SHUANGLN

工作汇报

屠晓涵

2016.10.27

本周工作

1，参加2016 CCF大数据与计算智能大赛

2， 导播系统大数据平台——预测功能

- 预测功能架构
- 预测用户学历
- 预测用户性别
- 预测用户年龄
- 预测用户性格

导播系统大数据平台——建立数据库

设计说明文件（见文件夹导播系统大数据平台数据库设计文件）：

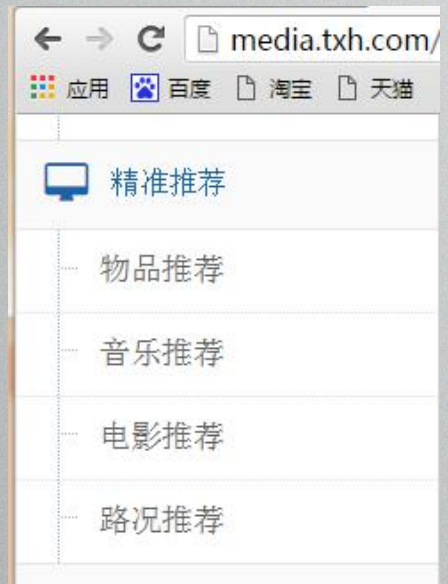
建立导播系统大数据平台数据库

FullName	NickName
(NULL)	张zzz鑫
(NULL)	litchiandapple

WeixinAccount
oRslswS7sUI0mMWEO45TFpWlqikU
oRslswUz0Jf_bLu-hFx5Ox3OuqW8
oRslswTAZPsFkDAD5a4EzkTrKhas
oRslswVWHTIDBS1Kh3viYBtDb74o
oRslswZpk4ps4EvlPvZvwL_HhxVA
oRslswYHkaDVeqO5Bdh3nRSkeeO8
oRslswdJvX-Oqb074-wOUprdfkCs
oRslswWVEmb1sIeEGz9dqRO6uRvI
oRslswdybACMhbiEkWJr5mY3yEGY
oRslswdybACMhbiEkWJr5mY3yEGY
o1Ia_tz1m6GsqcaYBsAIY8ZKPI6I
o1Ia_twHxOU62qh7rnMwavcv6u6U
o1Ia_t6mHUCiDtImwJxeldtGE3Jk
o1Ia_t2P5T8iv3mcw5JF6bww4Z6k
o1Ia_txYkkdSkRHeZnOVsM0FaA04
o1Ia_txYkkdSkRHeZnOVsM0FaA04
o1Ia_txYkkdSkRHeZnOVsM0FaA04
o1Ia_txYkkdSkRHeZnOVsM0FaA04
o1Ia_txYkkdSkRHeZnOVsM0FaA04

