大规模主题模型建模及其在腾讯业务中的应用

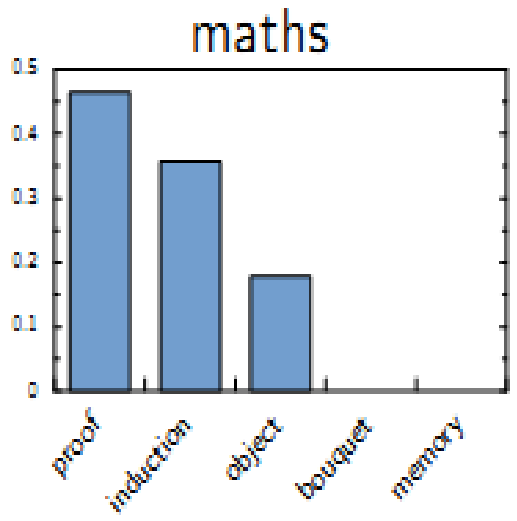
Rickjin(靳志辉)
腾讯SNG效果广告平台部

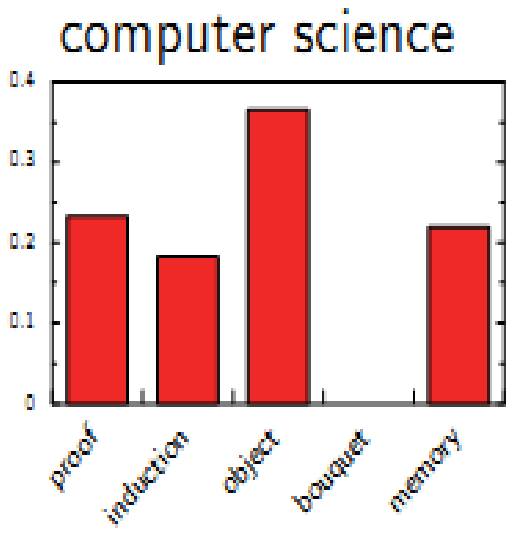
Outline

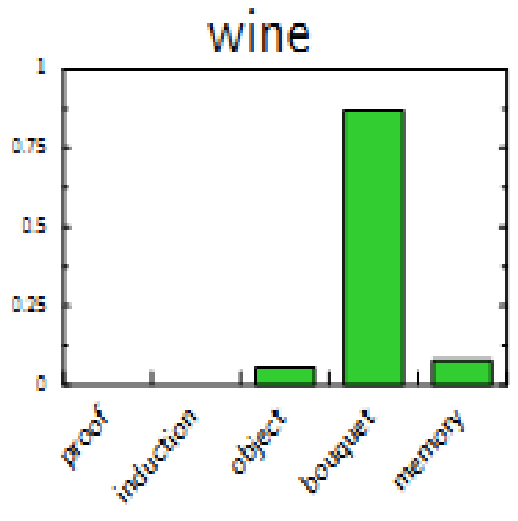
- **主题模型背景介绍**
- **大规模主题模型学习系统 Peacock**
- **Peacock 在腾讯业务中的应用**

Doc-Topic Structure

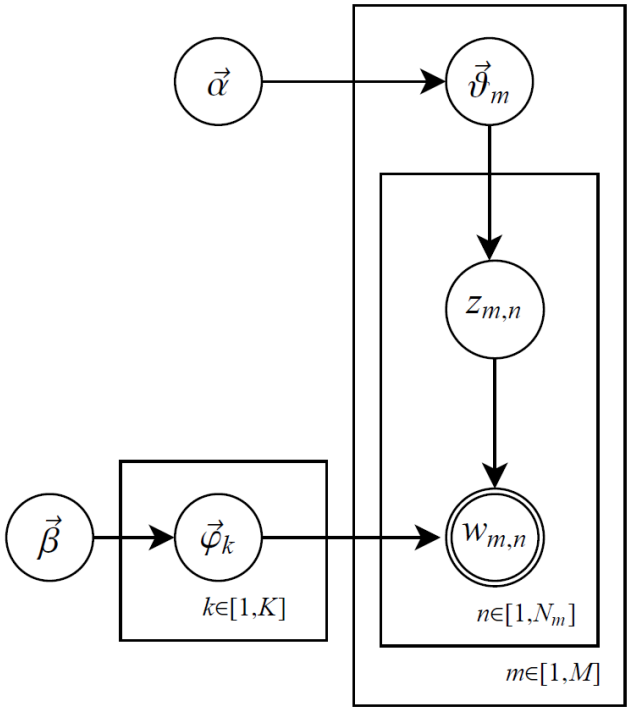
- Doc 是由 topic 组成的
- Topic 是 Vocab 上的概率分布 [Hofmann, 1999]



