Peacock: 大规模主题模型及其在腾讯业务中的应用

Rickjin(靳志辉)

腾讯SNG效果广告平台部

Outline

- **Peacock Demo**
- 主题模型背景介绍
- 大规模主题模型学习系统 **Peacock**
- **Peacock** 在腾讯业务中的应用

- 红酒木瓜汤
- 苹果
- 莫代尔

Peacock Demo

query:

红酒木瓜汤

Submit

tokens:

红酒(0.589038) 木瓜(0.582175) 汤(0.560452)

topics:

id(rank)	weight	topic_words
6147(3672)	0.904523	丰胸(0.170997) 产品(0.080866) 减肥(0.067258) 木瓜(0.048380) 效果(0.036604) 红酒(0.015934) 广告(0.009820) 女人(0.009058) 有用(0.008605) 快速
6338(325)	0.301545	糖尿病(0.081618) 血糖(0.033829) 高血压(0.028768) 孕妇(0.021932) 血压(0.021665) 患者(0.016755) 空腹(0.015495) 减肥(0.014270) 饮食(0.013987)
8009(3430)	0.301511	奇迹(0.247384) 世界(0.081658) 加点(0.037639) 木瓜(0.037639) mu(0.036604) 战士(0.023355) 私服(0.018500) 装备(0.011273) 武器(0.008356) 9
8443(1)	0.000121	游戏(0.268936) 下载(0.059112) 单机(0.057830) 双人(0.015077) 在线(0.010757) 网络(0.010431) 射击(0.010054) 好玩(0.008341) 连连看(0.008340) 免
9127(5)	0.000111	美女(0.077413) 视频(0.057143) 偷拍(0.045182) 做爱(0.043915) 自拍(0.037817) 图片(0.035592) mm(0.028053) 性感(0.025718) 激情(0.019556) 动漫(
5114(9)	0.000111	美女(0.112125) 丝袜(0.086679) 性感(0.064582) 视频(0.043439) 图片(0.040112) 裸体(0.016823) mm(0.014740) 高跟(0.014448) 日本(0.010446) 强奸(
6433(4)	0.000110	qq(0.248970) 农场(0.076421) 空间(0.076398) 好友(0.037156) 游戏(0.019570) 炫舞(0.014494) 日志(0.013880) 外挂(0.012754) 分组(0.012626) 9
3072(11)	0.000105	日本(0.118981) av(0.095897) 女优(0.059665) 电影(0.045194) 视频(0.019134) qvod(0.017863) 美女(0.017254) 成人(0.016873) 色情(0.015151) 下载(

query:

苹果

Submit

tokens:

苹果(1.000000)

topics:

id(rank)	weight	topic_words
4998(1487)	0.833025	苹果(0.234488) 手机(0.126480) iphone(0.025499) 电脑(0.017440) 价格(0.015992) 3gs(0.012293) 美国(0.011246) 报价(0.010444) mp3(0.010239)
6261(1002)	0.439990	范冰冰(0.115817) 苹果(0.086757) 电影(0.059883) 视频(0.034893) 修大为(0.031876) 近义词(0.029192) 反义词(0.025958) 电视剧(0.021859) 主演(0.02
5642(601)	0.243490	iphone(0.167452) 手机(0.070935) 3gs(0.039899) 苹果(0.033342) 3g(0.029012) 软件(0.025317) 下载(0.022439) 越狱(0.014688) wifi(0.011495)
2134(2601)	0.093624	千克(0.203649) 苹果(0.080570) 重量(0.027625) 大米(0.020498) 水果(0.015943) 面粉(0.015058) 质量(0.014013) 剩下(0.013518) 卖出(0.013049) 香
4926(452)	0.084451	蜂蜜(0.080695) 牛奶(0.043052) 面膜(0.030612) 好处(0.025836) 鸡蛋(0.024024) 孕妇(0.019607) 酸奶(0.018261) 减肥(0.016700) 黄瓜(0.014269) 作
4754(84)	0.065861	上网(0.094976) 无线(0.087973) 3g(0.051667) 手机(0.051194) 电信(0.040308) 上网卡(0.036969) 宽带(0.026859) 天翼(0.024700) 电脑(0.022207) 笔
8787(202)	0.056435	水果(0.097108) 蔬菜(0.076698) 批发(0.059384) 市场(0.050257) 价格(0.027530) 北京(0.007812) 盖子(0.007464) 减肥(0.006896) 礼品(0.006456) 种
4966(715)	0.056294	手机(0.184572) 步步高(0.083418) 电池(0.043816) 下载(0.041779) 音乐(0.028108) 诺基亚(0.023390) 主题(0.020511) qq(0.017884) 充电器(0.016438
3000(1601)	0.056240	开花(0.029634) 果树(0.026692) 修剪(0.025717) 技术(0.024415) 桃树(0.023246) 嫁接(0.020145) 管理(0.018708) 苹果树(0.018389) 梨树(0.018212)
628(563)	0.046953	电脑(0.127480) 笔记本(0.121242) 联想(0.103747) 手机(0.020950) 天逸(0.014518) 三星(0.013794) 东芝(0.013437) 系统(0.012250) 键盘(0.011759)
4427(1860)	0.046862	ipod(0.158706) touch(0.114535) nano(0.053839) pro(0.031299) 下载(0.031012) htc(0.023037) itunes(0.022234) 苹果(0.016381) 软件(0.01516

query:

苹果大尺度

Submit

---

tokens:

尺度(0.852807) 苹果(0.479783) 大(0.206226)

---

topics:

id(rank)	weight	topic_words
6261(1002)	0.995250	范冰冰(0.115817) 苹果(0.086757) 电影(0.059883) 视频(0.034893) 佟大为(0.031876) 近义词(0.029192) 反义词(0.025958) 电视剧(0.021859) 主演(0.020858)
5089(1730)	0.066110	沙发(0.174840) 真皮(0.030333) 图片(0.022596) 家具(0.020159) 价格(0.018055) 布艺(0.015374) 汽车(0.013144) 成都(0.011555) 颜色(0.011501) 皮沙发(0
6528(2373)	0.061118	风格(0.075817) 设计(0.051258) 图片(0.031166) 欧式(0.030241) 客厅(0.029511) 田园(0.019228) 大门(0.018108) 建筑(0.015066) 玄关(0.013679) 风水(0.0
6984(1353)	0.021215	尺寸(0.200698) 标准(0.052568) 规格(0.026346) 照片(0.022821) 大小(0.014162) 公差(0.014148) 图片(0.007800) 设计(0.006969) 螺纹(0.005850) 纸张(0.0
5275(1996)	0.021211	价值(0.199099) 药用(0.113664) 收藏(0.026221) 人生(0.015753) 植物(0.011953) 取向(0.010765) 用途(0.006986) 价格(0.006146) 人民币(0.006043) 中药(0
2743(7012)	0.011226	把握(0.176789) 机会(0.074915) 作文(0.018805) 教材(0.017101) 分析(0.016370) 人生(0.014362) 江映蓉(0.012293) 准确(0.011441) 重点(0.011137) 分寸(0
1206(3111)	0.008735	空间(0.178098) qq(0.163102) 模块(0.080496) 图片(0.033547) 留言(0.022237) 主流(0.021806) 动画(0.019730) 7
6799(5833)	0.008732	下载(0.155759) 播放器(0.083595) 视频(0.050489) mv(0.041266) 3gp(0.038037) mp4(0.034722) qvodplayer(0.030812) 手机(0.030089) 电影(0.028984)
7831(3322)	0.006241	灯笼(0.159440) 制作(0.083158) 大红(0.044755) 图片(0.019980) 手工(0.018532) 视频(0.016957) 歇后语(0.011745) 故事(0.010750) 爱情(0.009447) 批发(0
334(500)	0.005011	园林(0.069168) 绿化(0.043521) 树木(0.039447) 景观(0.037156) 植物(0.028347) 设计(0.025391) 花草(0.019898) 花卉(0.019174) 养护(0.014606) 种植(0.0
2134(2601)	0.004993	千克(0.203649) 苹果(0.080570) 重量(0.027625) 大米(0.020498) 水果(0.015943) 面粉(0.015058) 质量(0.014013) 剩下(0.013518) 卖出(0.013049) 香蕉(0.0

query:

苹果价格

Submit

---

tokens:

苹果(0.835558) 价格(0.549402)

---

topics:

id(rank)	weight	topic_words
4998(1487)	0.501161	苹果(0.234488) 手机(0.126480) iphone(0.025499) 电脑(0.017440) 价格(0.015992) 3gs(0.012293) 美国(0.011246) 报价(0.010444) mp3(0.010239) 下载(
53(413)	0.469074	手机(0.110538) 报价(0.077922) 诺基亚(0.071693) 三星(0.067755) 水货(0.062980) 行货(0.056647) 价格(0.022659) 最新(0.019562) 索爱(0.015023) 9
4160(2781)	0.462599	电脑(0.169847) 笔记本(0.131029) 英寸(0.035949) 分辨率(0.031167) 显示屏(0.016895) 联想(0.014707) 屏幕(0.014678) 报价(0.012591) 索尼(0.012418) 苹
9186(1339)	0.353394	技术(0.102576) 栽培(0.096336) 种植(0.046962) 视频(0.017818) 管理(0.015950) 玉米(0.015184) 食用菌(0.014333) 平菇(0.011256) 高产(0.008436) 西瓜(0
3281(154)	0.215327	批发(0.137911) 市场(0.129952) 服装(0.025175) 北京(0.016002) 广州(0.014742) 价格(0.011758) 上海(0.008191) 建材(0.008096) 深圳(0.007738) 地址(0.0
510(914)	0.196001	男装(0.110479) 服饰(0.027987) 专卖店(0.020380) 劲霸(0.019943) 服装(0.017449) 价格(0.016347) 品牌(0.013548) 夹克(0.013416) 加盟(0.012974) 休闲(0
563(2401)	0.160638	工艺(0.084145) 制作(0.053216) 工艺品(0.038609) 木制(0.025492) 塑料(0.014225) 加工(0.012635) 设计(0.012179) 流程图(0.010823) 陶瓷(0.010084) 水晶
1817(752)	0.144609	胶囊(0.150875) 价格(0.032905) 效果(0.023731) 减肥(0.020117) 蜂胶(0.014502) 中药(0.011425) 北京(0.010851) 作用(0.010698) 茯苓(0.010642) 药店(0.0
1655(4535)	0.138141	批发(0.106813) 饰品(0.093192) 义乌(0.027270) 广州(0.027214) 市场(0.025927) 服饰(0.018655) 郑州(0.017089) 家居(0.014656) 汽车(0.013621) 北京(0.0

query:

莫代尔

Submit

---

tokens:

莫代尔(1.000000)

---

topics:

id(rank)	weight	topic_words
4051(159)	0.929355	内衣(0.169274) 保暖(0.024903) 情趣(0.022491) 性感(0.022092) 视频(0.020301) 模特(0.017861) 内裤(0.015672) 透明(0.014533) 美女(0.012648) 品牌(0.0
5214(4425)	0.256063	纤维(0.189105) 竹炭(0.049936) 膳食(0.017621) 价格(0.013941) 产品(0.012464) 高斯贝尔(0.009052) 辐射(0.007709) 保暖(0.006769) 作用(0.006447) 纺织
5970(2592)	0.202387	涤纶(0.051665) 价格(0.026307) 面料(0.024020) 尼龙(0.019313) 化纤(0.017226) 锦纶(0.015207) 纤维(0.014484) 纱线(0.013749) 长丝(0.013227) 粘胶(0.0
1109(35)	0.132384	女装(0.066362) 品牌(0.032049) 淘宝网(0.027243) 服饰(0.019601) 服装(0.015364) 新款(0.015183) 外套(0.014774) t恤(0.011847) 春装(0.011017) 图片(0.
3595(2806)	0.070228	面料(0.102435) 针织(0.049609) 服装(0.037397) 印花(0.022289) 招聘(0.015971) 市场(0.008687) 工艺(0.008146) 广州(0.007693) 公司(0.007181) 真丝(0.0
7748(571)	0.053765	内裤(0.078704) 衣服(0.043388) 女人(0.041948) 美女(0.037248) 视频(0.025772) 胸罩(0.025339) 图片(0.020984) 少女(0.020088) 脱掉(0.016180) 游戏(0.0
8721(56)	0.037367	搭配(0.063732) 大衣(0.036574) 颜色(0.028024) 女装(0.019412) 流行(0.017402) 黑色(0.017399) 图片(0.016212) 服装(0.015273) 外套(0.014755) 围巾(0.0
4733(97)	0.037340	手工(0.069565) 毛衣(0.067347) 编织(0.059632) 制作(0.050857) 视频(0.020713) 毛线(0.020550) 图解(0.018679) 儿童(0.016230) 帽子(0.016213) 图片(0.0

