火光摇曳

夜幕降临之际，火光摇曳妩媚、灿烂多姿，是最美最美的… …

# Peacock：大规模主题模型及其在腾讯业务中的应用

2015/03/02 分布式计算、机器学习、自然语言处理 LDA、Peacock、数据并行、模型并行 xueminzhao

Peacock：大规模主题模型及其在腾讯业务中的应用

作者：赵学敏 王莉峰 王流斌 孙振龙 严浩 靳志辉 王益

## 摘要

如果用户最近搜索了“红酒木瓜汤”，那么应该展示什么样的广告呢？从字面上理解，可能应该返回酒水或者水果类广告。可是你知道吗？“红酒木瓜汤”其实是一个民间丰胸秘方。如果机器能理解这个隐含语义，就能展示丰胸或者美容广告——这样点击率一定很高。在广告、搜索和推荐中，最重要的问题之一就是理解用户兴趣以及页面、广告、商品等的隐含语义。

让机器能自动学习和理解人类语言中近百万种语义，以及从海量用户行为数据中归纳用户兴趣，是一个已经持续了20年的研究方向，称为主题建模（Latent Topic Modeling）。目前业界的各种系统中最为突出的是Google Rephil，在Google AdSense广告系统中发挥了重要作用。

追随Google的脚步，腾讯SNG效果广告平台部（广点通）的同学们成功的研发了Peacock大规模主题模型机器学习系统，通过并行计算可以高效地对10亿x1亿级别的大规模矩阵进行分解，从而从海量样本数据中学习10万到100万量级的隐含语义。我们把Peacock系统应用到了腾讯业务中，包括文本语义理解、QQ群的推荐、用户商业兴趣挖掘、相似用户扩展、广告点击率转化率预估等，均取得了不错的效果。

## 一、为什么我们要开发大规模主题模型训练系统Peacock？

### 1.1 短文本相关性

表1 短文本相关性

| |
|---|
| **Q1** (关于"苹果"水果) apple pie |
| **Q2** (关于"苹果"公司) iphone crack |
| **D1** (关于"苹果"公司) Apple Computer Inc. is a well know company located in California, USA. |
| **D2** (关于"苹果"水果) The apple is the pomaceous fruit of the apple tree. |

之所以会出现这种差异，是因为上述文档特征向量构建方法没有"理解"文档的具体语义信息，单纯的将文档中的词表示为一个ID而已。通过主题模型，文档可以表示为一个隐含语义空间上的概率分布向量（主题向量），文档主题向量之间的余弦夹角就可以一定程度上反映文档间的语义相似度了。

## 1.2 推荐系统

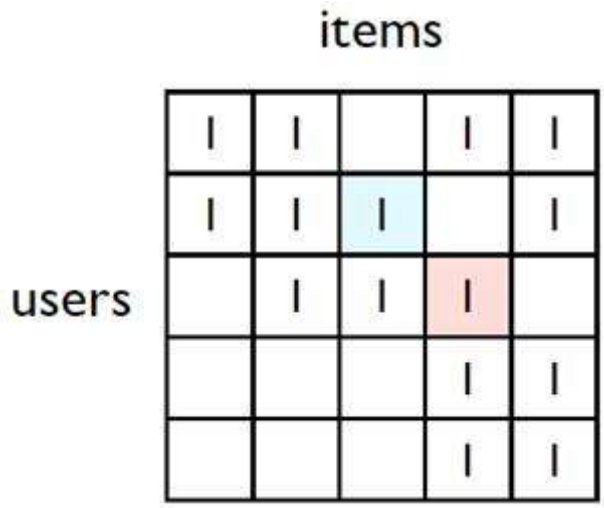
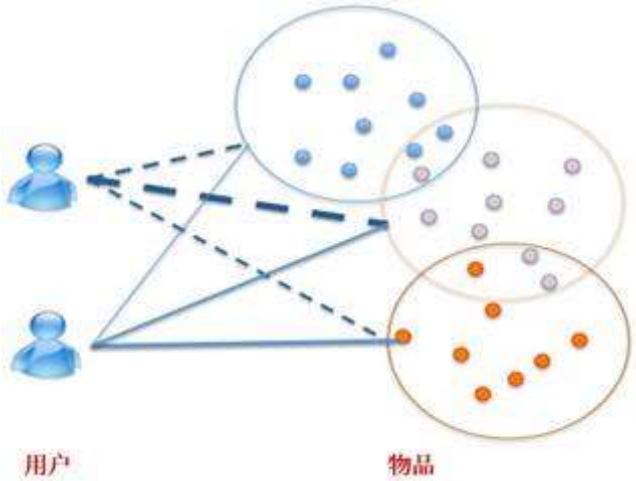



图1 用户-物品矩阵

主题模型的另一个主要应用场景是推荐系统。不管是电商网站的商品推荐，还是各大视频网站的视频推荐等，都可以简化为如下问题：给定用户-物品矩阵（图1，矩阵中用户*u*和物品*i*对应的值表示*u*对*i*的偏好，根据用户行为数据，矩阵会得到部分"初始"值），如何"填满"矩阵中没有值的部分。



以直接利用用户-物品矩阵的*u*行和*v*行，比如计算它们的余弦夹角。然而，真实的互联网数据中，用户-物品矩阵通常都非常稀疏，直接计算不能得到准确的结果。此时，常见的做法是对用户（或物品）进行聚类或者将矩阵投影到更低维的隐空间（图2、3），在隐空间计算用户相似度可以更加准确。主题模型可以用来将用户-物品矩阵投影到隐空间。