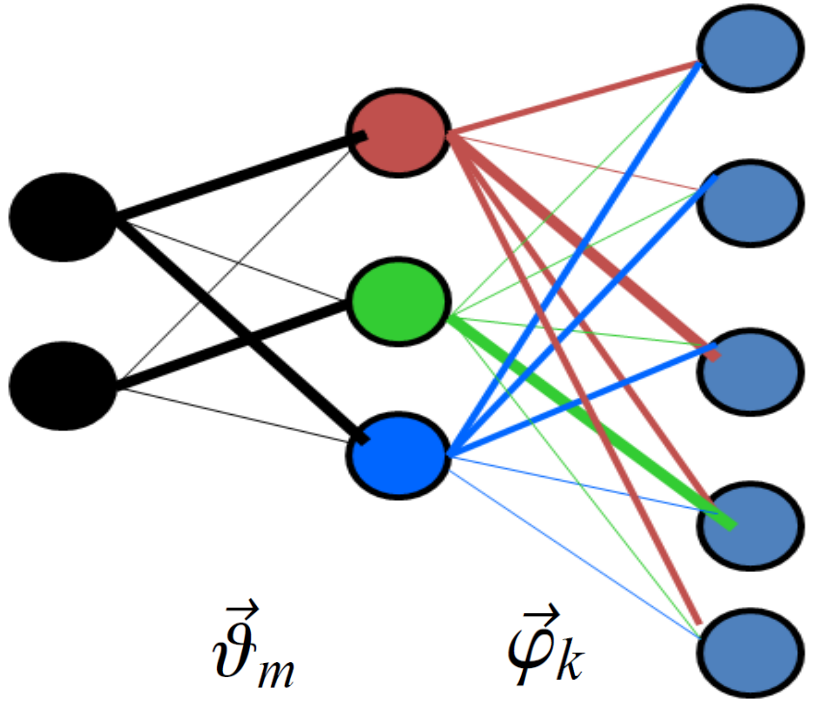




LDA Topic Modeling



Docs Topics Words

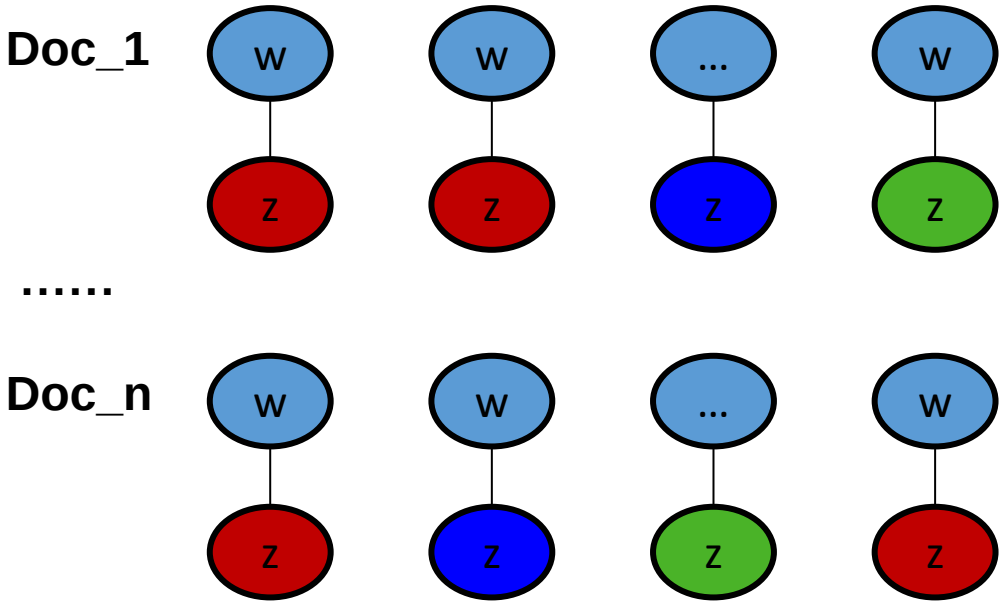


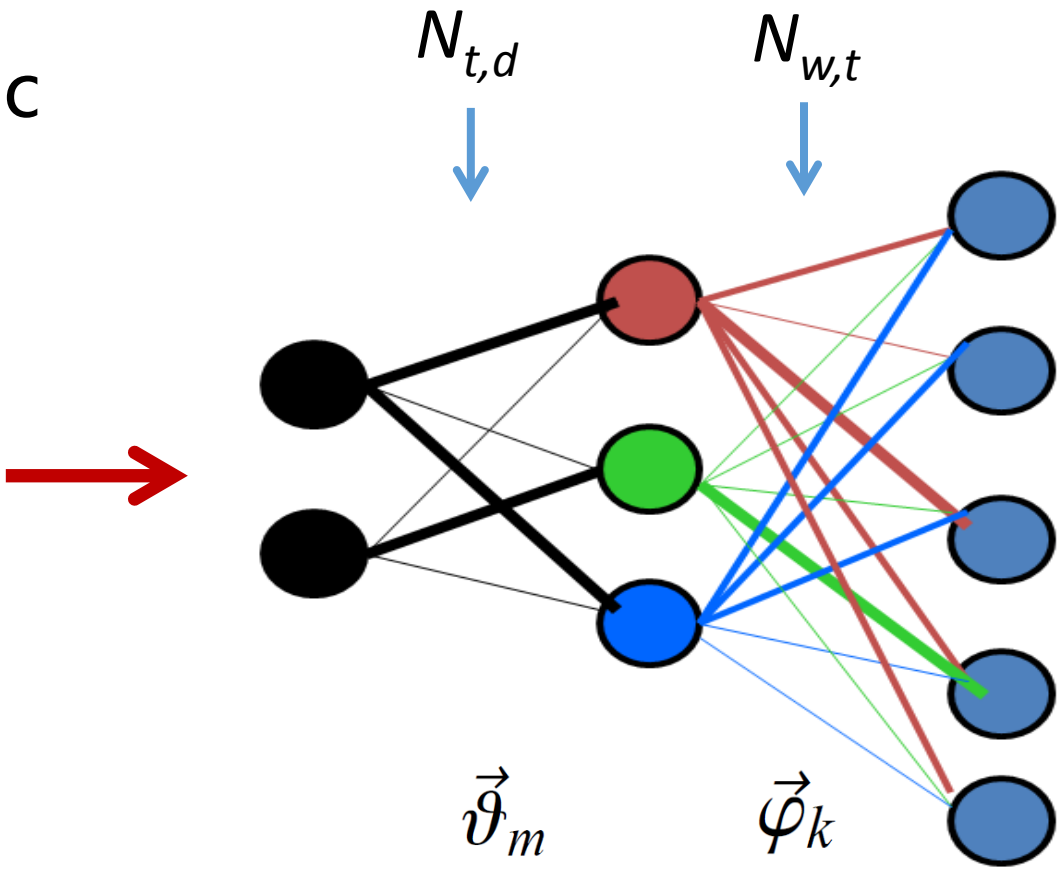
$$p(z_i{=}k|\vec{z}_{\neg i}, \vec{w}) \propto \frac{n^{(t)}_{m,\neg i} + \alpha_k}{[\sum_{k=1}^{K} n^{(k)}_m + \alpha_k] - 1} \cdot \frac{n^{(t)}_{k,\neg i} + \beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i} + \beta_t}$$

P(topic|doc) P(topic|word)

LDA Model Training (1)