Peacock Team

Peacock: Learning Long-Tail Topic Features for Industrial Applications

ACM Transactions on Intelligent Systems and Technology, 2014

Yi Wang

Xuemin Zhao

Zhenlong Sun

Hao Yan

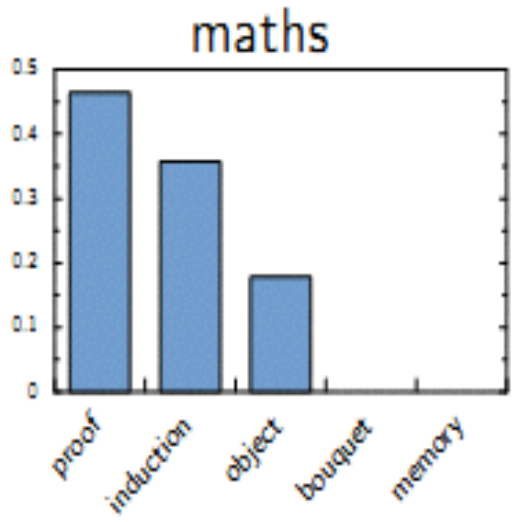
Lifeng Wang

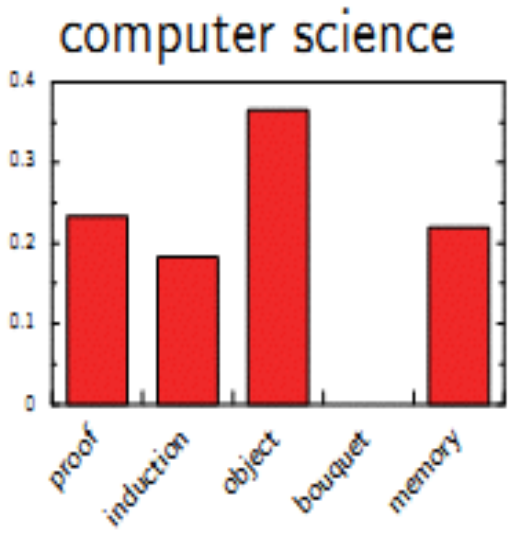
Zhihui Jin

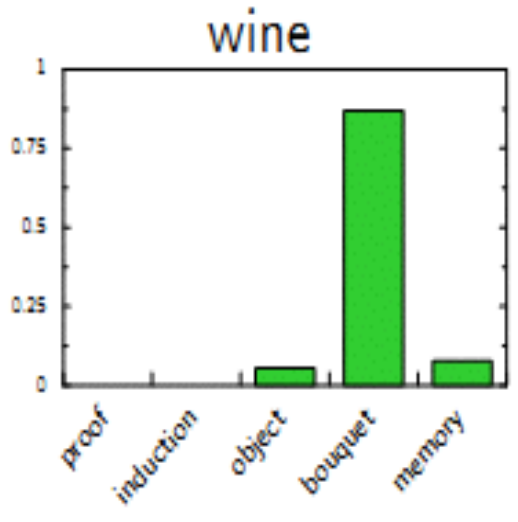
Liubin Wang

Doc-Topic Structure

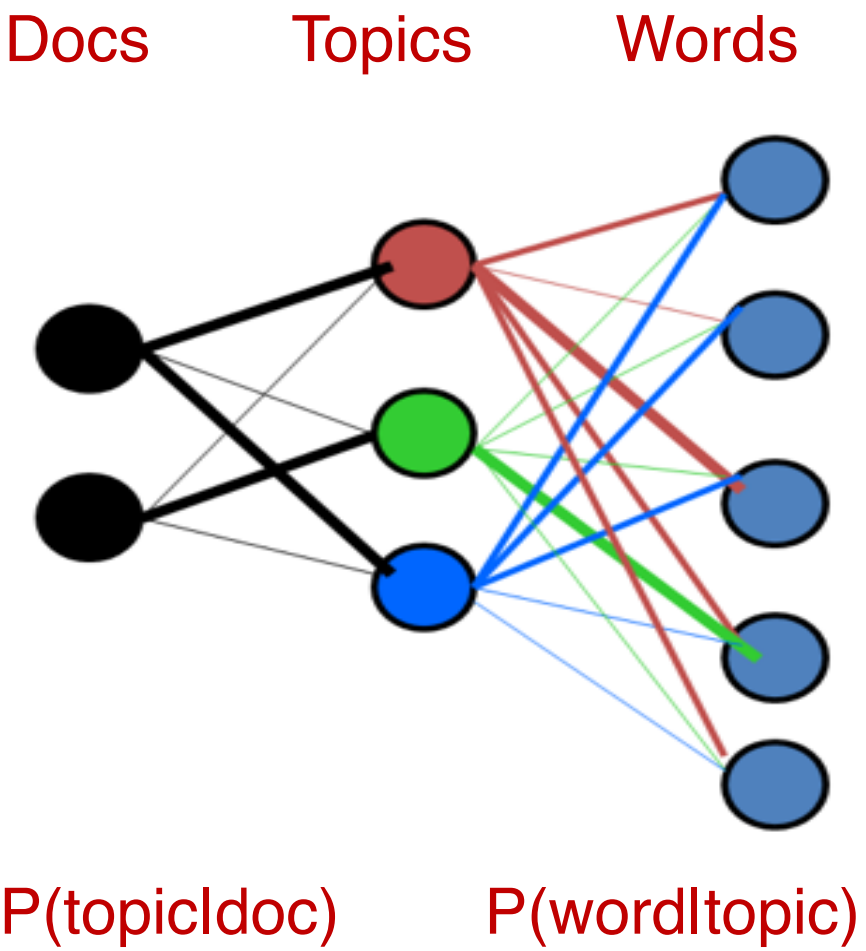
- Doc 是由 topic 组成的
- Topic 是 Vocab 上的概率分布 [Hofmann, 1999]



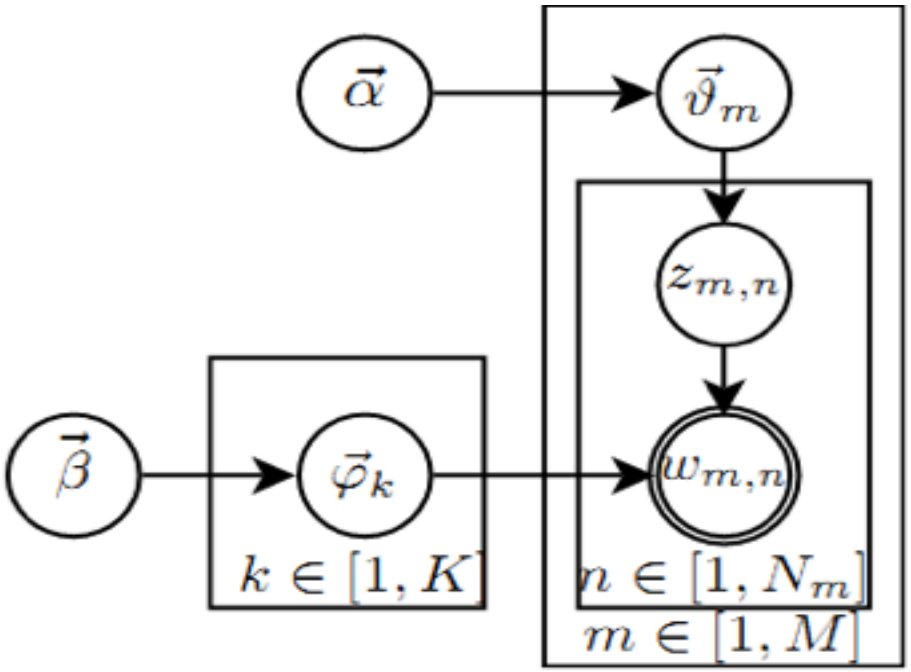


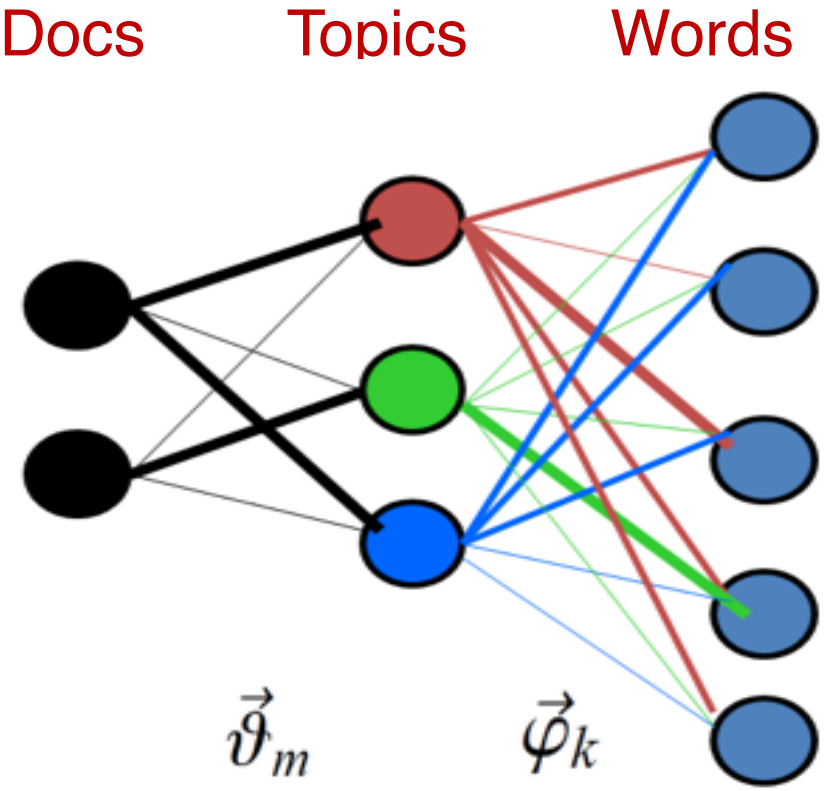


PLSA Topic Modeling



LDA Topic Modeling



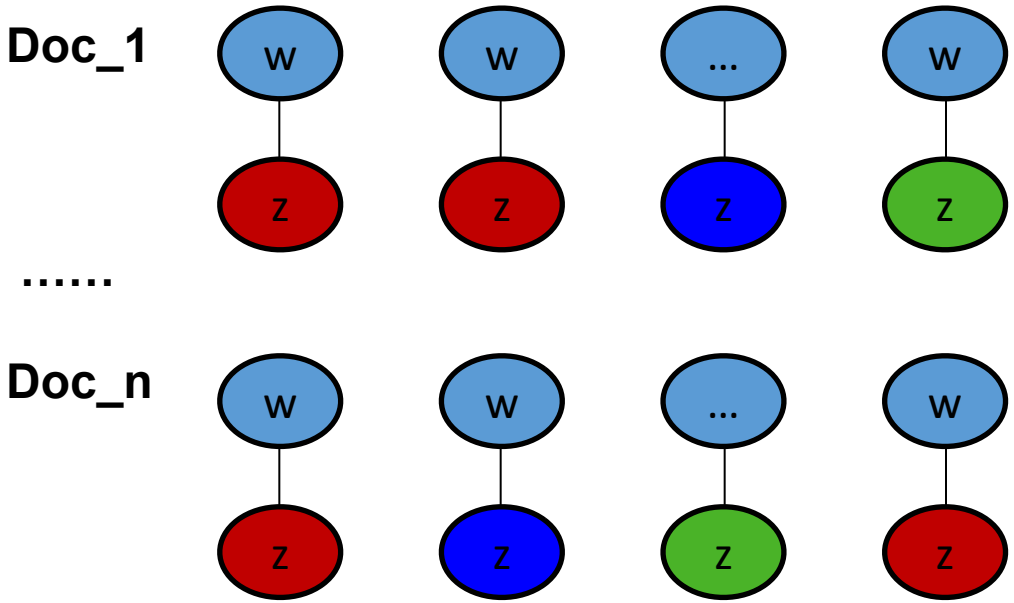


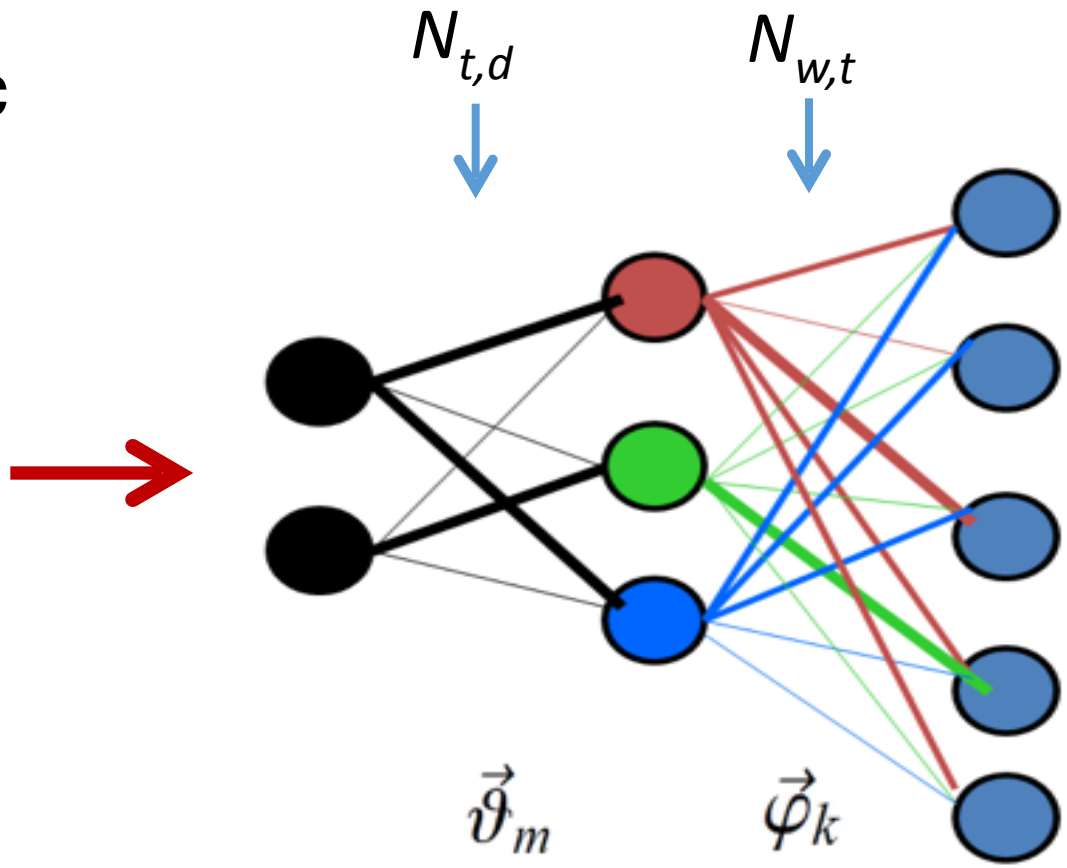
P(topicldoc)

P(wordltopic)