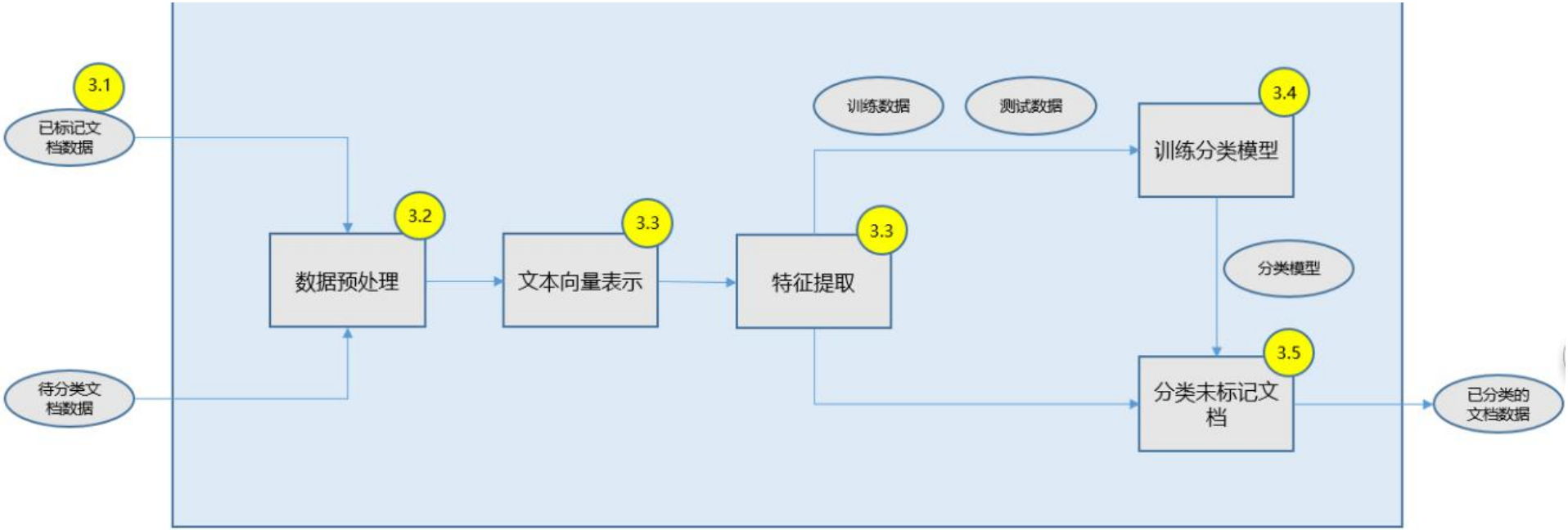
导播系统大数据平台——功能分类

- ✓ 预测功能——利用导播系统历史数据，预测用户属性（性别，年龄，学历，性格，兴趣等），进行用户评级，建立用户标签
- ✓ 信息提取功能——利用导播系统发布内容，实现关键词提取，短语提取，摘要生成。
- ✓ 精准营销功能——利用导播系统用户数据



导播系统大数据平台——预测功能架构

- ✓ 利用历史数据，预测新增用户的人口属性（性别，年龄，学历等），实现精准推荐。
- ✓ 导播系统应用平台预测功能架构图如下：



导播系统大数据平台功能——预测用户学历

数据清洗和抽取：将数据载入到 Spark 系统，抽象成为一个 RDD。

- ✓ 数据去重可以用 distinct 方法。
- ✓ 数据转换主要是用了 map 方法。
- ✓ 数据过滤使用 filter 方法，它能够保留判断条件为真的数据。

```
val text = sc.textFile(args(0)) //"D:/sougou-data/test.txt"
//println(text.count())

val cnt=text.map(line => line.split("\t")).map{
  line=>
    (line(args(1).toInt),line.slice(4,line.length)) // 1 2 3
}.map {
  line=>
    var str=""
    line._2.foreach(
      x=>
        str+=(x+"\t")
    )
    (line._1,str)
}
```

导播系统大数据平台功能——预测用户学历

分词：Spark + ansj分词——基于Spark平台，效率比较高；

```
val rateDocument=originData.map(line=> line.split('\t')).filter(
//rateDocument.repartition(1)
val rate=rateDocument.map{
  s=> s(0).toInt
}
val document=rateDocument.map(s=> s(1))

val words=document.map{
  x=>
    var str=""
    val temp = ToAnalysis.parse(x)
    //加入停用词
    FilterModifWord.insertStopWords(Arrays.asList("r", "n"))
    //加入停用词性
    FilterModifWord.insertStopNatures("w", null, "ns", "r", "u",
    val filter = FilterModifWord.modifResult(temp)
    val word = for (i <- Range(0, filter.size()) if(!filter.get(
      yield filter.get(i).getName+" "
    str+=word.mkString("")
    str
}.map(line=>line.split(" ").toList)
```

```
List(快手, 直播)
List(文字, 图片)
List(朱茵色, 戒)
List(朝歌, 是, 现在, 的)
List()
List(粉, 幼, 木耳)
List(苹果)
List(2016, 漂流瓶, 文, 爱, 截图)
List(蓝牙, 车位锁)
List(百度, 云, 资源, 搜索)
List(唐朝, 首都, 是, 还是, 长安)
```

导播系统大数据平台功能——预测用户学历

词频计算：用特征哈希（HashingTF）来计算，特征哈希是一种处理高维数据的技术，通过哈希方程对特征赋予向量下标，所以在不同情况下，同样的特征就能得到相同的向量下标，这样不需要维护一个特征值及其下标的向量。

数据的特征提取：用 TF-IDF 算法抽取文本特征。将输入的文本数据转化为向量，让计算能够“读懂”文本。

词频矩阵：

```
val hashingTF=new HashingTF()
val tf=hashingTF.transform(words)
tf.cache()
```