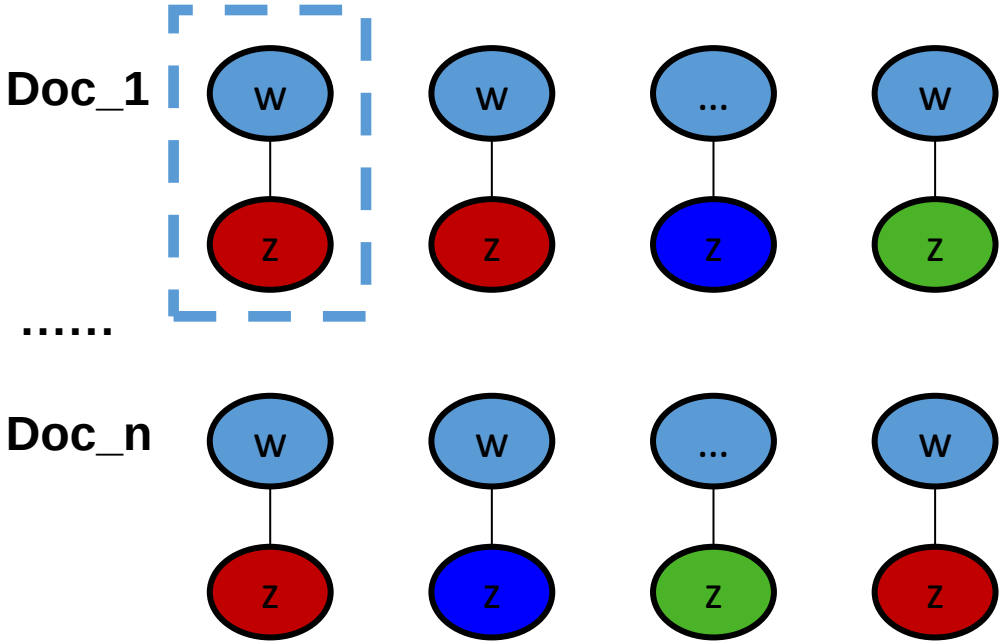
Step1: 随机初始化每个词的 topic

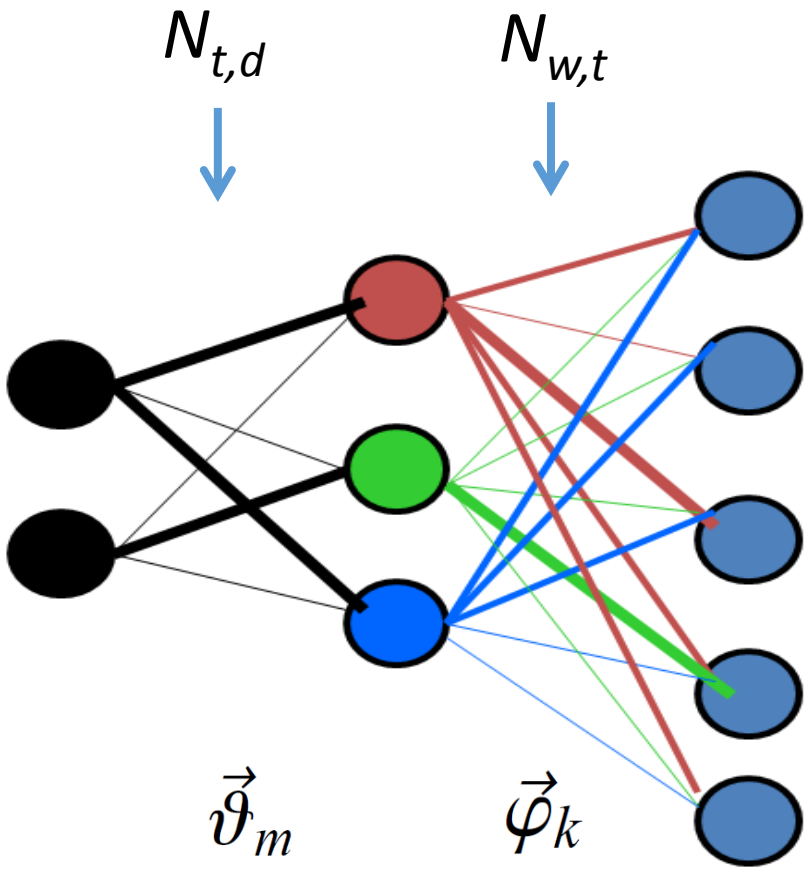




LDA Model Training (2)