- 隐含语义模型 (Latent Factor Model, LFM)[2]。该类方法本质上和主题模型是一致的，直观的理解是将用户-物品矩阵分解为用户-隐含语义（主题）矩阵和隐含语义（主题）-物品矩阵（图3），通过更低维度的上述两个矩阵，来重构原始用户-物品矩阵，重构得到的矩阵将不再稀疏，可以直接用于推荐。具体例子可以参看“QQ群推荐”应用。

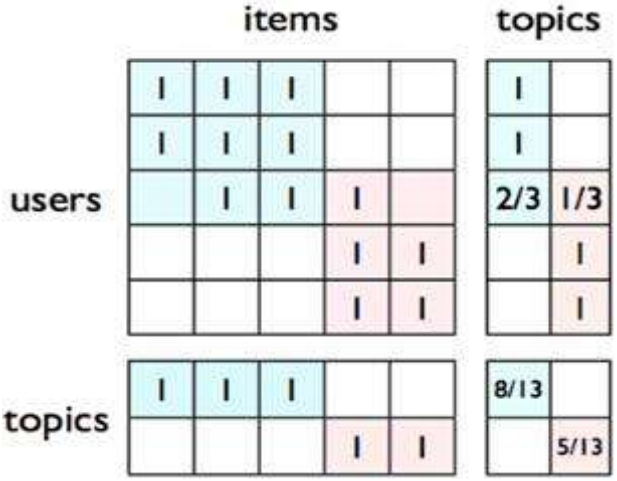



图**3** 用户-物品矩阵分解

实际上，从以上的讨论中我们容易发现，当使用BOW模型处理文本，把文档数据表示成文档-词（Doc-Word）矩阵的时候，其表示结构和用户-物品（User-Item）矩阵结构是完全一致的。因此这两类数据可以使用同样的算法进行处理。使用隐含主题模型处理文档-词矩阵的时候，可以理解为把词聚类为主题，并计算各个文档和词聚类之间的权重。类似地，处理用户-物品矩阵的时候，可以理解为把物品聚类为主题，然后计算每个用户和各个聚类之间的权重。图2是这个过程的一个形象描述，而这个过程如图3所示，可以理解为把原始矩阵分解为两个较小的矩阵：左下的Topic-Item矩阵描述了物品聚类，每行一个主题（Topic）表示一个聚类；而右侧的User-Topic矩阵每一行为主题权重向量，表示每个用户和每个主题的紧密关系。

## 1.3 Peacock是什么？

从上面两个小节我们已经看到，主题模型在互联网产业中具有非常重要的应用。而Peacock系统着手开发时（2012年11月），一些开源以及学术界的主题模型训练系统[5,6,7,8]，要么只能处理小规模的训练语料得到“小模型”，要么模型质量不佳。基于这种状况，我们设计并开发了Peacock系统（更多有关Peacock系统的设计哲学和开发进程，可以参考王益的博客[3]和图灵访谈文章[4]）。Peacock是一个大规模主题模型训练系统，它既可以从数十亿的网络语料中学习出百万级别的隐含语义（主题），也可以对数十亿乘以几亿规模的矩阵进行“分解”。我们的工作总结成论文“Peacock: Learning Long-Tail Topic Features for Industrial Applications”发表在ACM Transaction on Intelligent System and Technology (2015)[15]。


图**4 Peacock**文档语义推断系统**Demo**

文档主题权重 P(topic|doc):

0.170434 4998 苹果(0.230855) 手机(0.12452) iphone(0.0251039) 电脑(0.0171699)
0.086149 6261 范冰冰(0.114506) 苹果(0.0857748) 电影(0.0592054) 视频(0.034497
0.058666 5642 iphone(0.166094) 手机(0.0703596) 3gs(0.0395753) 苹果(0.0330713
0.025660 2134 千克(0.198335) 苹果(0.0784678) 重里(0.0269041) 大米(0.0199628)
0.014673 4966 手机(0.182923) 步步高(0.0826732) 电池(0.0434243) 下载(0.041405
0.012849 4926 蜂蜜(0.0801284) 牛奶(0.0427497) 面膜(0.0303977) 好处(0.0256547
0.011031 3898 圣诞节(0.0992735) 圣诞(0.0514906) 礼物(0.0351853) 祝福语(0.034
0.011005 9480 下载(0.164783) mp4(0.103005) 电影(0.0636672) 视频(0.05098) 格式
0.009197 805 windows(0.0892736) xp(0.088124) 系统(0.0509883) 下载(0.0424573
0.009190 8787 水果(0.0966501) 蔬菜(0.076337) 批发(0.0591047) 市场(0.0500198)

文档词权重 P(word|doc):

0.054041 苹果
0.041369 手机
0.014812 下载
0.014351 iphone
0.010200 电脑
0.009983 范冰冰
0.008309 电影
0.007954 视频
0.007053 价格
0.005963 软件

文档主题权重 P(topic|doc):

```
0.286885  2134    千克(0.198335) 苹果(0.0784678) 重里(0.0269041) 大米(0.0199628)
0.286885  532     桃子(0.0500153) 苹果(0.0364393) 李子(0.0283832) 南方(0.026182
0.104240  6338    糖尿病(0.0811359) 血糖(0.0336291) 高血压(0.0285981) 孕妇(0.021
0.095629  5691    咳嗽(0.0890044) 宝宝(0.0802794) 止咳(0.0497114) 小孩(0.031472
0.095629  3000    开花(0.029144) 果树(0.0262511) 修剪(0.0252923) 技术(0.0240111)
0.073634  1666    年级(0.0567557) 数学(0.0522898) 难题(0.0430788) 答案(0.041869
0.011476  177     奥比岛(0.261746) 奥比(0.0410186) 邮递员(0.0381663) 考试(0.027
0.000959  6156    移动(0.0850844) 套餐(0.0559763) 联通(0.0540878) 动感地带(0.034
0.000957  4998    苹果(0.230855) 手机(0.12452) iphone(0.0251039) 电脑(0.0171699)
```

