**淘宝行业大类的被搜索数的月度预测**

**——计算思路**

**1. 目的**

预测未来一段时间内（暂定6个月）淘宝各行业大类的关键词搜索数的月度总数。

主要方法：12期滑动平均 + 季节调整

**2. 数据描述**

**1. 数据来源**

测试的原始数据是来自于192.168.7.7下的MySql数据库asto\_ec\_origin中的表*datag\_symc\_shu*与库asto\_ec\_web中的表*linkage*。

**2. In-put 数据形式**

请从原始数据中生成包含如下变量的in-put数据。在前5个变量上保持唯一性。

|  |  |
| --- | --- |
| **变量** | **含义** |
| seg | 淘宝行业大类名称 |
| seg\_detail | 淘宝行业细类类名称 |
| word | 关键词名称 |
| year | 年 |
| month | 月 |
| shu | 各关键词的月度被搜索总次数 |

**2. 数据所需的时间长度**

目前可以使用2011年至今的全部数据。

当我们有更长时间的数据积累后，可以考虑只利用最近5年的数据。

**3.计算过程**

**3.1 在行业大类内部去除完全重复的关键词**

操作：删除变量seg\_detail，然后对全数据去重，保证数据在seg,word,year,month的level上唯一。

说明：各个year, month下，不同seg/seg\_detail下存在一些相同的 word(关键词)，且对应的shu的值完全相同。猜测这是同一 word 在不同seg/seg\_detail下搜索数的总和。

Input file:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| seg | seg\_detail | word | year | month | shu |
| 服饰 | 男装 | **羽绒服** | 2013 | 10 | 8893951 |
| 服饰 | 女装/女士精品 | **羽绒服** | 2013 | 10 | 8893951 |
| 母婴 | 童鞋/亲子鞋 | 连衣裙 | 2013 | 10 | 15707011 |
| 母婴 | 童鞋/亲子鞋 | **羽绒服** | 2013 | 10 | 8893951 |
| 母婴 | 孕妇装/孕产妇用品/营养 | 打底衫 | 2013 | 10 | 8017499 |
| 母婴 | 孕妇装/孕产妇用品/营养 | **羽绒服** | 2013 | 10 | 8893951 |
| 鞋类箱包 | 女鞋 | 靴 | 2013 | 10 | 12506441 |
| 运动户外 | 运动服/休闲服装 | 女装 | 2013 | 10 | 8172394 |

图1 tempSegAndShu

Output file:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| seg | word | year | month | shu |
| 服饰 | **羽绒服** | 2013 | 10 | 8893951 |
| 母婴 | 连衣裙 | 2013 | 10 | 15707011 |
| 母婴 | **羽绒服** | 2013 | 10 | 8893951 |
| 母婴 | 打底衫 | 2013 | 10 | 8017499 |
| 鞋类箱包 | 靴 | 2013 | 10 | 12506441 |
| 运动户外 | 女装 | 2013 | 10 | 8172394 |

图2 segAndShu

**3.2 把重复的关键词的搜索数分配给不同的行业大类**

操作：

3.2.1)不考虑seg变量，将同一year、month下word值唯一的数据存入NoDup（按图2计算）；将word值有重复的数据存入 Dup。

NoDup（输出文件）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| seg | word | year | month | shu |
| 母婴 | 连衣裙 | 2013 | 10 | 15707011 |
| 母婴 | 打底衫 | 2013 | 10 | 8017499 |
| 鞋类箱包 | 靴 | 2013 | 10 | 12506441 |
| 运动户外 | 女装 | 2013 | 10 | 8172394 |

图3 noDup

Dup（输出文件）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| seg | word | year | month | shu |
| 服饰 | **羽绒服** | 2013 | 10 | 8893951 |
| 母婴 | **羽绒服** | 2013 | 10 | 8893951 |

图4 dup1

3.2.2)对file NoDup，取近五年的数据, 在seg level下求shu\_t = sum(shu)（按seg分组后对shu求和），存为SegSum。