特征提取：

```
val idfModel = new IDF().fit(tf)
val tfidf = idfModel.transform(tf)
//tfidf.foreach(println)
val zipped=rate.zip(tfidf)
//zipped.foreach(println)
```

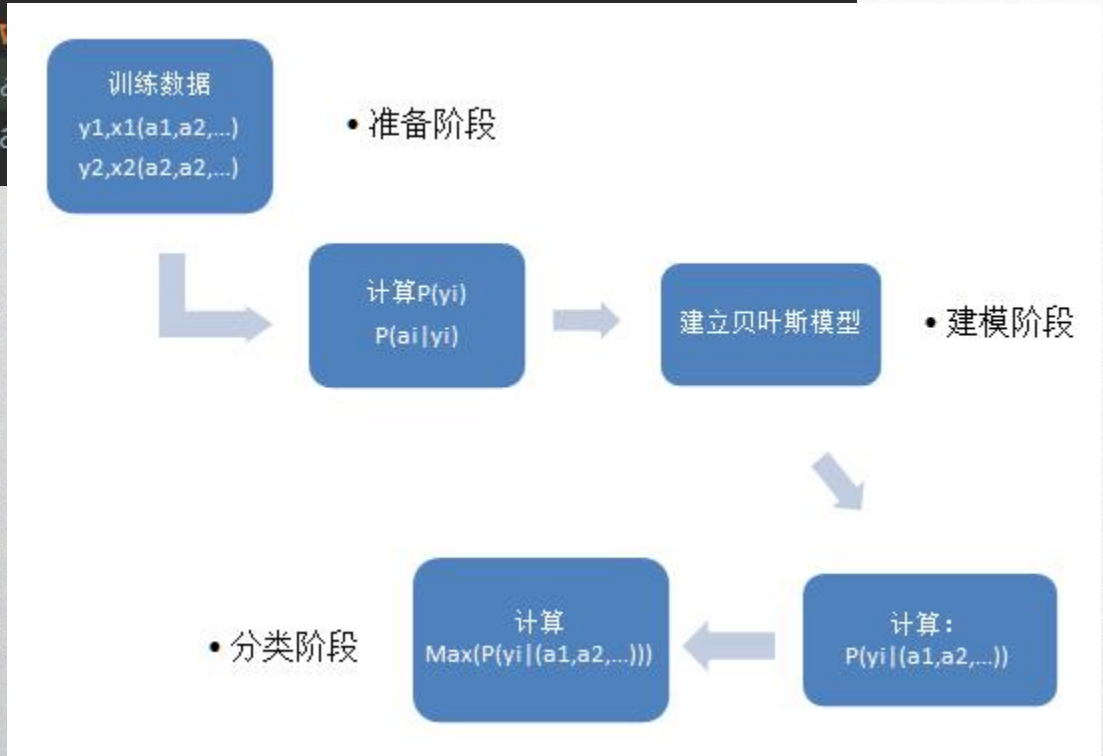
导播系统大数据平台功能——预测用户学历

生成训练集和测试集：

```
val data=zipped.map(line=>ne
val arr=data.randomSplit(Arra
val (training,test)=(arr(0),a
```

✓ 在数据不全面的情况下，贝叶斯方法是一种很好的利用经验帮助作出更合理判断的方法。

✓ 利用贝叶斯算法预测用户学历，流程图如右所示：



导播系统大数据平台功能——预测用户学历

训练贝叶斯分类模型：

```
val NBmodel=NaiveBayes.train(data,1.0)
val predictionAndLabel=data.map(p=>
  (NBmodel.predict(p.features),p.label))
//predictionAndLabel.foreach(println)
```

计算准确率：

```
val accuracy=1.0 * predictionAndLabel.filter{
```

字段	说明
ID	加密后的ID
age	0：未知年龄; 1：0-18岁; 2：19-23岁; 3：24-30岁; 4：31-40岁; 5：41-50岁; 6： 51-999岁
Gender	0：未知1：男性2：女性
Education	0：未知学历; 1：博士; 2：硕士; 3：大学生; 4：高中; 5：初中; 6：小学
Query List	搜索词列表