文档词权重 P(word|doc):

```
0.058426  千克
0.036056  苹果
0.015995  桃子
0.008970  李子
0.008635  咳嗽
0.008599  糖尿病
0.008419  宝宝
0.008340  水果
0.008181  南方
0.007925  重里
```

图**6 Peacock**文档语义推断示例**2**：“苹果 梨子”

文档主题权重 P(topic|doc):

```
0.465717  6261    范冰冰(0.114506) 苹果(0.0857748) 电影(0.0592054) 视频(0.0344978)
0.206565  8602    家具(0.178903) 红木(0.0262665) 价格(0.0161301) 图片(0.0147936) 北
0.095628  4853    医生(0.0711662) 医院(0.0379315) 全身(0.0236984) 儿科(0.0210252) :
0.095628  56      减肥(0.112415) 噪声(0.0355498) 运动(0.0313898) 污染(0.0256226) cu
0.020083  2010    电影(0.16116) 情色(0.0753046) 在线(0.0676254) av(0.0481669) 伦理
0.015301  7563    边城(0.101929) 陆风(0.0856327) 沈从文(0.0575565) x8(0.0409549) xu
0.013388  2743    把握(0.157607) 机会(0.0667866) 作文(0.0167649) 教材(0.0152458) 分
0.010520  5275    价值(0.195102) 药用(0.111382) 收藏(0.0256952) 人生(0.0154372) 植物
0.004781  6264    帝王(0.14446) 慵懒(0.0376641) 超女(0.0317217) 妖孽(0.0275368) su
0.002870  7876    饮料(0.127533) 食品(0.0703096) 公司(0.0695525) 有限(0.040378) 碳酸
```

文档词权重 P(word|doc):

```
0.053949  范冰冰
0.042701  苹果
0.037216  家具
0.031803  电影
0.018056  视频
0.014876  减肥
0.014846  佟大为
0.013595  近义词
0.012089  反义词
0.010294  电视剧
```

图**7 Peacock**文档语义推断示例**3**：“苹果大尺度”

下面我们分别给定一些具体的例子，让大家对Peacock有一些直观上的认识：

- 自然语言处理的例子。图4给出了Peacock在线推断系统Demo的主要界面，手动输入文档

Peacock可以比较准确的理解不同文档的具体含义，这将有助于我们完成一系列自然语言处理和信息检索的任务。

- 用户-物品矩阵分解的例子。这个例子中，“用户”（相当于“文档”）为QQ，“物品”（相当于“词”）为这部分用户加入的QQ兴趣群（在数据预处理中，我们会将QQ群分为关系群、兴趣群等，兴趣群可以比较好的反映用户的兴趣）。取非常活跃的5亿用户和非常活跃的1亿QQ兴趣群，得到一个5亿x1亿的矩阵，使用Peacock分解该矩阵后获得Topic-Item矩阵（即主题-QQ群矩阵），图8、9、10分别给出了该矩阵中的三个主题（只显示权重最高的主要QQ群）。为了方便理解，同时将QQ群的描述信息显示在群ID之后。可以看到，Peacock学习得到的主题含义比较明确，一定程度上可以反映出Peacock在处理用户-物品矩阵上的有效性。