Step2: 重新采样每个 topic, 更新计数



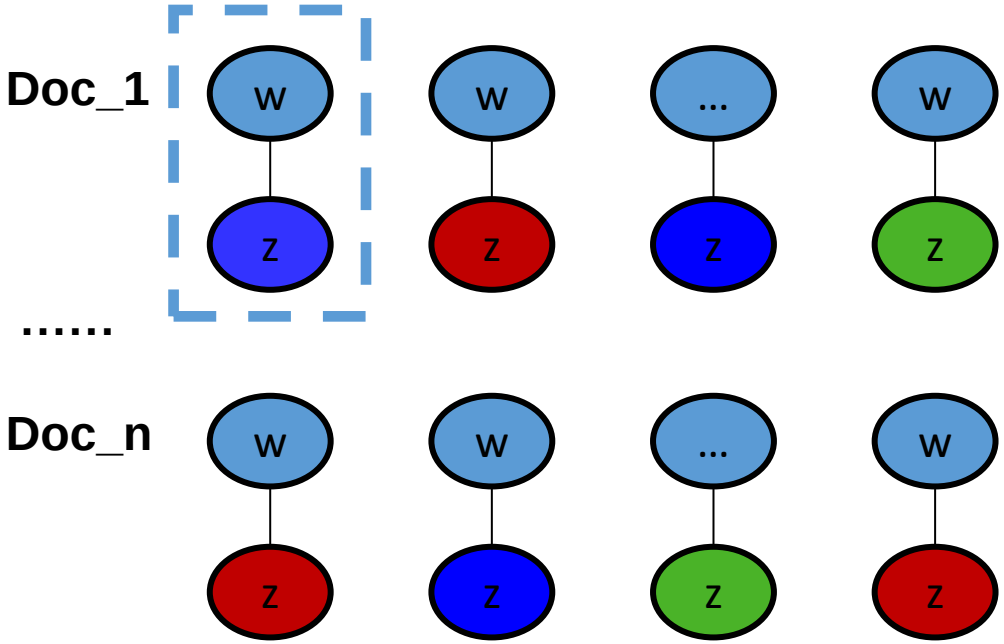


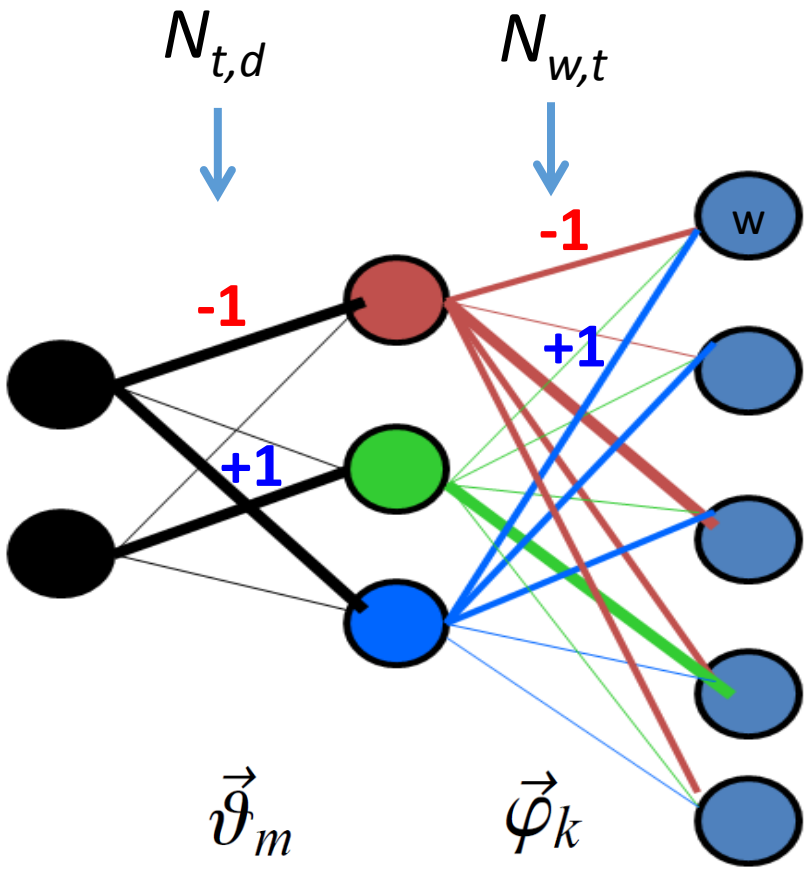
$$p(z_i{=}k|\vec{z}_{\neg i},\vec{w}) \propto \frac{n^{(t)}_{m,\neg i}+\alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m}+\alpha_k]-1} \cdot \frac{n^{(t)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i}+\beta_t}$$

P(topic|doc) P(topic|word)

LDA Model Training (3)