LDA Model Training

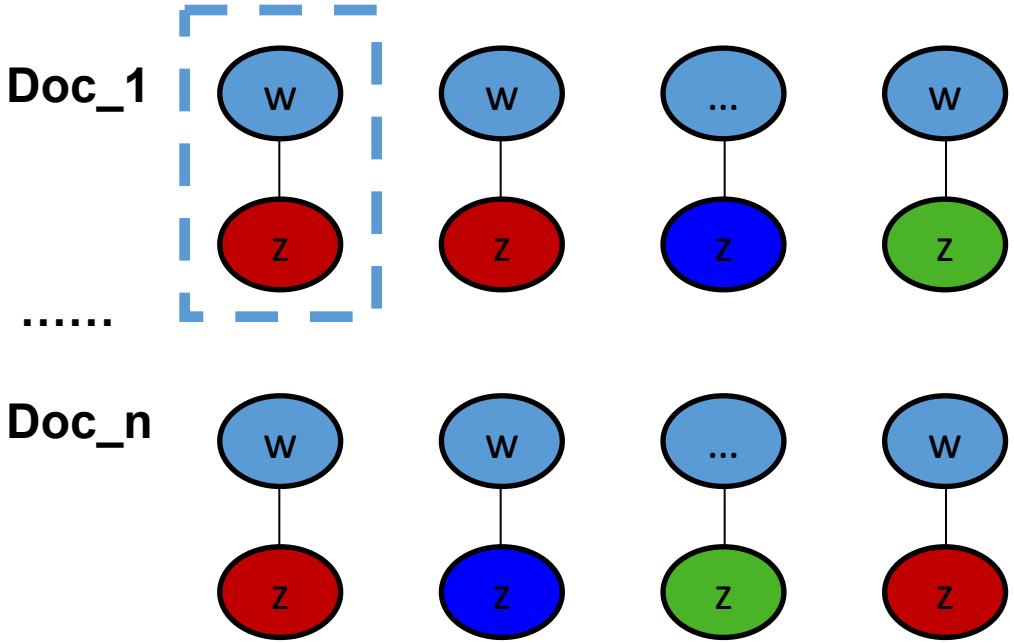
Step1: 随机初始化每个词的 topic

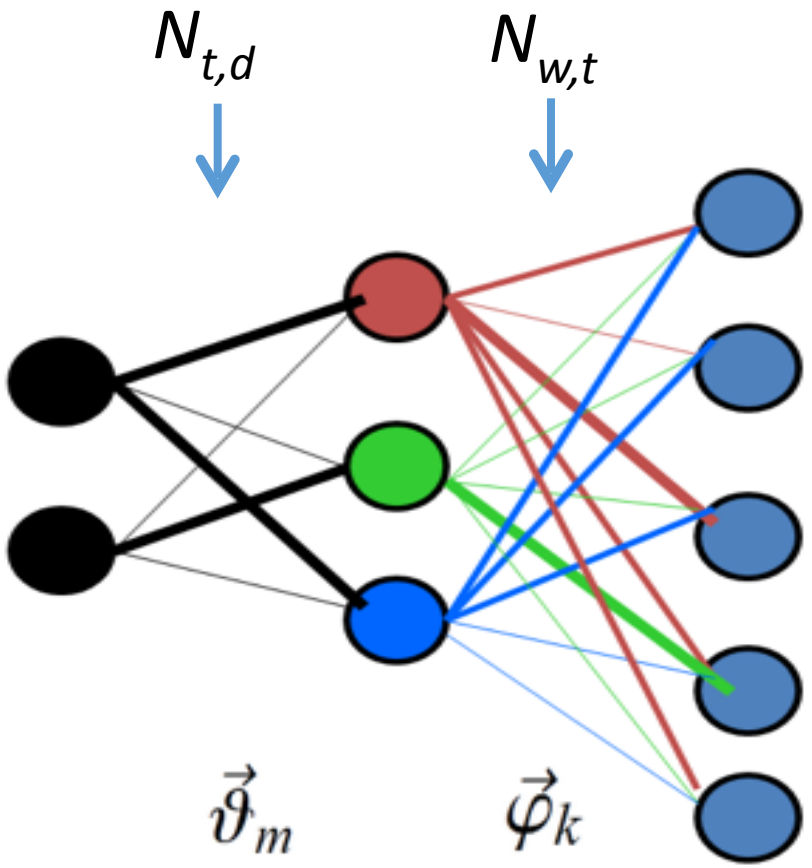




LDA Model Training

Step2: 重新采样每个 topic, 更新计数



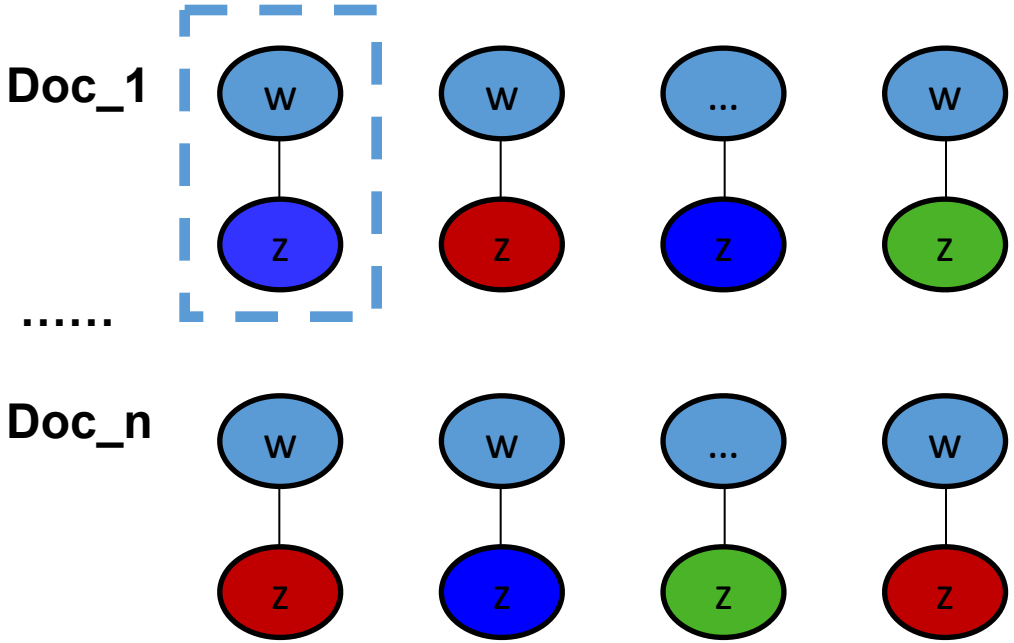


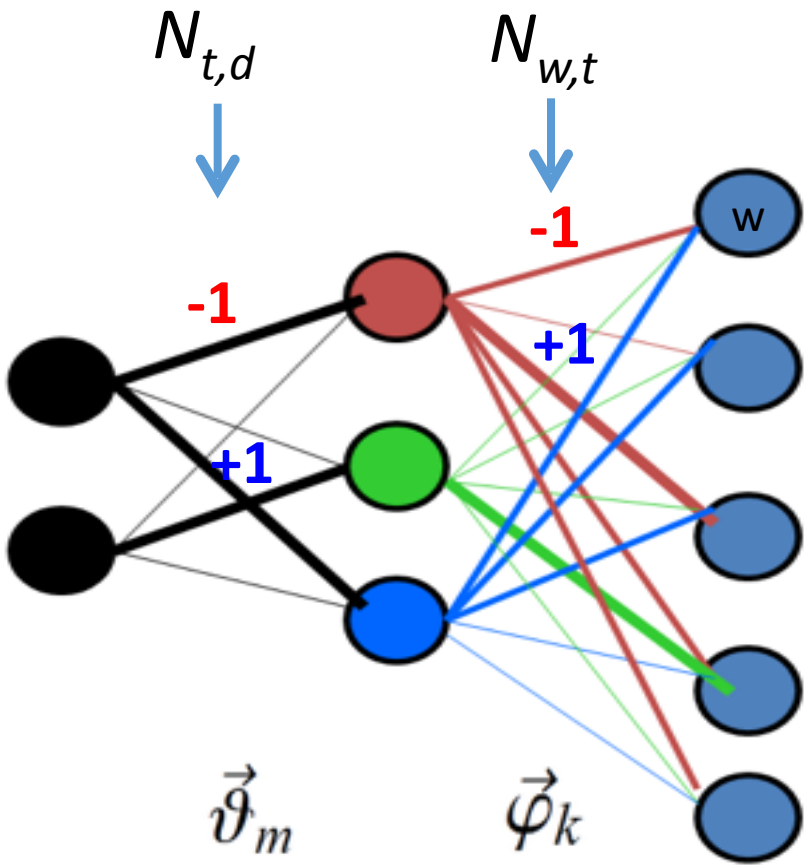
$$p(z_i=k|\vec{z}_{\neg i},\vec{w}) \propto \frac{n^{(t)}_{m,\neg i}+\alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m}+\alpha_k]-1} \cdot \frac{n^{(t)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i}+\beta_t}$$

P(topicldoc) P(wordltopic)

LDA Model Training

Step3: 重新采样每个 topic, 更新计数



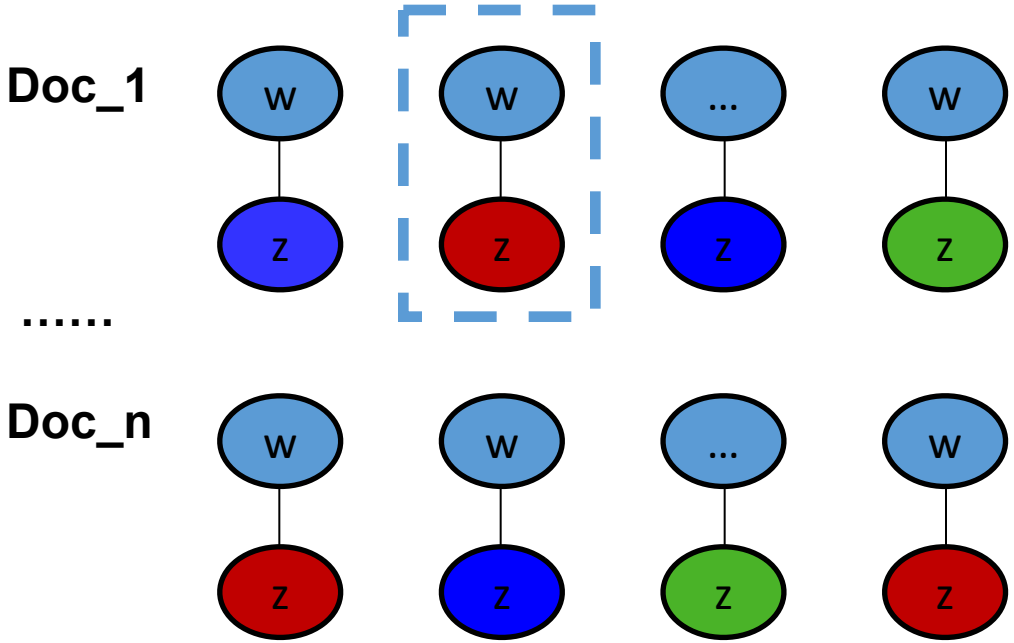


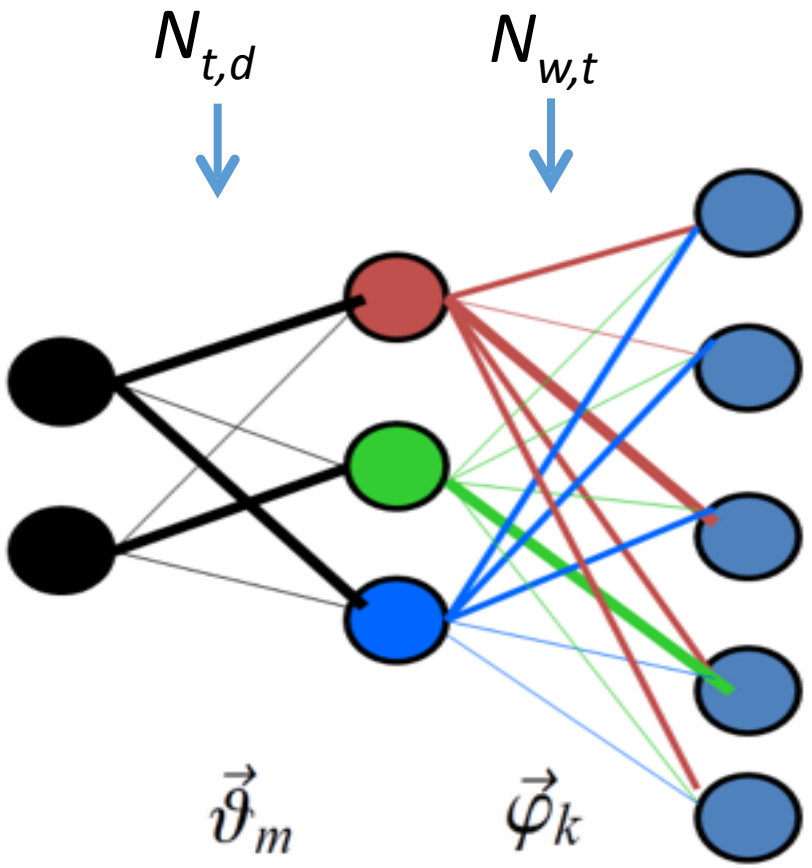
$$p(z_i=k|\vec{z}_{\neg i},\vec{w}) \propto \frac{n^{(t)}_{m,\neg i}+\alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m}+\alpha_k]-1} \cdot \frac{n^{(t)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i}+\beta_t}$$

P(topicldoc) P(wordltopic)

LDA Model Training

Step4: 重复 step2&3, 直到模型收敛

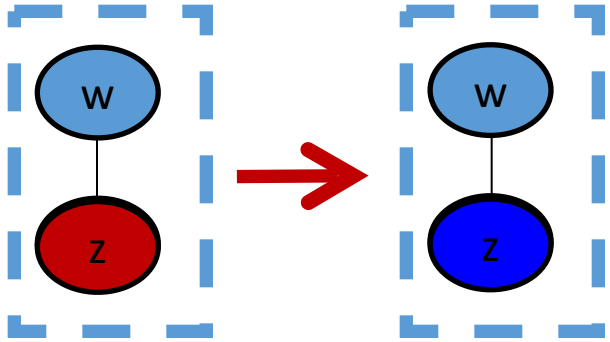




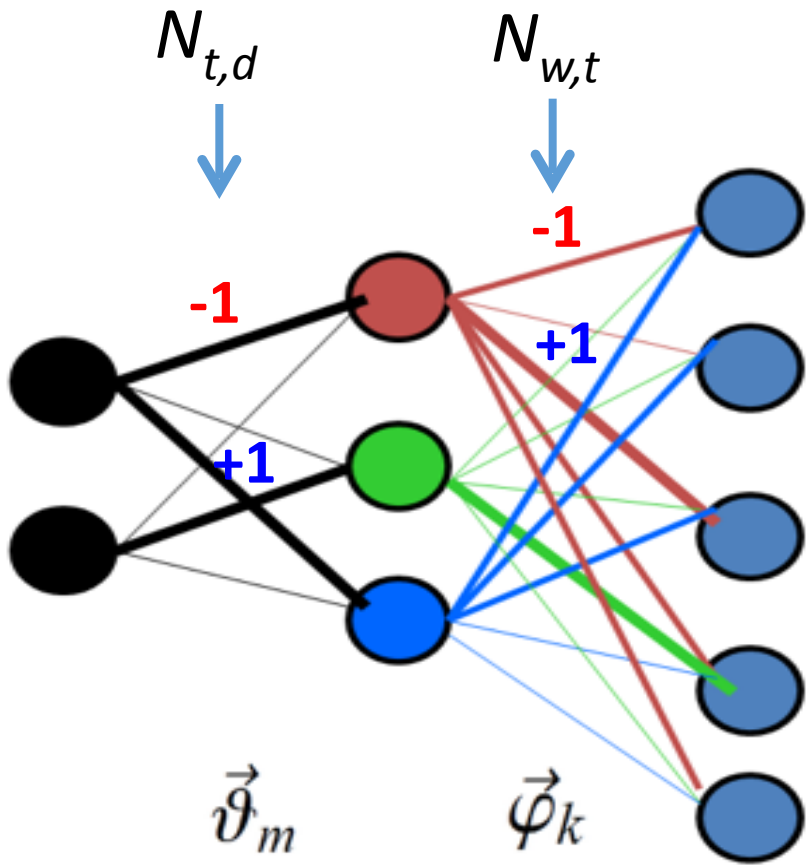
$$p(z_i{=}k|\vec{z}_{\neg i},\vec{w}) \propto \frac{n^{(t)}_{m,\neg i}+\alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m}+\alpha_k]-1} \cdot \frac{n^{(t)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i}+\beta_t}$$

P(topicldoc)　　P(wordltopic)