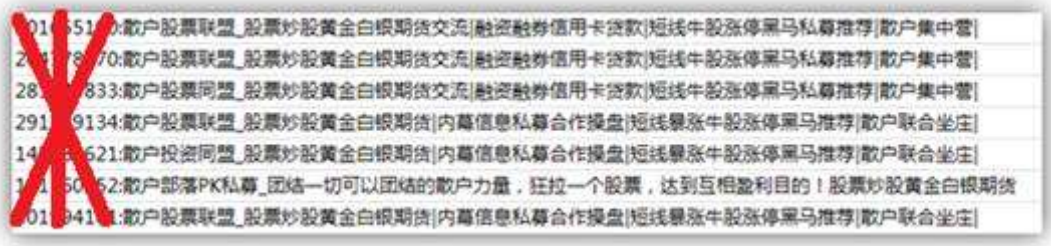



图**8** 基于**QQ-QQ**群**Peacock**矩阵分解示例：炒股类主题

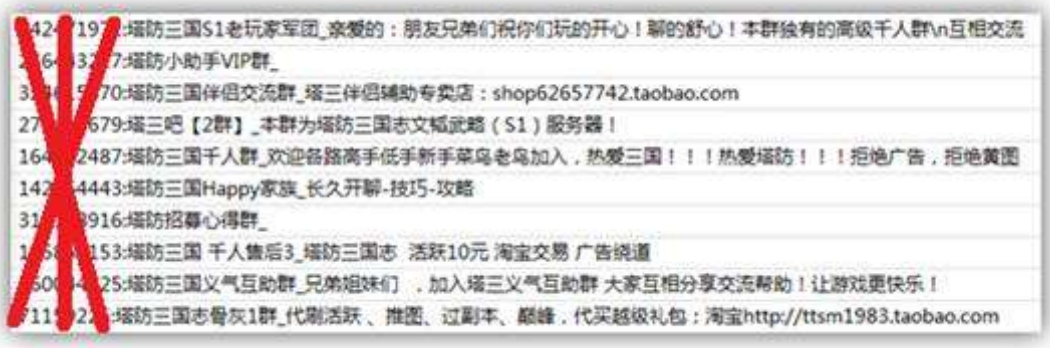



图**9** 基于**QQ-QQ**群**Peacock**矩阵分解示例：塔防三国游戏类主题

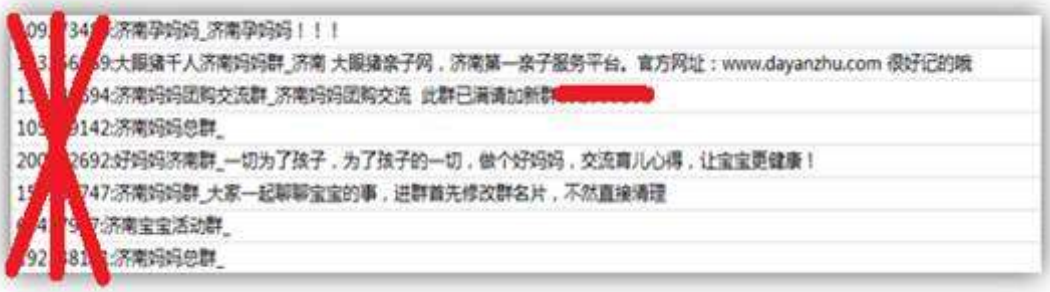



图**10** 基于**QQ-QQ**群**Peacock**矩阵分解示例：济南母婴类主题

通过一些具体的例子直观的介绍了主题模型之后，接下来第二章将主要从算法的角度来回答“什么是主题模型”这个问题，第三章介绍对主题模型并行化的一些思考以及Peacock的具体做法，最后第四章介绍主题模型在腾讯业务中的具体应用。

主题模型一般包含了三个重要的过程：生成过程、训练过程以及在线推断。生成过程定义了模型的假设以及具体的物理含义，训练过程定义了怎样由训练数据学习得出模型，在线推断定义了怎样应用模型。下面分别进行简要介绍。

一般来说，主题模型是一种生成模型（生成模型可以直观的理解为给定模型，可以生成训练样本）。给定模型，其生成过程如图11：

- 模型有2个主题，主题1关于银行（主要的词为loan、bank、money等），主题2关于河流（主要的词为river、stream、bank等）。
- 文档1内容100%关于主题1，主题向量为<1.0, 0.0>，文档中每一个词的生成过程如下：以100%的概率选择主题1，再从主题1中以一定的概率挑选词。
- 文档2内容50%关于主题1，50%关于主题2，主题向量为<0.5, 0.5>，文档中每一个词的生成过程如下：以均等的概率选择主题1和2，再从选中的主题中以一定的概率挑选词。
- 文档3内容100%关于主题2，主题向量为<0.0, 1.0>，文档中每一个词的生成过程如下：以100%的概率选择主题2，再从主题2中以一定的概率挑选词。

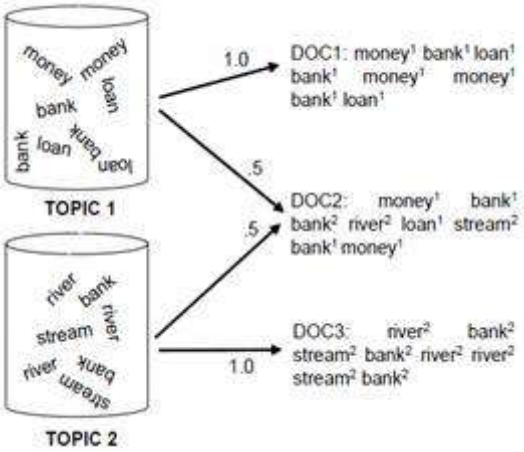
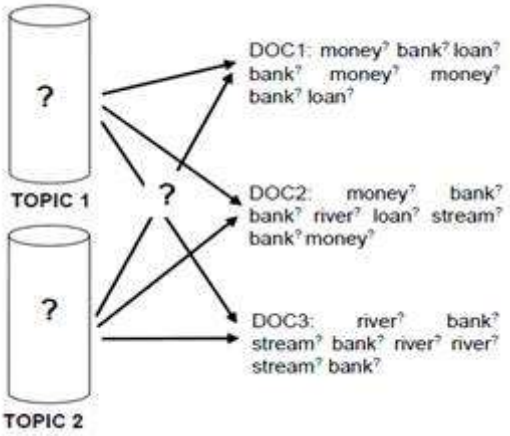



图**11** 主题模型的生成过程[9]

现实的情况是我们没有模型，只有海量的互联网文档数据，此时我们希望有机器学习算法可以自动的从训练文档数据中归纳出主题模型（如图12），即得到每个主题在词表上的具体分布。通常来说，训练过程还会得到一个副产品——每篇训练文档的主题向量。



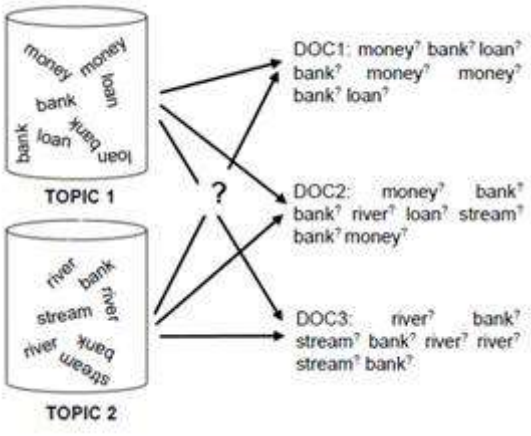



图**13** 主题模型的在线推断

三个过程中，训练过程是难点，后文将进行重点介绍。

## 2.2 LDA模型及其训练算法

LDA（Latent Dirichlet Allocation）[10]作为一种重要的主题模型，自发表以来就引起了学术界和产业界的极大关注，相关论文层出不穷。LDA的训练算法也多种多样，下面以吉布斯采样[11,12]为例，进行简要介绍。