Step3: 重新采样每个 topic, 更新计数



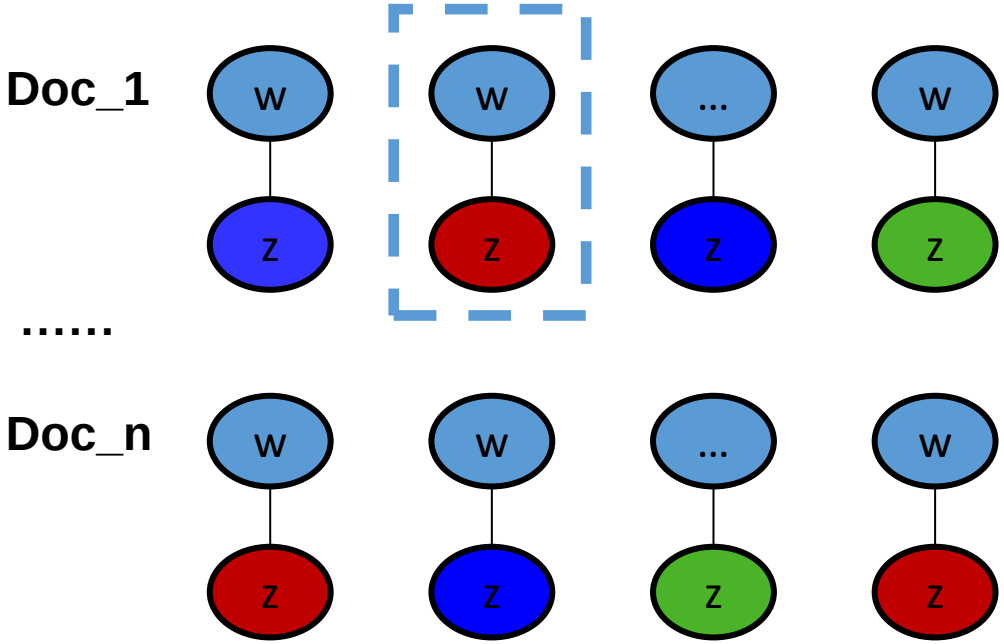


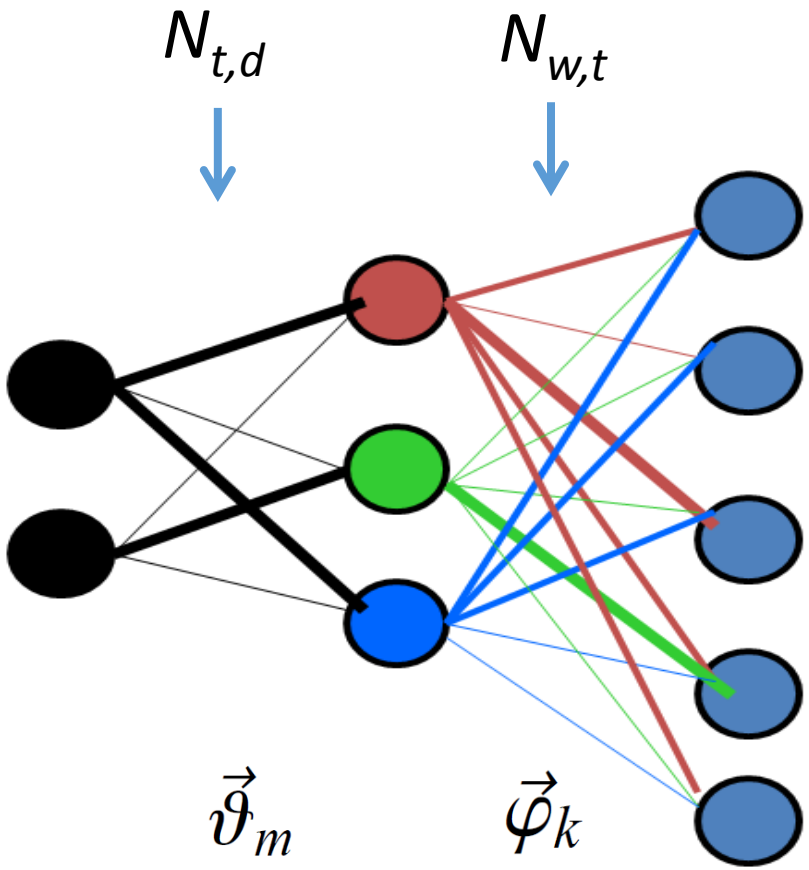
$$p(z_i{=}k|\vec{z}_{\neg i},\vec{w}) \propto \frac{n^{(t)}_{m,\neg i}+\alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m}+\alpha_k]-1} \cdot \frac{n^{(t)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i}+\beta_t}$$