Large-scale LDA Modeling

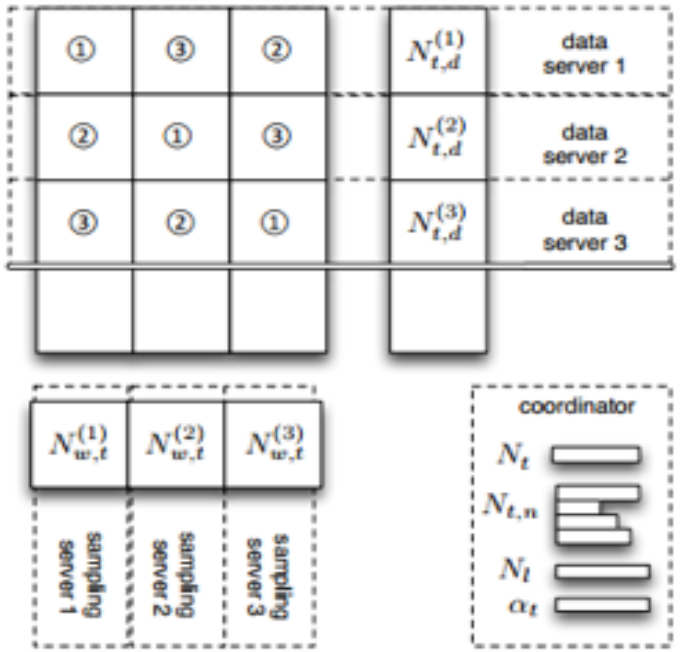


- Q1: 如何提升 Gibbs Sampling 速度
  - 标准采样算法太慢
- Q2: 如何支持大数据、大模型
  - 十亿文档，百万词汇，百万 topic
- Q3: 如何调参优化模型质量
  - alpha，beta 如何选取
  - topic 个数如何考虑



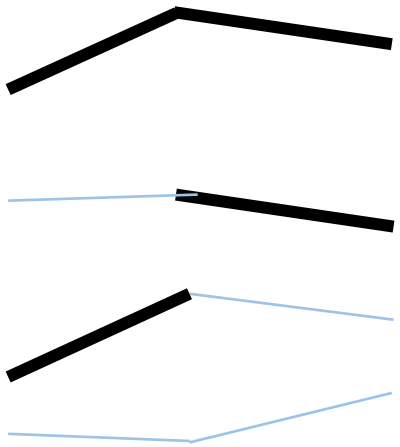
Peacock: Large-scale Topic Modeling

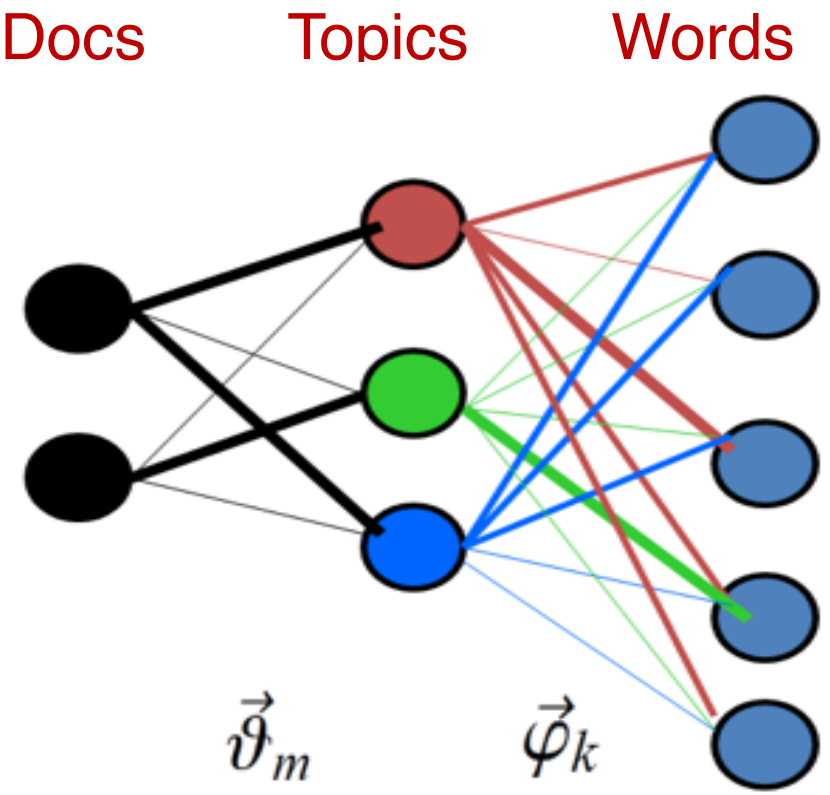
- Q1: 如何提升 Gibbs Sampling 速度
  - 使用 SparseLDA 算法做 Gibbs Sampling
  - 比标准 **LDA** 快**30**倍
- Q2: 如何支持大数据、大模型
  - 基于 Go 语言实现
  - 矩阵分块并行计算
  - 可以支持**10**亿 **x 1**亿的矩阵分解
  - 可以支持**100**万 **topics** 计算
  - 类似 Google Rephil 系统，挖掘长尾语义
- Q3: 如何调参优化模型质量
  - 每轮迭代对超参数做优化，**智能训练 topics** 个数



Q1: 采样速度

- 标准 LDA 采样
  - 计算所有路径的累积概率
  - 计算速度慢
- 概率路径是 sparse 的



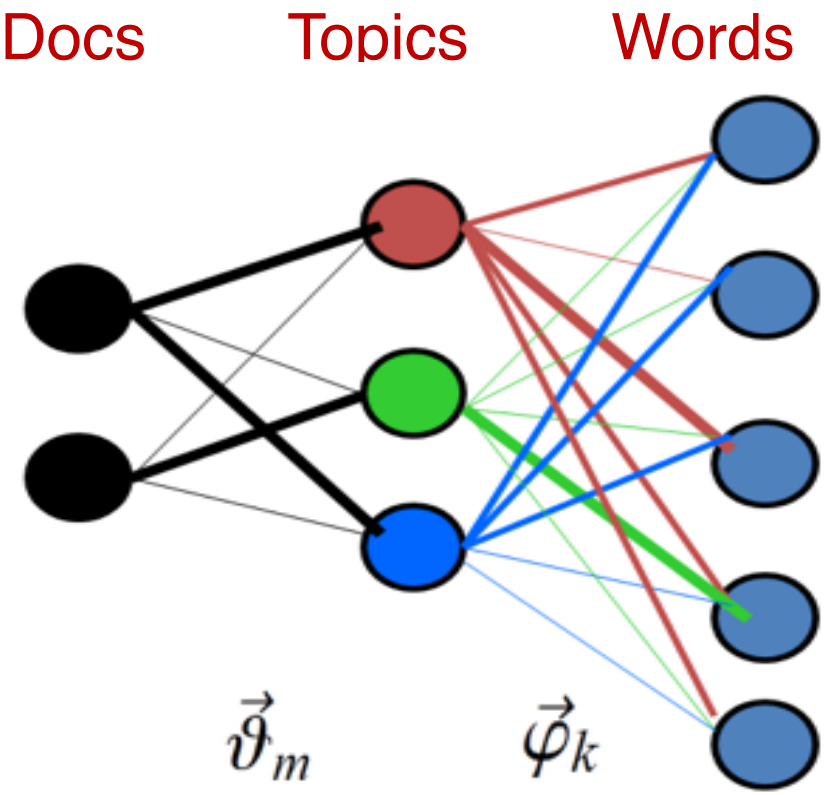


$$p(z_i{=}k|\vec{z}_{\neg i},\vec{w}) \propto \frac{n^{(t)}_{m,\neg i}+\alpha_k}{[\sum_{k=1}^{K} n^{(k)}_m+\alpha_k]-1} \cdot \frac{n^{(t)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i}+\beta_t}$$

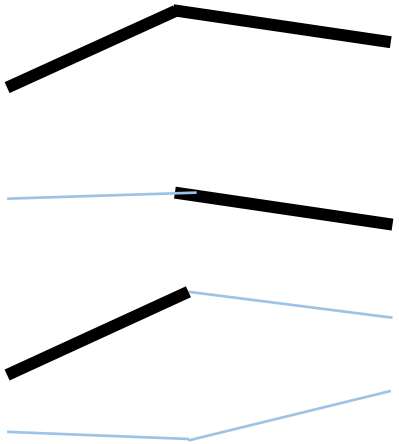
P(topicIdoc)　　P(wordItopic)

SparseLDA

- 按照路径类型计算概率分布
- 先按路径类型概率分布采样
- 在类型内部采样路径

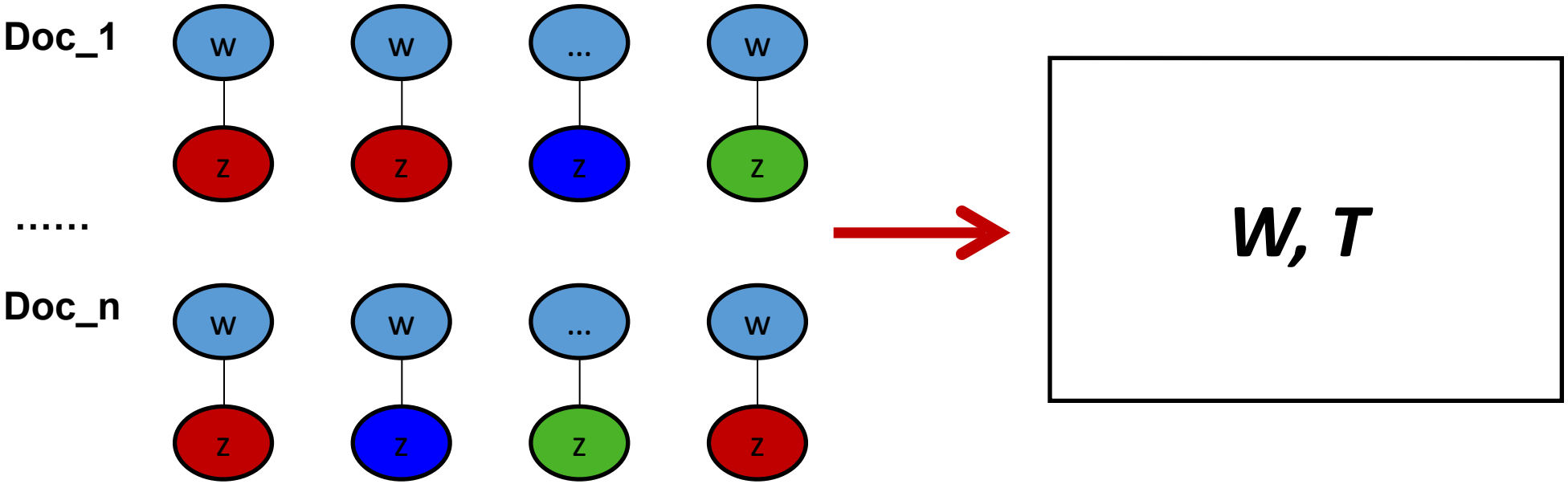


Limin Yao, David Mimno, and Andrew McCallum. ***Efficient Methods for Topic Model Inference on Streaming Document Collections***. KDD 2009.