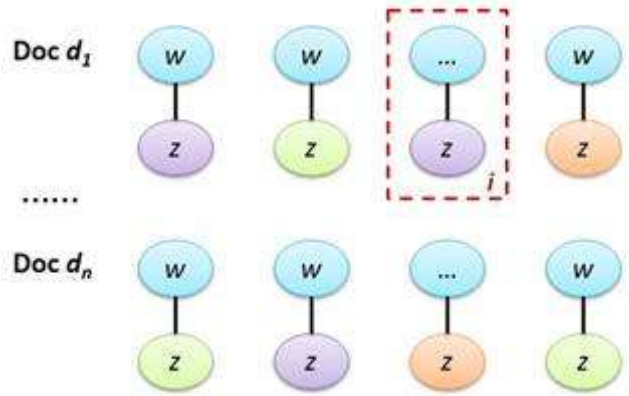



图**14 LDA**训练过程

跳过复杂的数学推导，基于吉布斯采样的LDA训练过程如图14所示（每个词用*w*表示，每个词对应的主题用*z*表示，图中节点*z*的不同颜色表示不同的主题）：

- Step1: 初始时，随机的给训练语料中的每一个词*w*赋值一个主题*z*，并统计两个频率计数矩阵：Doc-Topic计数矩阵$N_{td}$，描述每个文档中的主题频率分布；Word-Topic计数矩阵$N_{wt}$，表示每个主题下词的频率分布。如图15所示，两个矩阵分别对应于图中的边上的频率计数。
- Step2: 遍历训练语料，按照概率重新采样其中每一个词*w*对应的主题*z*，同步更新$N_{wt}$和$N_{td}$。
- Step3: 重复 step2，直到$N_{wt}$收敛。

St 2中重新采样词 对应主题 时 采样公式为

数，$N_{td}$表示文档*d*中主题*t*的出现次数，上角标¬表示剔除当前采样词*w*的影响（比如$N_{td}^{\neg}$表示减去当前采样词对应的主题后，文档*d*中主题*t*的出现次数）。

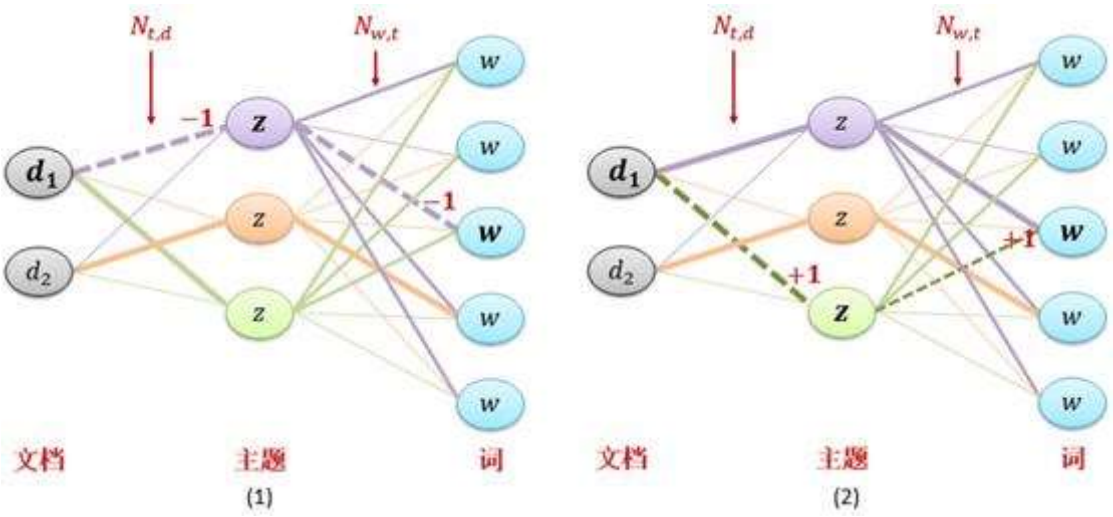



图**15** 文档$d_1$中词***w***主题重新采样

事实上，以上对文档*d*中词w的主题z进行重新采样的公式有非常明确的物理意义，表示P(w|z)P(z|d)，可以如图15直观的表示为一个“路径选择”的过程：

- 对当前文档*d*中的当前词*w*（图15中黑体表示），词*w*的“旧”主题*z*给出了*d*-*z*-*w*的一条路径（图15（1）虚线）；
- 剔除词*w*对应的“旧”主题*z*，更新在$N_{wt}$和$N_{td}$中的计数（图15（1）在旧路径对应的两条边上做 “-1”操作）；
- 计算*d*-*z*-*w*的每一条可能路径的概率，*d*-*z*-*w*路径的概率等于*d*-*z*和*z*-*w*两部分路径概率的乘积即P(z|d)P(w|z)，P(z|d)和$N_{td}$有关，P(w|z)和$N_{wt}$有关（图15（1））；
- 依据概率对*d*-*z*-*w*路径进行采样，得到词*w*的“新”主题*z*（图15（2）虚线）；
- 增加词*w*对应的“新”主题*z*，更新在$N_{wt}$和$N_{td}$中的计数（图15（2）在新路径对应的两条边上做“+1”操作）。