SegSum

|  |  |
| --- | --- |
| seg | shu\_t |
| 服饰 | sum1 |
| 母婴 | sum2 |
| 鞋类箱包 | sum3 |
| 运动户外 | sum4 |

图5 segSum

3.2.3)将file SegSum按拼接到Dup上（按seg关键字拼接）。此时Dup形式为：

Dup2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| seg | word | year | month | shu | shu\_t |
| 服饰 | **羽绒服** | 2013 | 10 | 8893951 | sum1 |
| 服饰 | **羽绒服** | 2013 | 11 | 10893951 | sum1 |
| 母婴 | **羽绒服** | 2013 | 10 | 8893951 | sum2 |
| 母婴 | **羽绒服** | 2013 | 11 | 10893951 | sum2 |
| 运动户外 | **羽绒服** | 2013 | 11 | 10893951 | sum4 |

图6 dup2

3.2.4)对file Dup2(图6)，在year、month、word的level下对shu\_t求和，记为变量shu\_tt。计算新变量shu\_n = shu \* (shu\_t / shu\_tt)。

Dup3

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| seg | word | year | month | shu | shu\_t | shu\_tt | shu\_n |
| 服饰 | **羽绒服** | 2013 | 10 | 8893951 | sum1 | sum1+sum2 | shu \* (shu\_t / shu\_tt) |
| 服饰 | **羽绒服** | 2013 | 11 | 10893951 | sum1 | sum1+sum2+sum4 | shu \* (shu\_t / shu\_tt) |
| 母婴 | **羽绒服** | 2013 | 10 | 8893951 | sum2 | sum1+sum2 | shu \* (shu\_t / shu\_tt) |
| 母婴 | **羽绒服** | 2013 | 11 | 10893951 | sum2 | sum1+sum2+sum4 | shu \* (shu\_t / shu\_tt) |
| 运动户外 | **羽绒服** | 2013 | 11 | 10893951 | sum4 | sum1+sum2+sum4 | shu \* (shu\_t / shu\_tt) |

图7 dup3

3.2.6)对file Dup3（图7），删除变量shu，shu\_t，shu\_tt，重命名shu\_n为shu （只需要shu\_n）。

3.2.7)将file NoDup与 Dup纵向拼接(union操作)。记为AllData。

AllData

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| seg | word | year | month | shu |
| 母婴 | 连衣裙 | 2013 | 10 | 15707011 |
| 母婴 | 打底衫 | 2013 | 10 | 8017499 |
| 鞋类箱包 | 靴 | 2013 | 10 | 12506441 |
| 运动户外 | 女装 | 2013 | 10 | 8172394 |
| 服饰 | **羽绒服** | 2013 | 10 | 8893951\*sum1/(sum1+sum2) |
| 母婴 | **羽绒服** | 2013 | 10 | 8893951\*sum2/(sum1+sum2) |

图8 allData

3.2.8)对AllData，在seg、year、month的level下（按seg、year、month分组后）对shu求和，记为变量shu\_seg。生成新file ModelData，保留变量seg、year、month、shu\_seg，在seg、year、month上唯一（如下图）。

ModelData就是清理完毕的数据。

|  |  |  |  |
| --- | --- | --- | --- |
| seg | year | month | shu\_seg |
| 服饰 | 2013 | 10 | sum of Oct |
| 服饰 | 2013 | 11 | sum of Nov |
| 母婴 | 2013 | 10 | sum of Oct |
| 母婴 | 2013 | 11 | sum of Nov |
| 鞋类箱包 | 2013 | 10 | sum of Oct |
| 鞋类箱包 | 2013 | 11 | sum of Nov |
| 运动户外 | 2013 | 10 | sum of Oct |
| 运动户外 | 2013 | 11 | sum of Nov |

图9 modelData

**3.3 计算各行业大类的季节指数**

操作：

3.3.1)对ModelData，在seg、month的level下计算新变量shu\_m = mean(shu\_seg) （对shu\_seg求平均），存为新file SeasonIndex。 File SeasonIndex包含变量seg、month、shu\_m，在seg、month上唯一。