P(topic|doc) P(topic|word)

LDA Model Training (4)

Step4: 重复 step2&3, 直到模型收敛

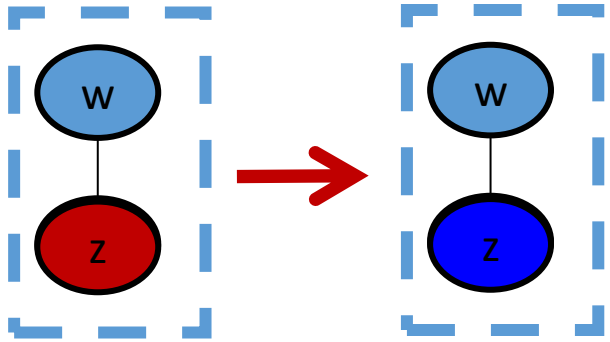




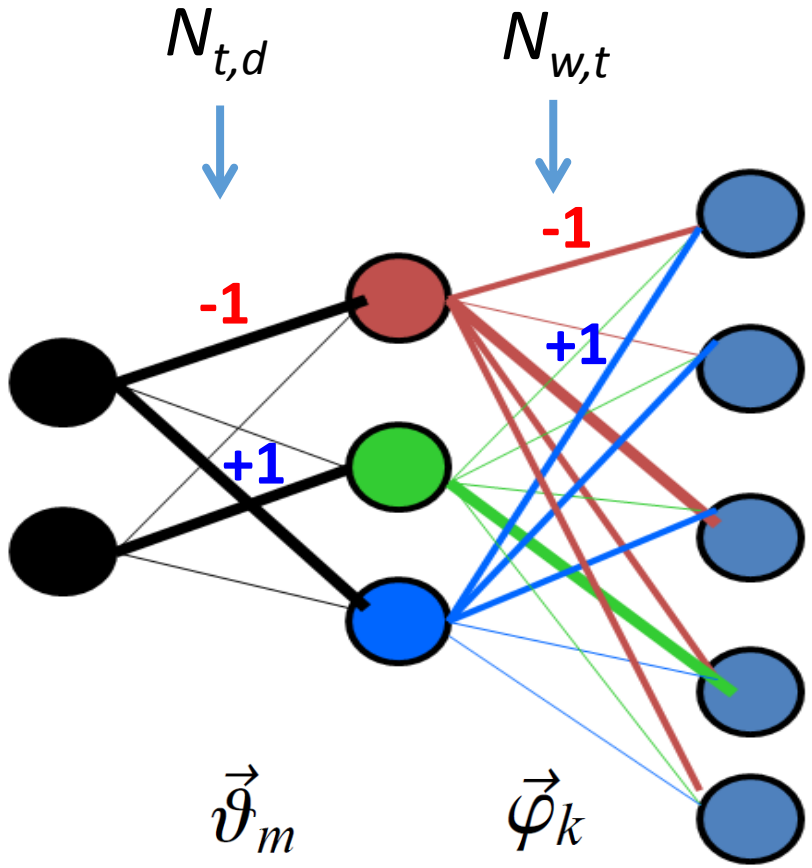
$$p(z_i{=}k|\vec{z}_{\neg i},\vec{w}) \propto \frac{n^{(t)}_{m,\neg i}+\alpha_k}{[\sum_{k=1}^{K} n^{(k)}_{m}+\alpha_k]-1} \cdot \frac{n^{(t)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i}+\beta_t}$$

P(topic|doc) P(topic|word)