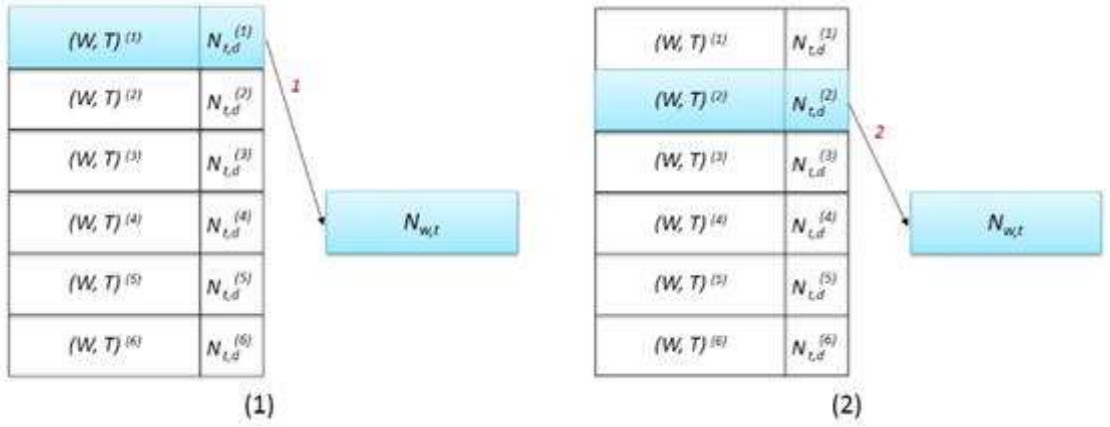



图**16** 单机版**LDA**训练过程

在训练模型时，为了包含尽可能多的隐含语义（主题）同时保证效果，通常会使用海量的训练语料。这些互联网原始文档语料经过切词、停用词过滤、文档过滤（长度）等预处理步骤后（通常

- 采样完成一个数据块后，训练器将更新后的*(W,T)*和$N_{td}$序列化到磁盘上，供下一个迭代加载采样；
- 所有迭代结束，$N_{wt}$收敛，训练器根据$N_{wt}$计算出模型并输出。

基于吉布斯采样的LDA在线推断过程与训练过程（图14）类似：给定文档，采样更新其中每一个词*w*对应的主题*z*（采样公式同上，采样过程中可以保持模型$N_{wt}$不变）；重复上述过程，直到文档主题直方图$N_{td}$收敛，使用$\alpha_t$对其进行简单平滑即为文档主题向量。

# 三、十亿文档、百万词汇、百万主题？

从上一个小节的算法描述中，我们可以看到LDA 的训练算法貌似并不复杂，主要的工作就是在维护两个频率计数矩阵$N_{wt}$和$N_{td}$。然而在这个时代，我们要面对的是互联网的海量数据，想象一下，如果在图15中，左边的文档节点是十亿、中间的主题个数是百万、右边不同的词的个数也是百万，我们将需要处理一张多大的图！！！在实际应用中，我们希望使用更多的数据训练更大的模型，这包含了两重意思：

- “更多的数据”，我们希望训练器能处理海量的训练数据，因为更多的数据蕴含着更加丰富的隐含语义，同时模型也更加准确，效果更好。上一小节提到单机版LDA训练器显然是处理不了海量数据的，使用它训练模型，我们估计要等到天荒地老了。
- “更大的模型”，我们希望训练器能归纳出更多更具体更长尾的隐含语义，比如一百万主题。抛开标准LDA算法本身的问题，更大的模型意味着矩阵$N_{wt}$规模更大。$N_{wt}$的大小为*V*x*K*，*V*表示词表大小，*K*表示主题个数。取*V=1,000,000*且*K=1,000,000*，$N_{wt}$需要消耗3000G以上内存（假设int型密集存储，因为模型随机初始化并不稀疏），显然单机内存是无法满足需求的，必须对模型进行切分。

下面分别从数据并行和模型并行两个方面来介绍怎样解决上述两个问题。“数据并行”和“模型并行“是Google大神Jeff Dean在深度学习训练系统DistBelief[13]中新提出的两个概念，尽管Peacock系统开发的时候，DistBelief还没有真正对外公布。随着深度学习的持续升温，大家现在已经逐渐熟悉了这两个形象的名词，此处请允许我们借用一下这两个概念。

## 3.1 数据并行——处理更多的数据

“数据并行”通俗的理解：通过多任务（每个任务都包含一份完整的模型）并行的处理数据训练模型，任务之间的模型或同步或异步的进行融合。借用王益[3]的说法，“如果一个算法可以做数据并行，很可能就是可扩展的了”。幸运的是，David Newman团队发现基于吉布斯采样的LDA训练算法可以“数据并行”，并给这个算法取了一个名字叫AD-LDA[14]。

注意，AD-LDA算法是吉布斯采样的近似算法，因为严格的吉布斯采样要求串行采样，不能并行。直观的理解就是语料中前一个词$w_1$采样更新后的$N_{wt}$和$N_t$应该应用于后一个词$w_2$的采样，而不是$w_1$和$w_2$的采样都基于相同状态的$N_{wt}$和$N_t$。AD-LDA算法会使得LDA的训练收敛速度变慢，但在多几轮迭代后，AD-LDA算法可以收敛到与串行吉布斯采样相同的点。

AD-LDA算法的整个过程和MapReduce的执行过程非常一致，所以早期有非常多的团队使用MapReduce来实现AD-LDA算法[5]：

- MapReduce的一个Job进行AD-LDA算法的一个迭代；
- 训练语料数据块*(W,T)*和$N_{td}$作为Job输入，Mapper加载上个迭代生成的$GN_{wt}$作为$LN_{wt}$，对数据块中的词进行主题采样；
- Reducer融合各个$LN_{wt}$，生成下一个迭代需要加载的$GN_{wt}$。

因为MapReduce使用磁盘进行数据交换，同时整个训练任务需要调度几百个Jobs，所以基于MapReduce的AD-LDA实现是非常低效的。

## 3.2 模型并行——训练更大的模型

上文提到，训练大模型时，$N_{wt}$太大而无法整体放入任务的内存，直观的解决方法如图18所示，将$N_{wt}$沿词的维度进行分片，每个采样任务只加载一个模型分片$N_{wt}^{(i)}$。相应的，语料数据块也需要做对应的词维度切分，因为单个任务*i*只能采样$N_{wt}^{(i)}$包含的词*w*。细心的童鞋可能已经发现，图18所示的模型并行方式在$N_{td}$上采用了类似AD-LDA算法的近似，$LN_{td}$间的融合与$LN_{wt}$间的融合类似，相应的算法也会减缓收敛（因为$N_{wt}$是所有训练语料上的聚合结果，而$N_{td}$只和具体文档*d*有关，后者变化比前者更加“快速”，$N_{td}$的并行近似采样更加“危险”，很容易造成训练不收敛）。