3.3.2)对SeasonIndex，在seg level下计算新变量shu\_y = mean(shu\_m)。

3.3.3)对SeasonIndex的每条记录，计算新变量ratio\_season = shu\_m / shu\_y。

新变量ratio\_season就是所需的季节指数。

SeasonIndex

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| seg | month | shu\_m | shu\_y | ratio\_season |
| 服饰 | 10 | Oct\_mean1 | mean( Jan\_mean1,…Dec\_mean1) | shu\_m/shu\_y |
| 服饰 | 11 | Nov\_mean1 | mean( Jan\_mean1,…Dec\_mean1) |
| 母婴 | 10 | Oct\_mean2 | mean( Jan\_mean2,…Dec\_mean2) |
| 母婴 | 11 | Nov\_mean2 | mean( Jan\_mean2,…Dec\_mean2) |
| 鞋类箱包 | 10 | Oct\_mean3 | mean( Jan\_mean3,…Dec\_mean3) |
| 鞋类箱包 | 11 | Nov\_mean3 | mean( Jan\_mean3,…Dec\_mean3) |
| 运动户外 | 10 | Oct\_mean4 | mean( Jan\_mean4,…Dec\_mean4) |
| 运动户外 | 11 | Nov\_mean4 | mean( Jan\_mean4,…Dec\_mean4) |

图10 seasonIndex

**3.4 预测各行业大类的趋势**

操作：

3.4.1)以变量seg, month 为key，将SeasonIndex横向拼接到ModelData上，记为新file TrendData。

3.4.2)对TrendData，计算新变量 trend = shu\_seg/ratio\_season; 添加新变量 flag = “actual”标识这是真实值。TrendData保留变量seg、year、month、trend，flag。

TrendData (红色的变量shu\_seg，ratio\_season要删除)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| seg | year | month | *shu\_seg* | *ratio\_season* | trend | flag |
| 服饰 | 2013 | 10 | *num1* | *aa* | shu\_seg/ratio\_season | actual |
| 服饰 | 2014 | 10 | *num3* | *aa* | actual |
| 服饰 | 2013 | 11 | *num2* | *bb* | actual |
| 服饰 | 2014 | 11 | *num4* | *bb* | actual |
| 母婴 | 2013 | 10 | *num5* | *cc* | actual |
| 母婴 | 2014 | 10 | *num7* | *cc* | actual |
| 母婴 | 2013 | 11 | *num6* | *dd* | actual |
| 母婴 | 2014 | 11 | *num8* | *dd* | actual |
| 鞋类箱包 | 2013 | 10 | *num9* | *ee* | actual |
| 鞋类箱包 | 2014 | 10 | *num11* | *ee* | actual |
| 鞋类箱包 | 2013 | 11 | *num10* | *ff* | actual |
| 鞋类箱包 | 2014 | 11 | *num12* | *ff* | actual |
| 运动户外 | 2013 | 10 | *num13* | *hh* | actual |
| 运动户外 | 2014 | 10 | *num15* | *hh* | actual |
| 运动户外 | 2013 | 11 | *num14* | *jj* | actual |
| 运动户外 | 2014 | 11 | *num16* | *jj* | actual |