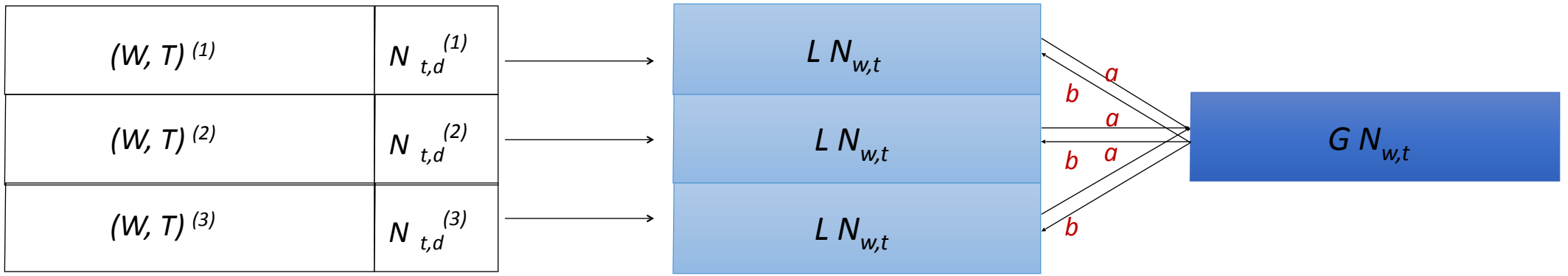


Path-Num	Probability
10	*0.8*
20	*0.1*
70	*0.09*
9900	*0.01*

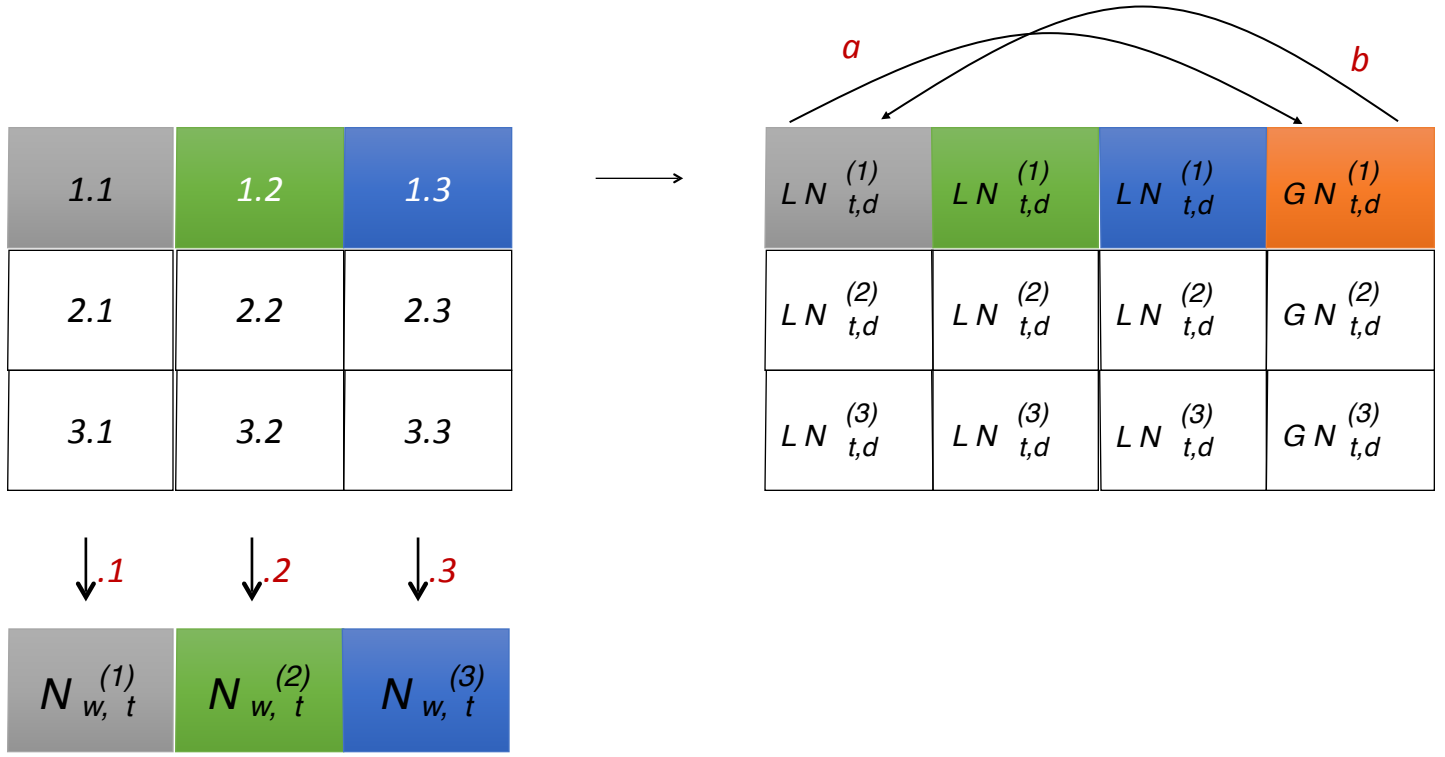
Q2: 十亿篇文档，百万词汇，百万 Topics



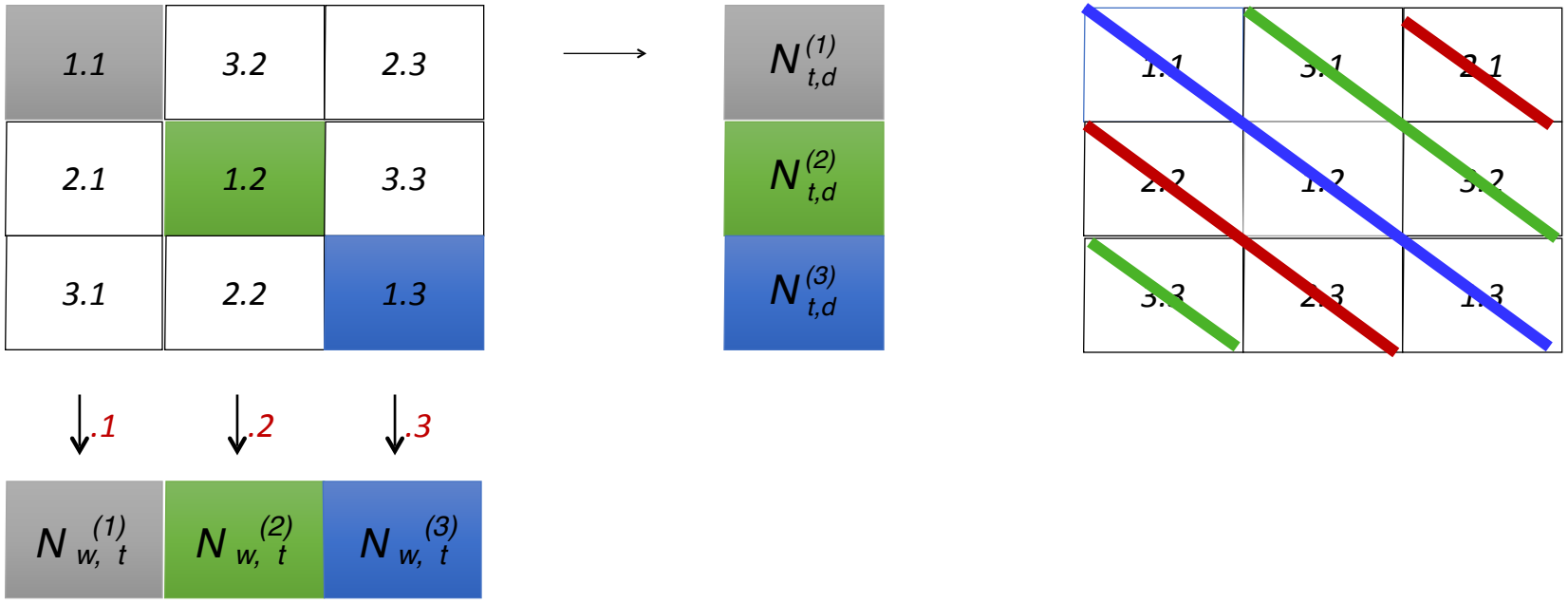
AD-LDA (Data Parallelism)



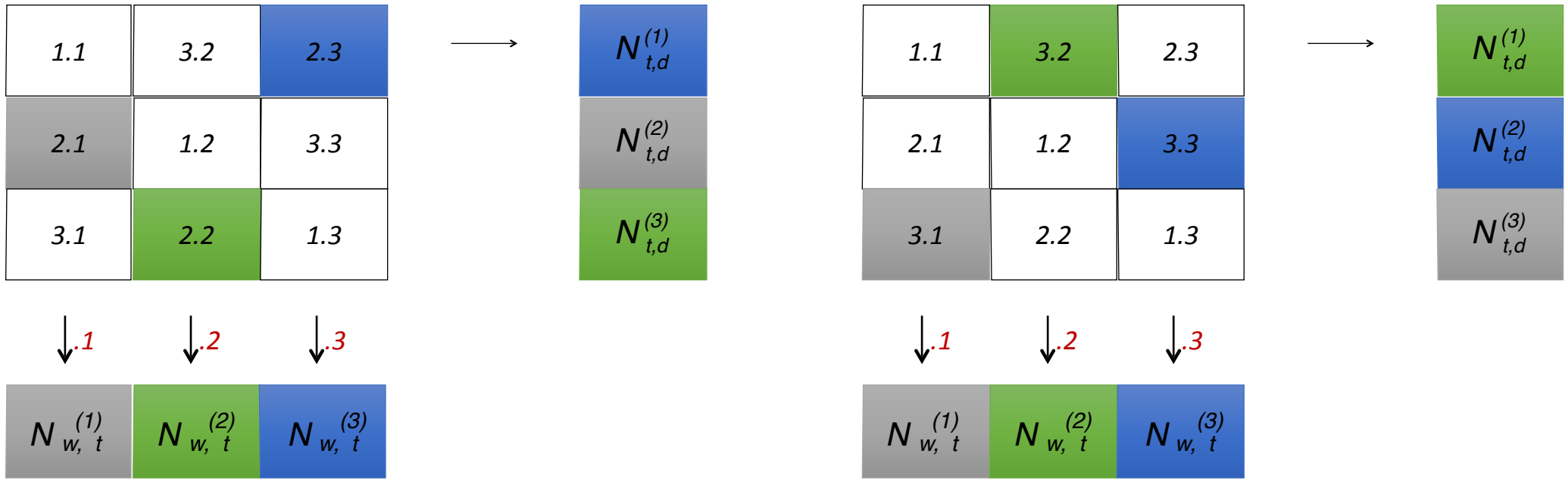
Model Parallelism



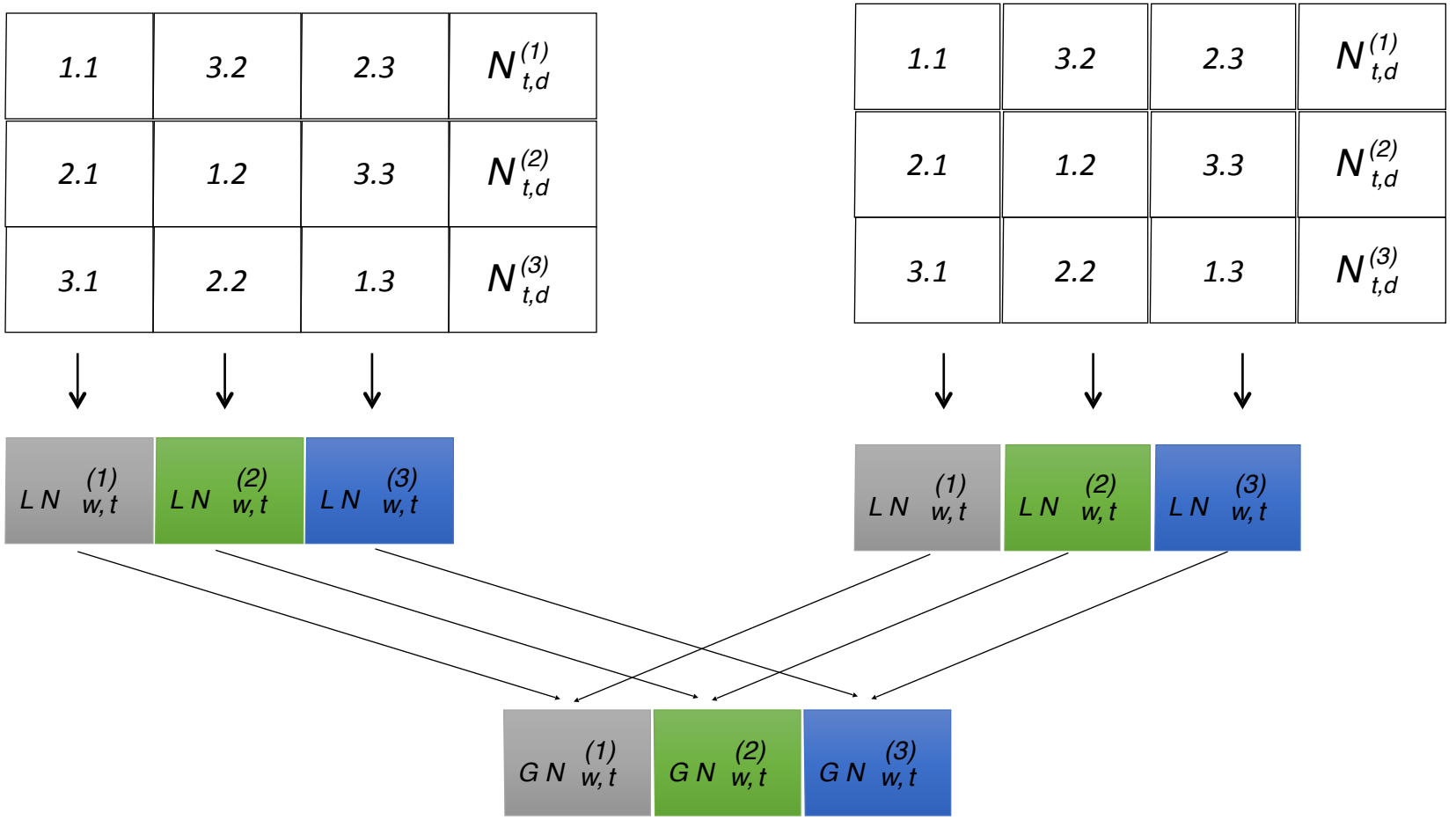
Lock-free Synchronization



Lock-free Synchronization



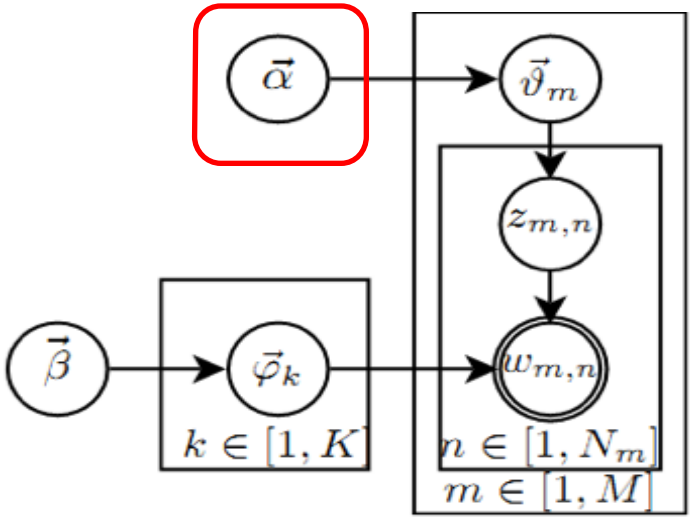
Model Parallelism + Data Parallelism

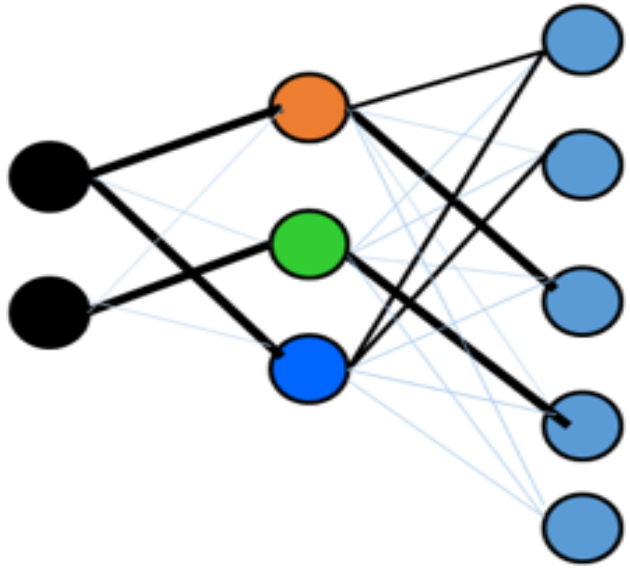


Q3: 优化模型质量

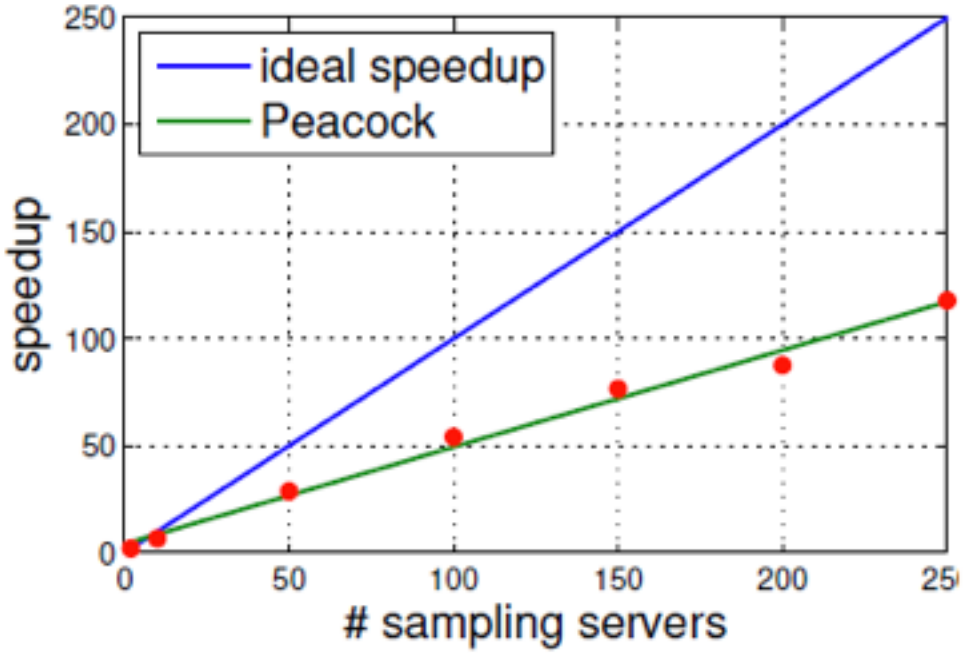
Hanna M. Wallach, David Mimno, and Andrew McCallum.
Rethinking LDA: Why Priors Matter. NIPS 2009.