准确率为：

```
16/10/24 13:41:34 INFO SparkContext: Invoking stop() from shutdown hook
16/10/24 13:41:34 INFO SparkUI: Stopped Spark web UI at http://192.168.31.1:4040
```

导播系统大数据平台功能——预测用户性别

训练SVM分类模型：

```
val SVMmodel=SVMWithSGD.train(training,numIterations = 10)
val SVMPredictionAndLabel=test.map(p=>(SVMmodel.predict(p.features),p.label))
val SVMAccuracy=1.0*SVMPredictionAndLabel.filter{
  x=>x._1==x.
}.count()/tes
println("SVMA
```

字段	说明
ID	加密后的ID
age	0：未知年龄; 1：0-18岁; 2：19-23岁; 3：24-30岁; 4：31-40岁; 5：41-50岁; 6： 51-999岁
Gender	0：未知1：男性2：女性
Education	0：未知学历; 1：博士; 2：硕士; 3：大学生; 4：高中; 5：初中; 6：小学
Query List	搜索词列表

计算准确率：

```
 PROCESS_LOCAL, 2
16/10/23 16:56:08
16/10/23 16:56:08
16/10/23 16:56:08 INFO BlockManager: Found block rdd_14_0 locally
16/10/23 16:56:08 INFO Executor: Finished task 0.0 in stage 7.0 (TID 7). 2082 bytes re
0.921875
16/10/23 16:56:08 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 7) in 75 ms
16/10/23 16:56:08 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all co
```

导播系统大数据平台功能——预测用户年龄

训练随机森林分类模型：

```
val numClasses = 2
val categoricalFeaturesInfo = Map[Int, Int]()
val numTrees = 3 // Use more in practice.
val featureSubsetStrategy = "auto" // Let the algorithm choose.
val impurity = "gini"
val maxDepth = 4
val maxBins = 32

val model = RandomForest.trainClassifier(training, numClasses, categoricalFeaturesInfo,
  numTree

// Evalua
val labe
  val pre
  (point.
}

val testE
println("
```

字段	说明
ID	加密后的ID
age	0：未知年龄; 1：0-18岁; 2：19-23岁; 3：24-30岁; 4：31-40岁; 5：41-50岁; 6： 51-999岁
Gender	0：未知1：男性2：女性
Education	0：未知学历; 1：博士; 2：硕士; 3：大学生; 4：高中; 5：初中; 6：小学
Query List	搜索词列表

大数据竞赛结果

A	B	C	D		
1A51112E076CC0F756997390F61B8BC3	1	1	5		
E7A8424B8852DF8F8F25D92839A3C687	3	1	3		
83D9E9DD405DE73D6BE6FA36724F1D7D	1	2	5		
D0F14597DD7B711F34B4E89D95544604	1	2	5		
CF1B31FC3EB47B797A9C92223583D1AB	4	2	3		
468710F5BC03F31D39AFD4ABF58AEBD5	2	2	1		
9F790C5E718ACAD51DCCFEB9512C72F5	1	2	5		
39093B39BA60BD9FEDD21111BC6E25D1	1	2	3		
7E591F1210CF2900A907B4F02E1E4E2F	4	1	3		
6A869EFE7142CD342CAC627041310552	2	2	5		
30B059D07DB35267A421EA7B2A16821E	1	1	5		
406D9BF56FB6ECF640E71BBC83A2DF1B	3	1	3		
DA51D72CFED67E63349FEB1C438FE63A	3	2	3		
362D72FD11939FD395B58C907C853C96	4	2	3		

导播系统大数据平台功能——用户性格预测

- ✓ 将用户交互信息，评论信息分词
- ✓ 将词语转化成 tf-idf 特征向量
- ✓ 用分类模型来分类。

- ➢ 当用户的交互信息，评论信息类似：“生活没有意思，真没劲, 唉”
  程序输出为“ Predict: 0.0” ，代表消极型。
- ➢ 当用户的交互信息，评论信息类似：“太精彩了，生活真的很美好，我要加油”
  程序输出为“Predict: 1.0” ，代表积极型。

计算准确率为： 74.83%

下周任务

下周任务： 导播系统应用平台功能开发

相关功能实现

补充完善功能点

SHUANGLN

2016

谢谢您的观看