site相关性

Query：简笔画大全
加入Query-Site相关性后：



Doc-Pic相关性

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。

Doc-Pic相关性

- 如何获得doc的title的语义表达，如何获得pic的语义表达？
- 如何计算title语义表达与pic的语义表达的相关性。
  - ✓首先将优质的80万pic通过CNN获得特征表达
  - ✓利用CNN特征聚类
  - ✓将聚类结果作为分类目标训练分类器，作为pic的分类器
  - ✓将pic对应的类别的title作为相应类别，训练文本分类器，作为title的分类器
  - ✓对于一个新的page，将pic和title分别过各自分类器，计算类别之间的相关关系作为判断page的title和pic是否相关的判断因子。

Doc-Pic相关性

效果：

Pic分类器：训练数据总共80万，其中图片分类目标200类，通过4096维度特征+mlp+softmax分类最终效果 top-1准确率93%。

文本分类器：对这1000类图像对应的文本进行聚类以及人工合并，合并之后为29类，利用liblinear进行分类实验，top -1准确率90%。

```
itle 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
```

Doc-Pic相关性


```
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,
```

```
节去哪玩云南最热  热门飞行地昆明排前十
**********
Image Prediction Result is:
     top 10 label : confidence
168:0.843125 199:0.121750 30:0.010662 146:0.
**********


**********
Text Prediction Result is:  5.000000
**********

[dicmap unlock success.]

**********  **********
The correlation coefficie t is 0.843125
**********
```

结合图像特征的关键词提取

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

[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 我爱边境 于 2007-8-23 19:58 编辑*]

都让开，我来了！.jpg (27.89 KB, 下载次数: 24)

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

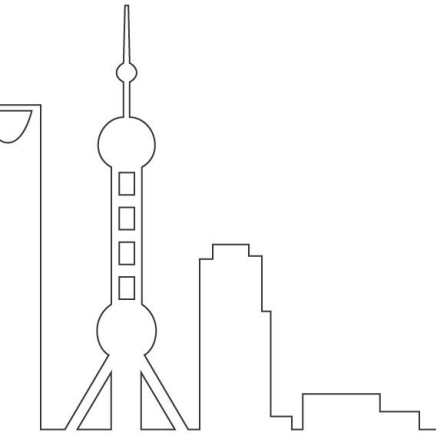
图像相关的特征：图像CNN特征、图像通过DISSM后的特征

CNN特征：将图片通过CNN，做哈希，作为图像特征

DISSM特征：分别将候选term、CNN特征通过训练好的DISSM模型，得到候选term与图像的相关性特征。

模型：LambdaMart

效果：加入图像相关的特征后，NDCG@3提升9%



Thanks!

International Software Development Conference

主办方 Geekbang. 极客邦科技 InfoQ