图**18** 模型并行**1**

有没有办法不进行$N_{td}$的并行近似采样，同时保持上述的模型切片方式呢？Peacock系统设计了图19所示的并行采样方式：加载了不同$N_{wt}^{(i)}$切片的任务并行的沿对角线方向对训练语料数据块*(W,T)*进行采样，一条对角线采样完成后，依次进行下一条对角线。这样在对同一个文档的不同数据块间的词进行采样时，仍然保持了“串行性”，应用了之前数据块中的词对$N_{td}$的更新。图19的模型并行采样方式收敛性同AD-LDA是一致的。

图**19** 模型并行**2**

## 3.3 大规模主题模型训练系统**Peacock**

为了“利用更多的数据训练更大的模型”，Peacock系统结合了上述的“数据并行”和“模型并行”（图20）：

- 多组“模型并行”任务之间采用“数据并行”的方式工作，“模型并行”任务组内部，依然保持图19所示的并行采样方式；
- 在迭代结束或任务处理训练语料数据块过程中，不同“模型并行”任务组之间或同步或异步的融合模型分片$LN_{wt}^{i}$。模型融合的方式可以类似MPI中的AllReduce，也可以借助全局的参数服务器$GN_{wt}^{i}$。

- 数据传输和文档采样之间的流水线。
- 图19所示的模型并行方式在每条对角线并行采样结束后都需要同步，怎样去掉这种同步？
- 怎样的模型$N_{wt}$分片方式，能尽可能的保证采样服务器之间的负载均衡？
- 我们是否需要每个迭代都重采样所有词的主题？
- 怎样快速的计算对数似然度？
- 怎样将模型的超参数$\alpha_t$和$\beta$优化融入Peacock系统？
- 除了标准的吉布斯采样，是否有更加快速的采样算法？
- 主题数$K$从100到1,000,000，系统的内部数据结构都保持不变么？

在我们的论文[15]中，部分的解答了上述问题，更详细的Peacock解密请关注我们的博客“火光摇曳”[16]^_^。

# 四、Peacock在腾讯都有哪些应用？

## 4.1 文本语义分析

为了理解互联网上海量、多样化、非结构化的自然语言描述的文本，我们通常会从词法、句法、语义等维度进行分析。受限于文本字面信息量小，存在歧义现象，词法和句法分析容易遭遇Vocabulary Gap的问题，从海量文本数据中归纳 “知识”，从语义角度帮助理解文本，是一种非常重要的途径。

图**21** 文本分析示例

例如，对于输入文本 “红酒木瓜汤效果怎么样？”，根据人的背景知识，很容易猜到这是一位女性用户在询问丰胸产品“红酒木瓜靓汤”的效果。对于机器而言，通常会先进行词法分析，对原始文本做切词、词性标注、命名实体识别等，然后使用词袋模型（Bag of Words，BOW）或提取关键词来表示文本。不难发现，从字面抽取的信息，很容易理解成“红酒”、“木瓜”等餐饮类语义，并非原始文本真实的意思。当然，我们可以对关键词做扩展，给出一些相似的词条，但是，更好的是直接理解语义。一种常见的方法是文本分类，由于对标注语料库的依赖，类别规模一般不会太大，粒度较粗。还有一种方法就是文本聚类，挖掘语义主题标签，更细粒度的理解文本意思，隐含语义分析技术逐渐发展成为常用的解决方案。能够从十亿级别的文档中归纳上百万语义的Peacock系统更是在腾讯广点通广告系统扮演着核心角色。这些不同维度的文本分析模块，包括词袋、关键词提取、关键词扩展、文本分类和Peacock等（图21），整合在一起构成了我们理解语言的基础文本分析平台TextMiner（图22）。

图**22** 文本分析平台**TextMiner**

### 4.1.1 文本分类器

文本分类是一个典型的有监督的机器学习任务，我们在做在线广告系统过程中遇到的任务就有许多，包括网页分类、广告分类、QQ群分类、用户兴趣分类等。在使用相同的标注数据集和机器

### 4.1.2 相关性计算

对给定的查询语句，搜索引擎会将检索到的网页进行排序，把相关性好的排在前面。同样的，在线广告系统应该保证展示给用户的广告与页面内容、用户兴趣相关，以尽量不影响用户体验。这里都涉及到一个共同的任务：排序学习。此问题通常被形式化为有监督的学习问题，我们会将查询、网页、用户、广告表示成语义特征向量，从而在语义空间里比较用户意图（查询、网页内容、用户历史行为）和网页、广告的相关性。

Peacock已成功应用在腾讯搜索广告和情境广告中，用于分析文本数据，归纳自然语言的语义，从而更好地匹配查询词和广告，以及页面内容和广告。在情境广告 Learning To Rank 相关性计算框架下，增加Peacock语义特征后，NDCG@5提升达8.92%，线上A/B Test实验 AdCTR 提升8.82%。相关性评估效果图24所示。

图**24** 情境广告相关性（相关性标注样本包括**4,000** 查询，**200,000**对**(**查询**,** 广告**)**，标注**0~3**四档打分）

## 4.2 广告CTR预估

广告点击率预估是预测给定场景下一个广告被点击的概率：P(click=1|ad,user,context)，user表示当前用户，context表示当前的环境信息，譬如当前所在的网页。点击率预估是在线广告系统最核心的技术之一，它决定着广告的排序和计价。