图11 trendData

3.4.3)用12期滑动平均生成趋势预测值：

1)在TrendData中，对每一个seg，取最后 12 个月的记录（数据里的最近12个月），求mean(trend)，作为该seg下个月的trend的预测值。

2)给新生成的trend预测值加上seg、year、month、flag，并给定flag=“forecast”，加到TrendData之中。

3)循环1)、2)两步，直至生成各seg未来6个月的trend的预测，存为TrendForecast。

TrendForecast

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| seg | year | month | trend | flag |
| 服饰 | 2011 | 7 | a | actual |
| … | … | … | … | … |
| … | … | … | … | … |
| 服饰 | 2014 | 6 | n1 | actual |
| 服饰 | 2014 | 7 | n2 | actual |
| 服饰 | 2014 | 8 | n3 | actual |
| 服饰 | 2014 | 9 | n4 | actual |
| 服饰 | 2014 | 10 | n5 | actual |
| 服饰 | 2014 | 11 | n6 | actual |
| 服饰 | 2014 | 12 | n7 | actual |
| 服饰 | 2015 | 1 | n8 | actual |
| 服饰 | 2015 | 2 | n9 | actual |
| 服饰 | 2015 | 3 | n10 | actual |
| 服饰 | 2015 | 4 | n11 | actual |
| 服饰 | 2015 | 5 | n12 | actual |
| 服饰 | 2015 | 6 | n13 = mean(n1,…,n12) | forecast |
| 服饰 | 2015 | 7 | n14 = mean(n2,…,n13) | forecast |
| 服饰 | 2015 | 8 | n15 = mean(n3,…,n14) | forecast |
| 服饰 | 2015 | 9 | n16 = mean(n4,…,n15) | forecast |
| 服饰 | 2015 | 10 | n17 = mean(n5,…,n16) | forecast |
| 服饰 | 2015 | 11 | n18 = mean(n6,…,n17) | forecast |

图12 trendForecast

**3.5 以季节指数调整趋势预测值**

3.5.1)以seg、month为key，将SeasonIndex横向拼接到TrendForecast上，记为FinalForecast。

3.5.2)对FinalForecast的每条记录，计算新变量shu\_frcst = trend \* ratio\_season。

FinalForecast

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| seg | year | month | trend | ratio\_season | shu\_frcst | flag |
| 服饰 | 2011 | 7 | a | r\_07 | trend \* ratio\_season | actual |
| … | … | … | … | … | … |
| … | … | … | … | … | … |
| 服饰 | 2014 | 6 | n1 | r\_06 | actual |
| 服饰 | 2014 | 7 | n2 | r\_07 | actual |
| 服饰 | 2014 | 8 | n3 | r\_08 | actual |
| 服饰 | 2014 | 9 | n4 | r\_09 | actual |
| 服饰 | 2014 | 10 | n5 | r\_10 | actual |
| 服饰 | 2014 | 11 | n6 | r\_11 | actual |
| 服饰 | 2014 | 12 | n7 | r\_12 | actual |
| 服饰 | 2015 | 1 | n8 | r\_01 | actual |
| 服饰 | 2015 | 2 | n9 | r\_02 | actual |
| 服饰 | 2015 | 3 | n10 | r\_03 | actual |
| 服饰 | 2015 | 4 | n11 | r\_04 | actual |
| 服饰 | 2015 | 5 | n12 | r\_05 | actual |
| 服饰 | 2015 | 6 | n13 | r\_06 | forecast |
| 服饰 | 2015 | 7 | n14 | r\_07 | forecast |
| 服饰 | 2015 | 8 | n15 | r\_08 | forecast |
| 服饰 | 2015 | 9 | n16 | r\_09 | forecast |
| 服饰 | 2015 | 10 | n17 | r\_10 | forecast |
| 服饰 | 2015 | 11 | n18 | r\_11 | forecast |
| 母婴 |  |  |  |  |  |  |
| 鞋类箱包 |  |  |  |  |  |  |
| 运动户外 |  |  |  |  |  |  |
| … |  |  |  |  |  |  |

图13 finalForecast

**4. Out-put 数据**

FinalForecast就是最后输出的预测数据，而对flag=“actual”的记录，变量shu\_frcst是真实值；对flag=“forecast”的记录，变量shu\_frcst就是预测的值。

**5. 关于可视化的一些想法**

5.1)行业大类的被搜索数的数值很大，可视化展现时可以考虑取万为单位。

5.2) 可视化时可以考虑让使用者选择部分或全部行业大类；也可以考虑加入鼠标悬停显示数值的功能。

5.3) 可视化时是否可以考虑选择添加趋势线的功能。（数据可以利用trend变量。）