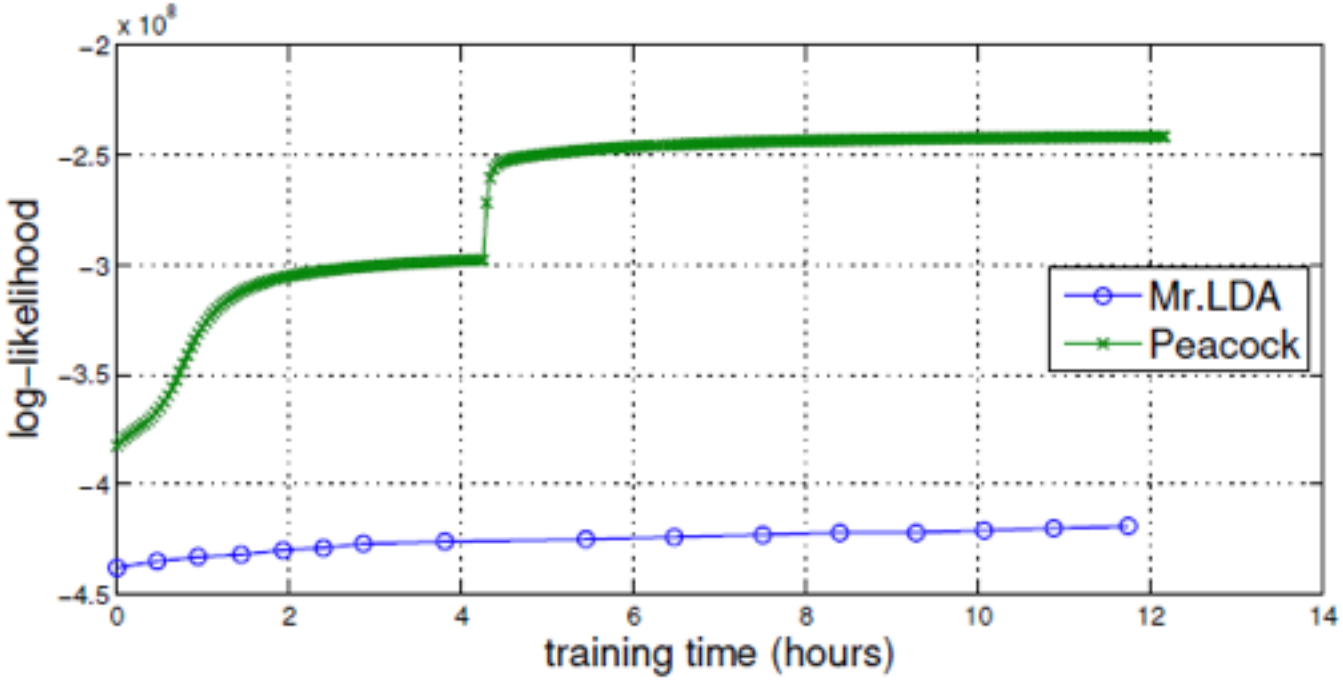
- 超参数 alpha 对模型质量有重要的影响
- 每轮迭代中，通过 MLE 估计优化 alpha



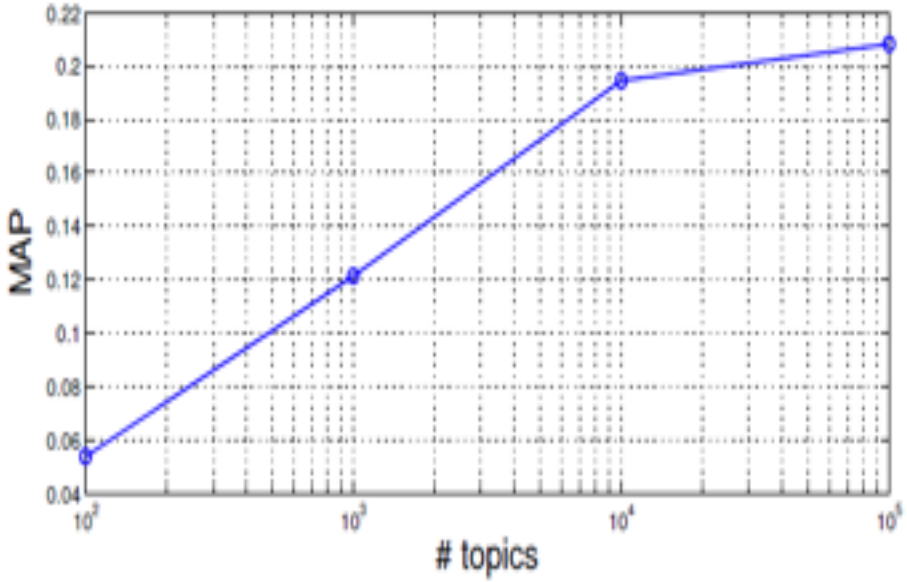


Peacock 性能

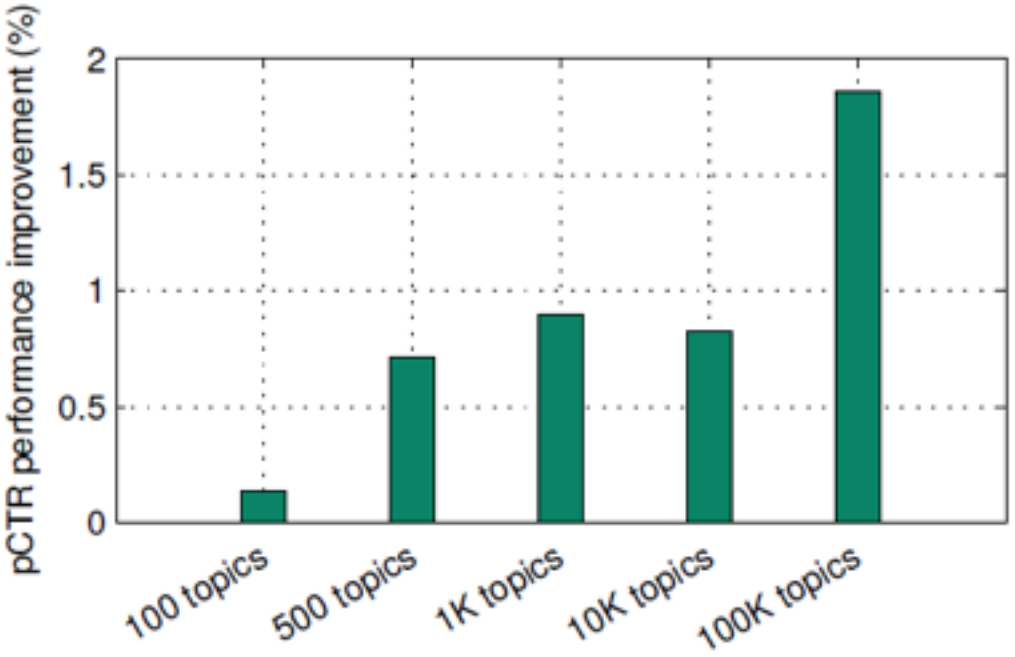

ideal speedup
Peacock
speedup
sampling servers
0
50
100
150
200
250


x 10^8
log−likelihood
training time (hours)
Mr.LDA
Peacock
−2
−2.5
−3
−3.5
−4
−4.5
0
2
4
6
8
10
12
14

为什么我们需要大模型



搜索相关性MAP

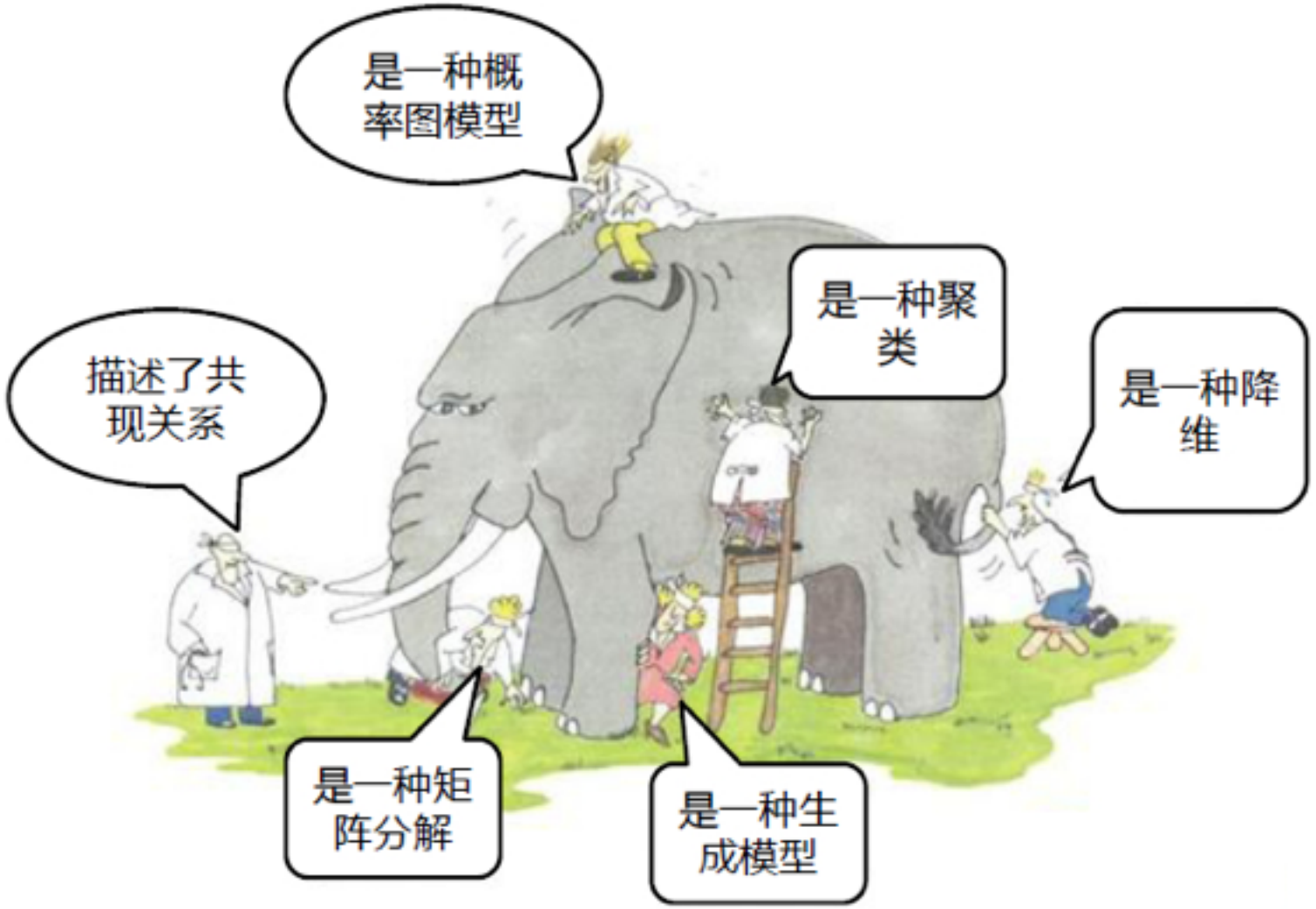


广告点击率模型 AUC

Peacock 学习长尾的 Topic

狗 dog	生 birth	孩子 child	小狗 puppy	病 ill	虫 parasites	怀孕 preganent	猫 cat
污水 polluted water	池 pool	厂 plant	石油 oil	页岩 shale	炼 refine		
血 blood	功能 function	甲状腺 thyroid	检查 test				
空间 space	图片 images	qq	制作 make	头像 head portait			
美 beautiful	雅 elegant	仕 gentle	上海 Shanghai	服饰 clothing			
店 restaurant	小吃 snack	永和	夜市 night market	好吃 delicious	豆浆 bean milk		
鲁鲁修	叛逆 rebellious	反叛 rebel	鲁路修	动画 animation			
小路 trail	飘 float	歌曲 song	云 cloud	歌词 lyrics			
草	社区 forum	榴	最新 most recent	地址 address	下载 download	黄色 porn	视频 video