业界一般做法是将广告展示、点击日志作为训练数据，抽取特征，通过机器学习方法拟合训练数据得到预估模型，进而做在线点击率预估。选取有效的特征对得到一个精准的点击率预估模型起着至关重要的作用。

Peacock是我们理解广告语义的关键技术，被引入到广告点击率预估模型中提升效果。具体的，与KDD Cup 2012 Track2的数据集产生过程类似，我们使用了腾讯情境广告系统的广告展示、点击日志，使用L1范数正则的逻辑回归训练预估模型，通过AUC评估模型精度。Baseline使用一些基础特征，优化实验分别在baseline特征集合的基础上引入主题规模为1000、10,000和100,000的Peacock Top-N语义特征。

图**25 pCTR**增加不同粒度**topic**特征模型**AUC**的提升

从图25可以看出，加入Peacock语义特征后AUC得到了显著提升，尤其当增加topic规模为100,000的Peacock语义特征时，AUC提升最大，约为1.8%，线上A/B Test实验AdCTR有8.82%的提升。

## 4.3 精准广告定向

们利用Peacock系统对上述用户-物品做矩阵分解（如图3），从不同数据来源，多视角理解用户兴趣，进而挖掘相似用户，提供给广告主丰富的定向策略，如用户商业兴趣定向、关键词定向和Look-Alike定向等。同时，获取到的用户特征，也可以作为广告CTR、CVR预估系统的重要特征。

## 4.4 QQ群推荐

图**26 QQ**群推荐

根据用户已加QQ群社交关系数据，利用Peacock对QQ-QQ群做矩阵分解，我们发现语义相近的QQ群被比较好的归到了相同的主题下，如图8、9、10所示。非常直观的，我们将Peacock 模型应用在QQ群消息面板推荐产品中（如图26），相比基于QQ好友关系链的推荐算法，推荐群的点击率和转化率（即点击后是否加入了该群）均有2~3倍的提升（图27）。

$$P(QQGroup \mid user) = \sum_{topic} P(QQGroup|topic) \cdot P(topic|user) \qquad (2)$$

图**27 QQ**群推荐效果

## 后记

LDA是一个简洁、优雅、实用的隐含主题模型，腾讯效果广告平台部（广点通）的工程师们为了应对互联网的大数据处理，开发了大规模隐含主题模型建模系统，并在腾讯的多个业务数据中得到了应用。本文由赵学敏、王莉峰、王流斌执笔，靳志辉、孙振龙等修订，相关工作由腾讯SNG效果广告平台部（广点通）质量研发中心Peacock团队王益、赵学敏、孙振龙、严浩、王莉峰、靳志辉、王流斌为主完成，苏州大学曾嘉教授、实习生高阳，香港科技大学杨强教授等持续大力支持，是多人合作的结果。

## 参考文献


- [1] Greg Linden, Brent Smith, and Jeremy York. *Amazon.com Recommendations: Item-to-Item Collaborative Filtering*. IEEE Internet Computing, 2003.
- [2] Simon Funk. *Netflix Update: Try This at Home*. http://sifter.org/~simon/journal/20061211.html.
- [3] 分布式机器学习的故事. http://cxwangyi.github.io/2014/01/20/distributed-machine-learning/.
- [4] *LinkedIn*高级分析师王益*:* 大数据时代的理想主义和现实主义*(*图灵访谈*).* http://www.ituring.com.cn/article/75445.
- [5] *PLDA and PLDA+*. https://code.google.com/p/plda/.
- [6] Arthur Asuncion, Padhraic Smyth, and MaxWelling. *Asynchronous Distributed Learning of Topic Models*. NIPS'2008.
- [7] *Yahoo LDA* https://github com/shravanmn/Yahoo LDA

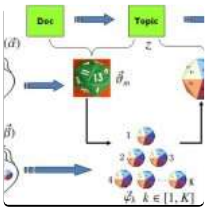
Ng. *Large Scale Distributed Deep Networks.* NIPS'2012.

- [14] David Newman, Arthur Asuncion, Padhraic Smyth, and MaxWelling. *Distributed Algorithms for Topic Models.* JMLR'2009.
- [15] Yi Wang, Xuemin Zhao, Zhenlong Sun, Hao Yan, Lifeng Wang, Zhihui Jin, Liubin Wang, Yang Gao, Ching Law, and Jia Zeng. *Peacock: Learning Long-Tail Topic Features for Industrial Applications.* TIST'2015.
- [16] 火光摇曳. http://www.flickering.cn/.

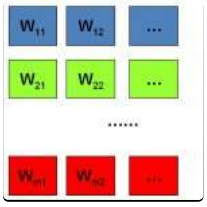
返回顶部



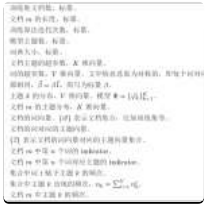
## 相关文章



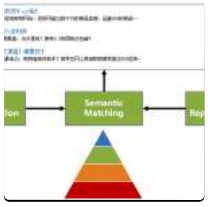

2014/06/17
[LDA数学八卦-5]LDA 文本建模




2014/06/17
[LDA数学八卦-4]文本建模



2014/07/21
[LDA工程实践之算法篇-1]算法实现正确性验证




2014/06/19
[我们是这样理解语言的-1]文本分析平台TextMiner

2015/01/28
Peacock: 大规模主题模型及其在腾讯业务中的应用

2015年3月9日 回复 顶 转发

返回顶部

每半小时汇报一次

LDA实际应用。

2015年4月2日 回复 顶 转发

linna_angle林娜

2015年10月26日 回复 顶 转发

宫保胖丁丁

讚

8月23日 回复 顶 转发

社交帐号登录:

微信 微博 QQ 人人 更多»

说点什么吧...

发布

火光摇曳正在使用多说

分享到...