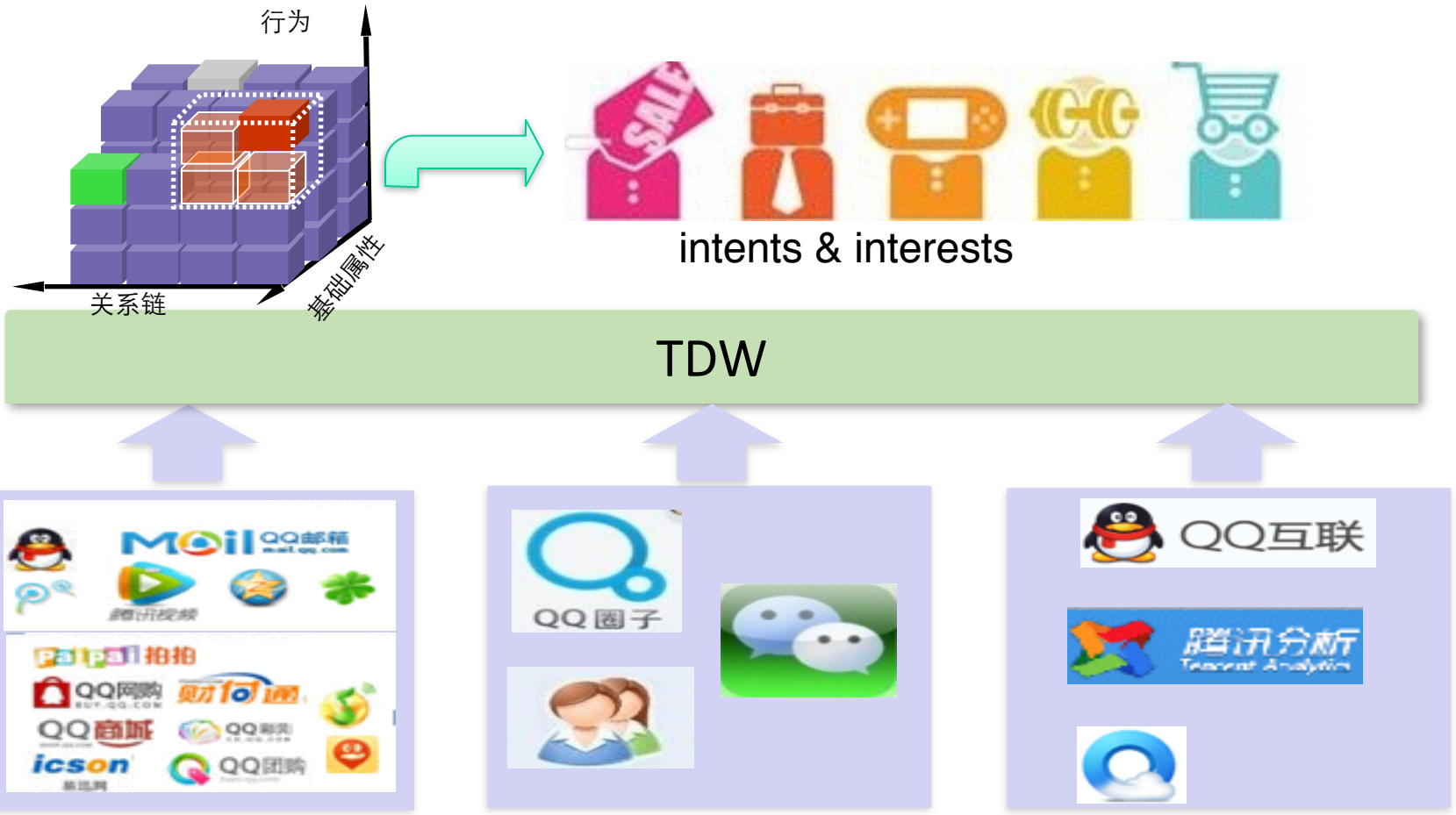
LDA Topic Modeling



Peacock 在腾讯业务中的应用

- 文本语义分析
- 广告相关性计算与 CTR 预估
- QQ 群分类与广告定向
- QQ 群推荐

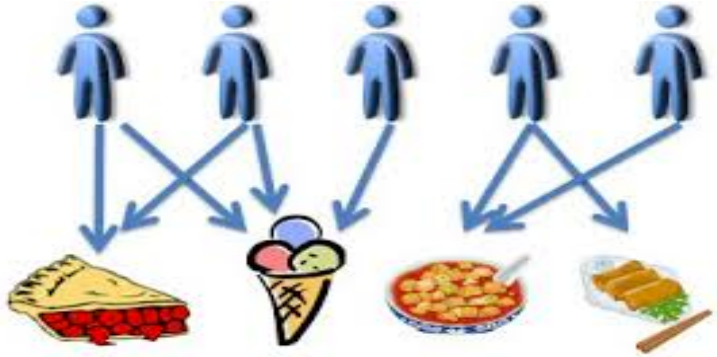
广点通用户数据挖掘与广告精准定向

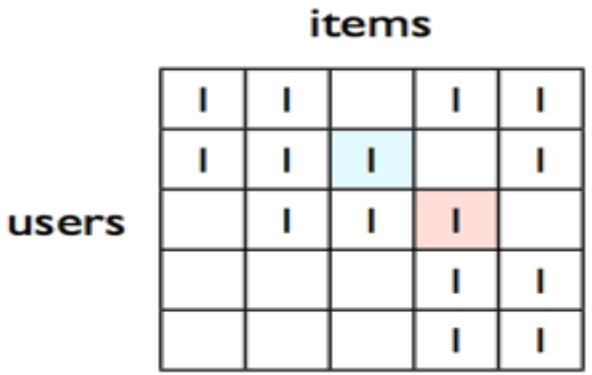


用户行为数据分析

- 文本语义分析

- RecSys: user-item 矩阵分解





Peacock 应用：文本语义分析

- 解决方案
  - 字面抽取：命名实体识别、关键词
    - 信息量小，有歧义，容易陷入 Vocabulary Gap
  - 语义分析：文本聚类（Topic），文本分类
    - 从海量文本数据中归纳“知识”，帮助理解语义
- 难点
  - 如何挖掘细粒度、长尾语义?

Peacock应用： 文本语义分析

红酒木瓜汤

0.397 [丰胸(0.1642) 产品(0.0776) 减肥(0.0645) 木瓜(0.0464)]
0.182 [饭后(0.1251) 饭前(0.0757) 服用(0.026) 减肥(0.022)]
0.162 [功效(0.0435) 山药(0.039) 作用(0.0379) 做法(0.0264)]
0.095 [糖尿病(0.0811) 血糖(0.0336) 高血压(0.0285)]
0.050 [蜂蜜(0.0801) 牛奶(0.0427) 面膜(0.0303) 好处(0.025)]
0.044 [做法(0.0598) 萝卜(0.0569) 排骨(0.0213) 牛肉(0.017)]

苹果

0.170 [苹果(0.23) 手机(0.124) iphone(0.025) 电脑(0.017)]
0.086 [范冰冰(0.114) 苹果(0.085) 电影(0.059) 佟大为(0.0315)]
0.058 [iphone(0.166) 手机(0.07) 3gs(0.039) 苹果(0.033)]
0.025 [苹果(0.078) 重量(0.027) 水果(0.015) 质量(0.013)]
0.014 [手机(0.183) 步步高(0.083) 电池(0.043)]
0.009 [windows(0.089) xp(0.088) 系统(0.05)]

苹果电影

0.588 [范冰冰(0.114) 苹果(0.085) 电影(0.059) 佟大为(0.0315)]
0.095 [电影(0.096) 在线(0.087) 观看(0.07) 视频(0.039)]
0.043 [苹果(0.23) 手机(0.124) iphone(0.025) 电脑(0.017)]
0.043 [ipod(0.156) touch(0.11) pro(0.03) itunes(0.02)]
0.020 [电脑(0.145) 关机(0.069) 自动(0.06) 开机(0.05)]

TextMiner 语义分析平台

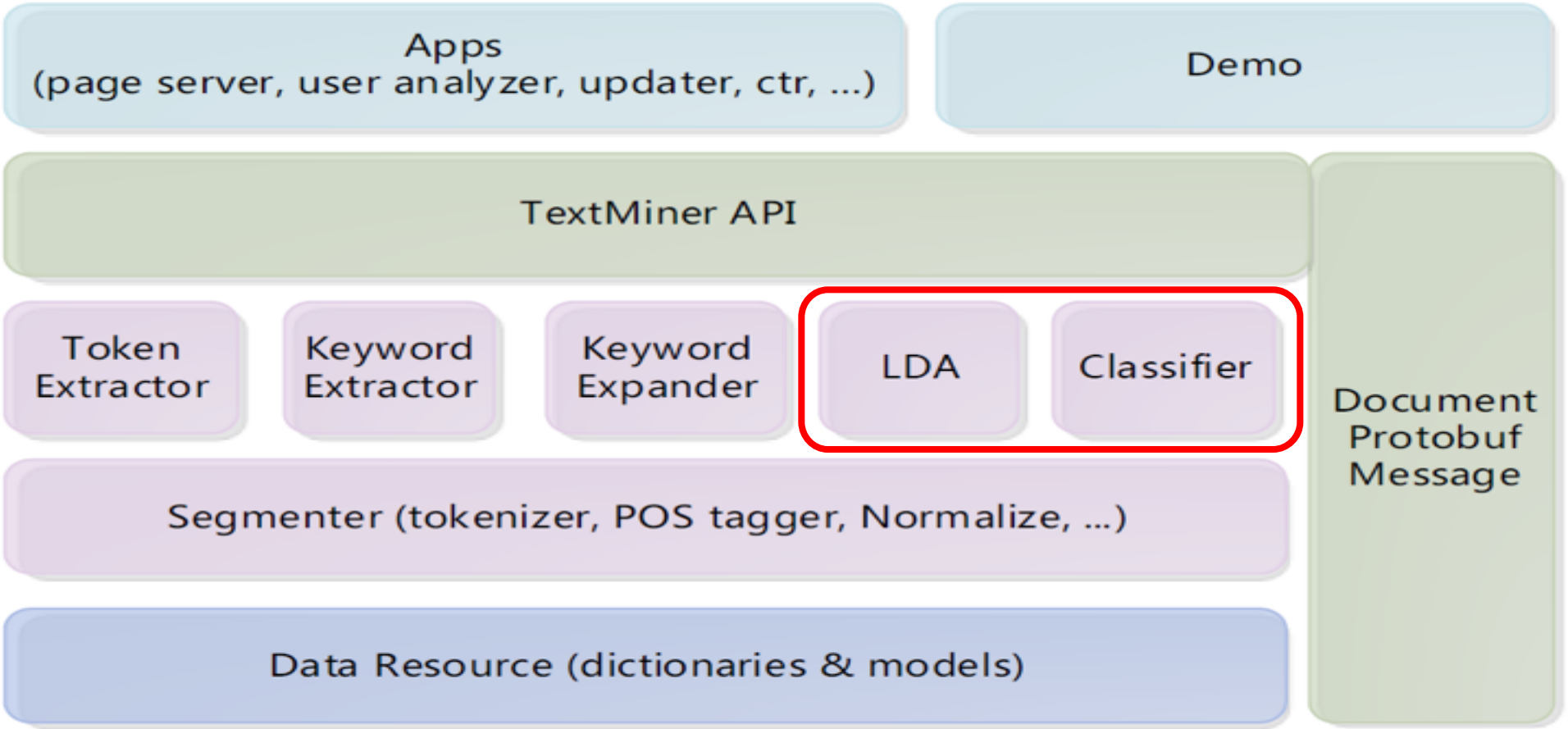


图 1 TextMiner 系统架构图

Peacock 应用： 情境广告相关性与CTR 优化
关键词定向广告

- **Peacock Model**

  10 亿 query log，20w words，10w topics，160 台机器，一周训练
- **相关性优化**

  3万 (query, doc) 相关性标注样本, LearningToRank Model

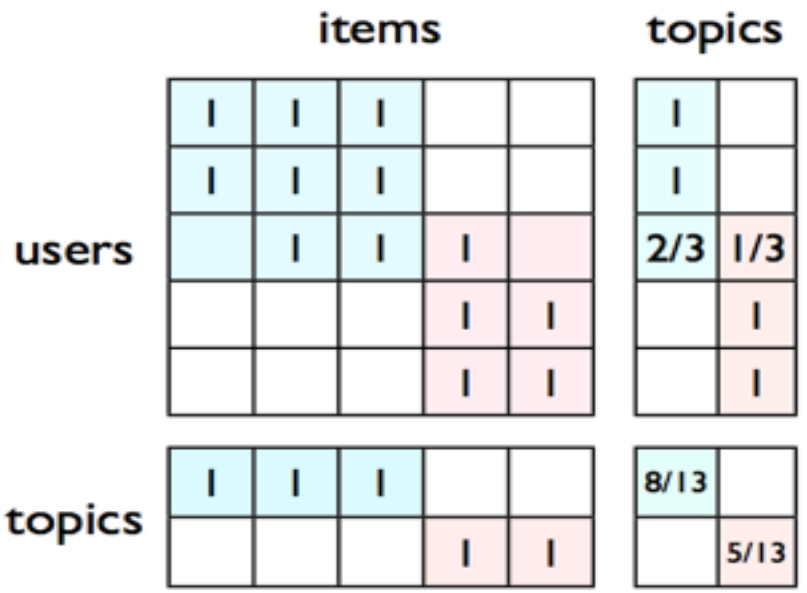
  NDCG@5提升 **8.92%**
- **CTR 优化**

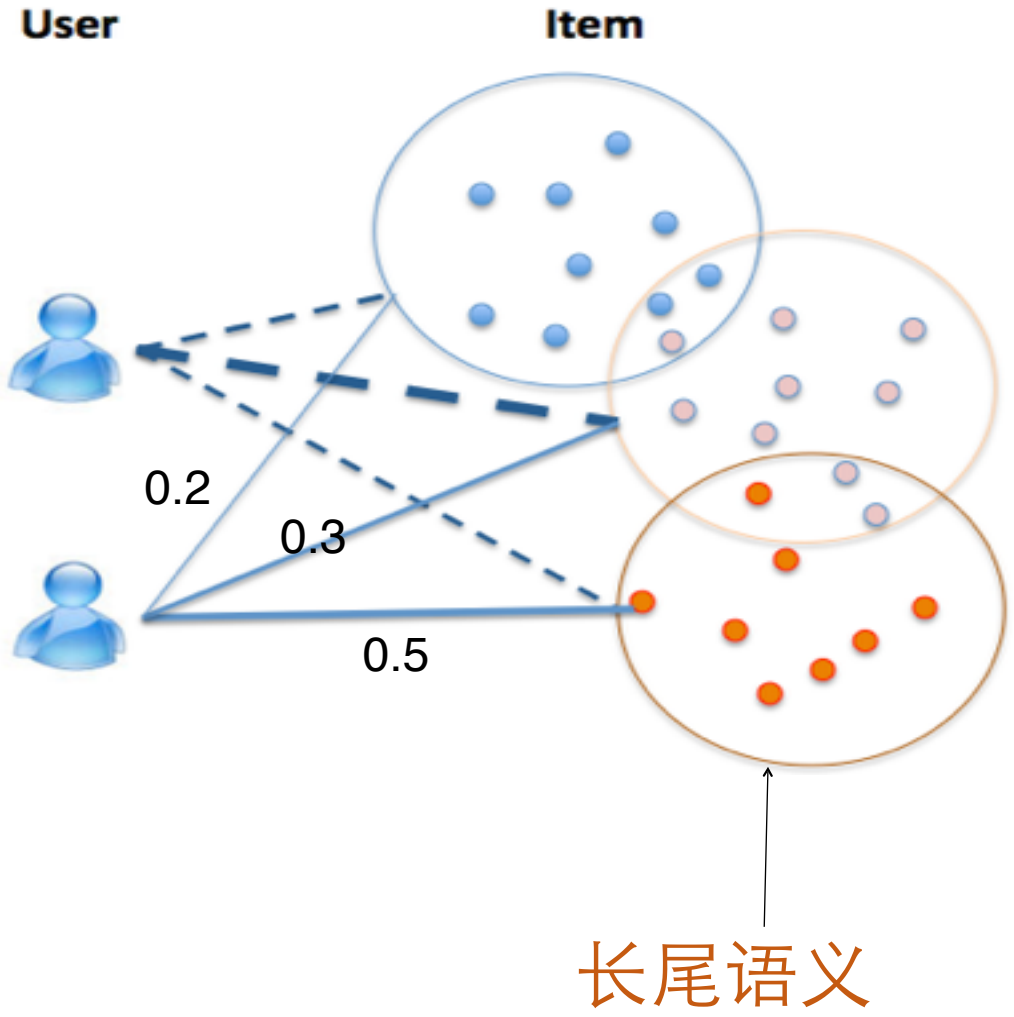
  2亿 pv / 天

  离线 pCTR AUC 提升 **1.8%**

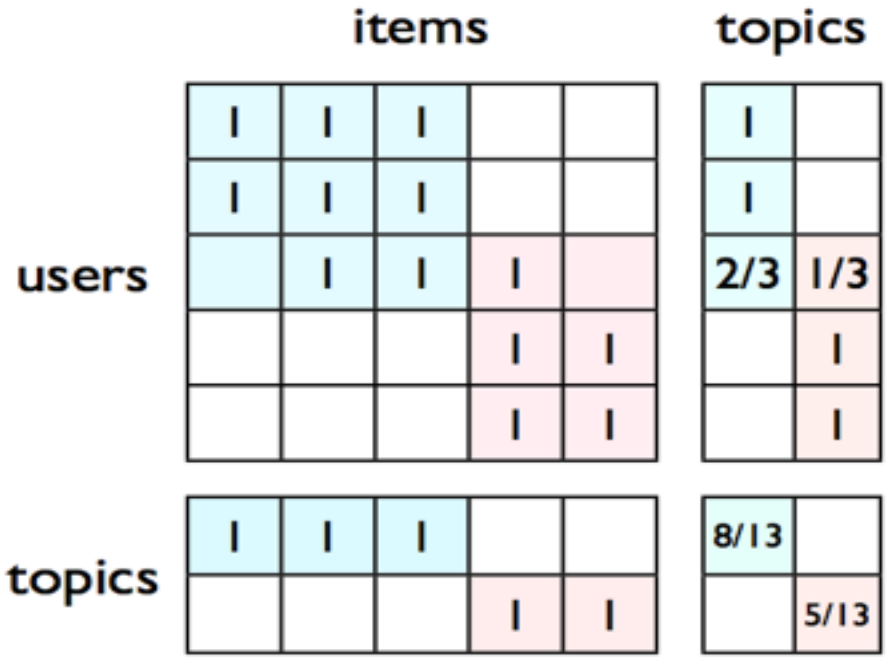
  在线实验 AdCTR 提升 **8.82%**

Recsys: user-item 语义挖掘



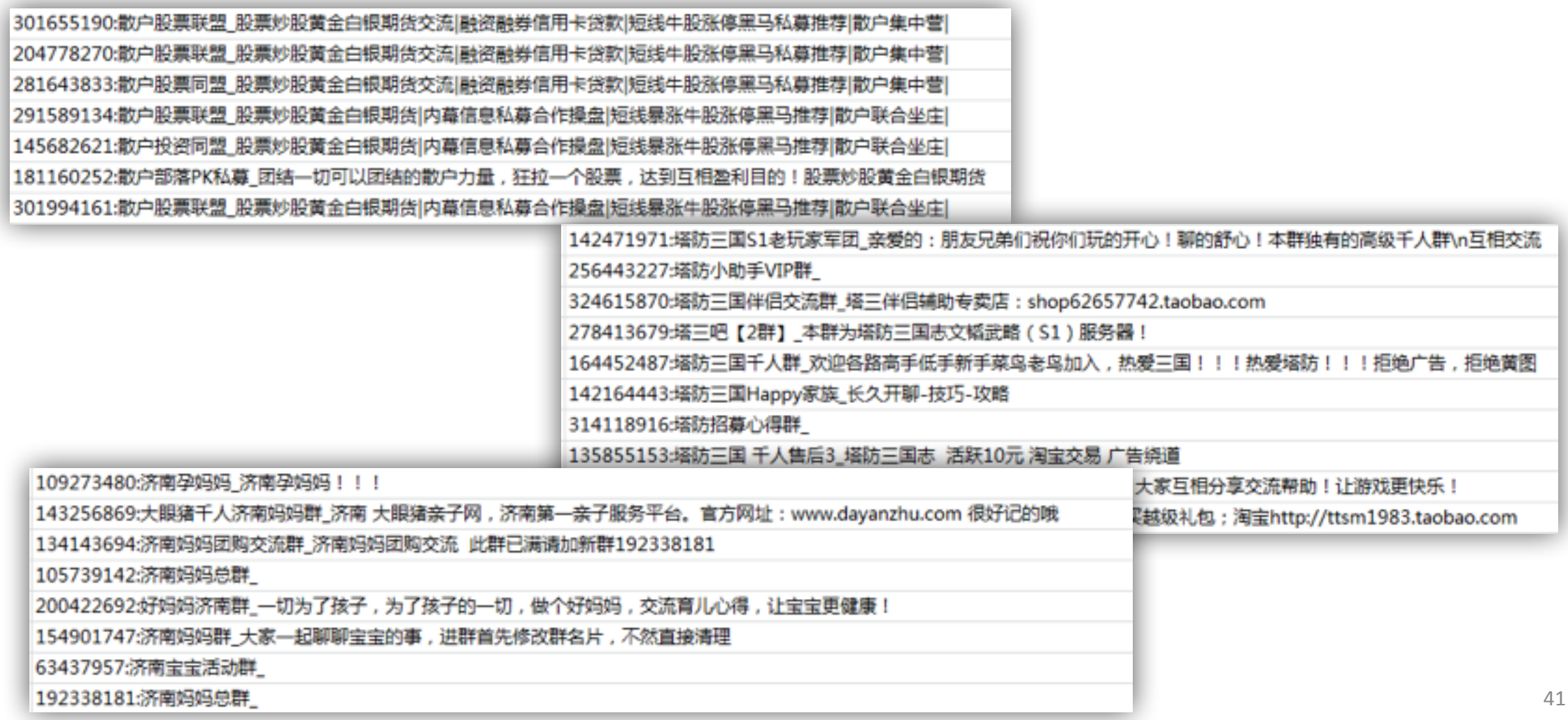


Peacock: 大规模矩阵分解

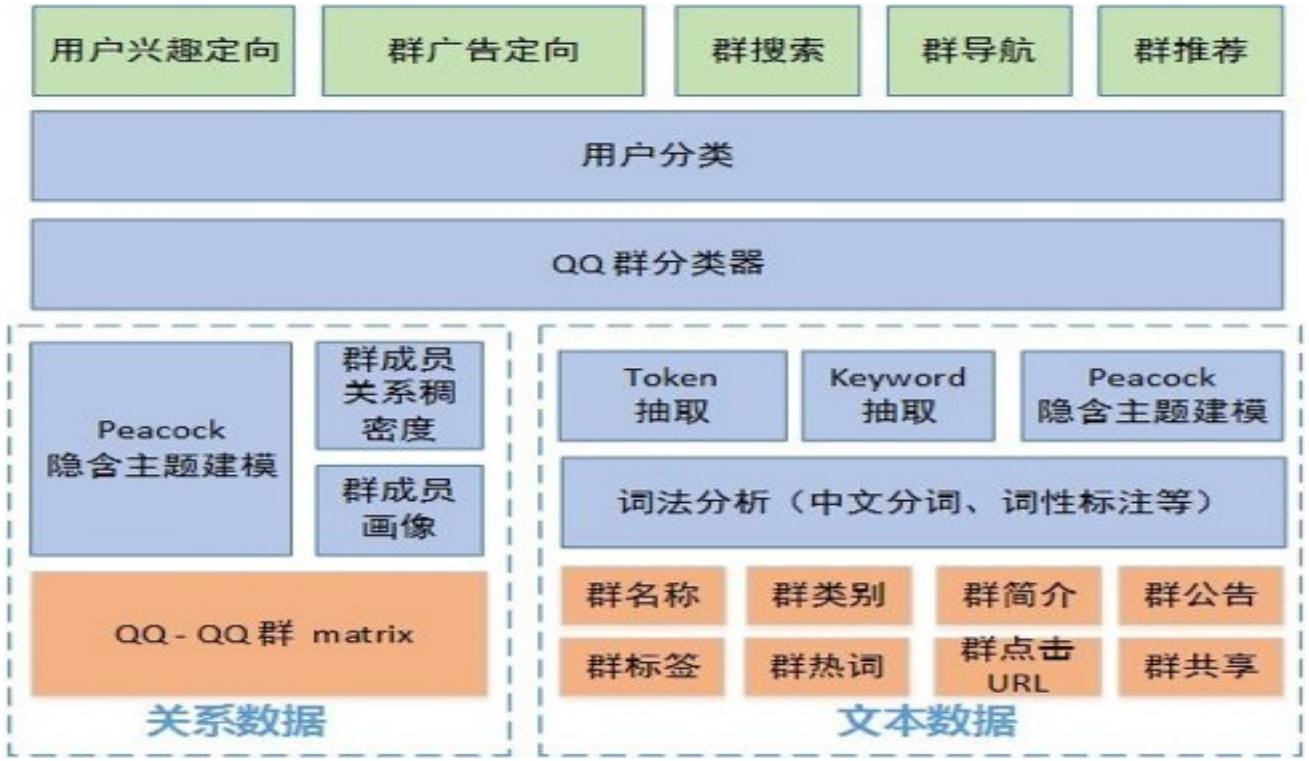


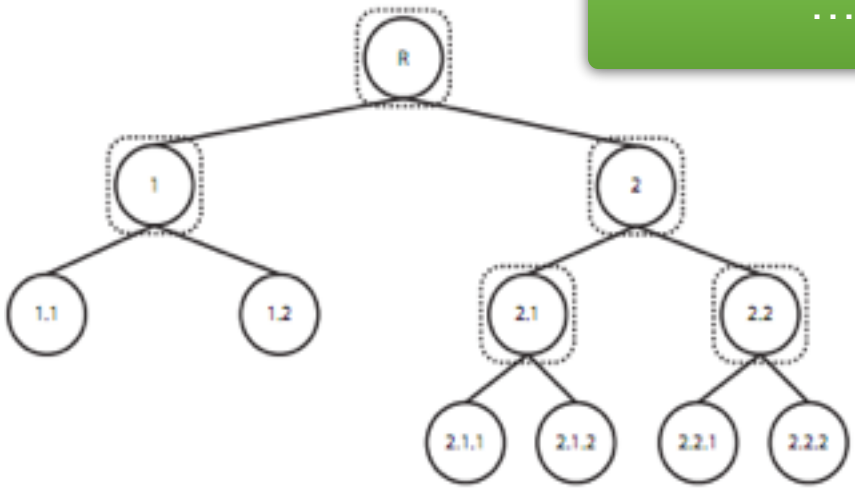
Matrix Type	Size
SearchQuery-word 矩阵	**10亿 x 20万**
QQ-QQ群 关系链矩阵	**7亿 x 2亿**
User-APP 安装列矩阵	**1亿 x 30万**
QQ-QQ 关系链矩阵	10亿x10亿
QQ-URL 点击矩阵	10亿x100亿

Peacock应用： QQ群语义挖掘，分解User-QQ群矩阵



QQ群语义挖掘：层次分类器





- 圆圈表示类别节点
- 二层分类体系，一共 100+结点
- 边表示类别节点间的父子关系
- 虚线椭圆表示训练的子分类器

教育

金融

.....

QQ群语义挖掘：QQ 群用户商业兴趣挖掘

- **Peacock 模型训练**

  文本类：10 亿 query log，20w words，10w topics，160 台机器，一周训练

  关系类：5 亿 QQ，1 亿 QQ 群, 1w topics，160 台机器，2 天训练

- **分类模型训练**

  二层分类体系，一共 100+结点，MaxEnt Model　标注8万 QQ 群

- **离线效果评测**

特征集	一级行业			二级行业		
	测试样本数	准确率	召回率	测试样本数	准确率	召回率
BOW(bag of words)	12987	82.33%	80.14%	12454	79.96%	79.96%
BOW+ peacock topics	12987	**86.82%**	**84.18%**	12454	**83.05%**	79.20%

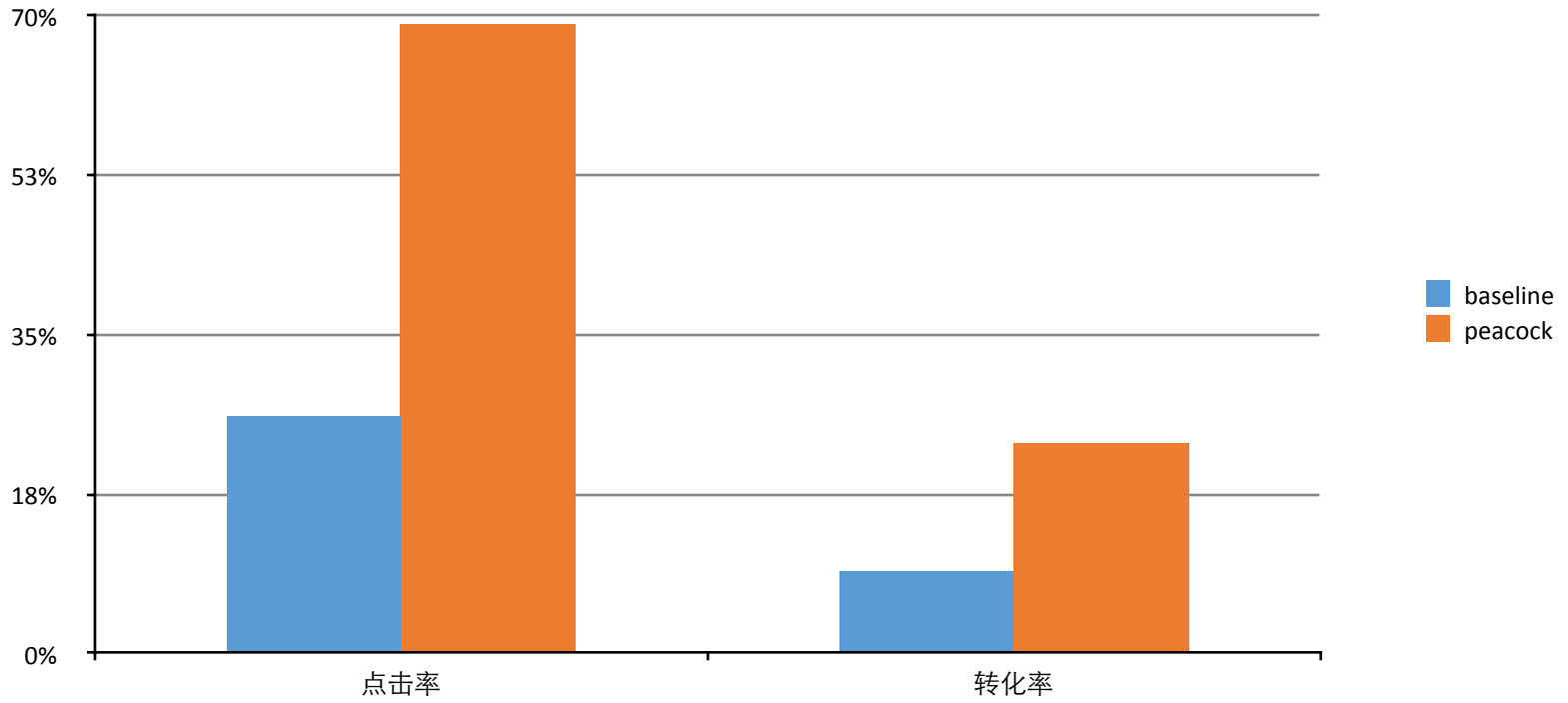
- **初步线上定向效果检验**

  引入5家广告主做线上 A/B test 投放测试，CTR **40% ↑**

QQ群推荐： online

$$p(\text{QQ群}|user) = \sum_{topic} p(\text{QQ群}|topic) \cdot p(topic|user)$$

QQ群推荐：效果



Thanks for your attentions!