Large-scale LDA Modeling

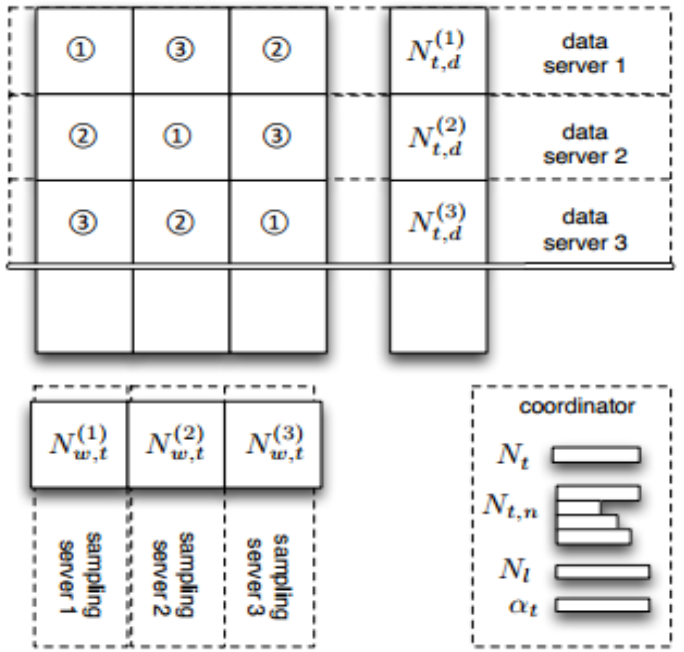


- Q1: 如何提升 Gibbs Sampling 速度
  - 标准采样算法太慢
- Q2: 如何支持大数据、大模型
  - 十亿文档，百万词汇，百万 topic
- Q3: 如何调参优化模型质量
  - alpha，beta 如何选取
  - topic 个数如何考虑



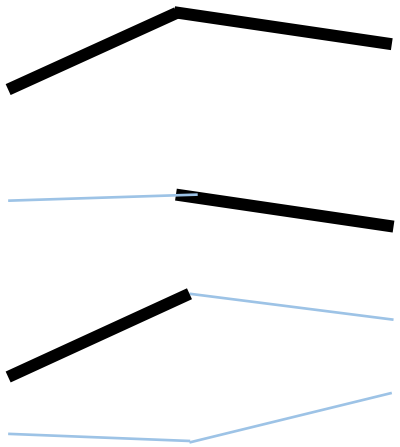
Peacock: Large-scale Topic Modeling

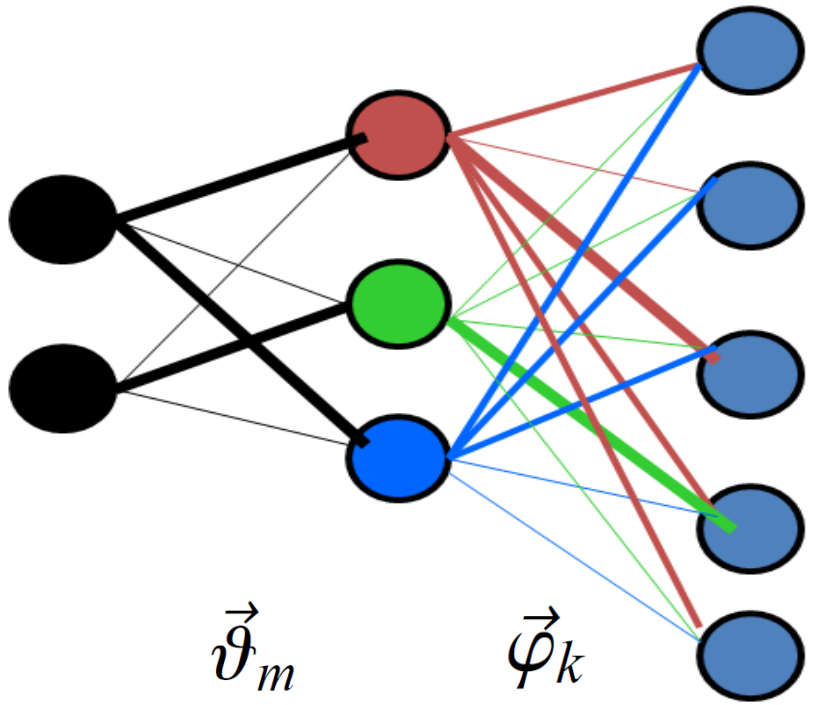
- Q1: 如何提升 Gibbs Sampling 速度
  - 使用 SparseLDA 算法做 Gibbs Sampling
  - **比标准 LDA 快30倍**
- Q2: 如何支持大数据、大模型
  - 基于 Go 语言实现
  - 矩阵分块并行计算
  - 可以支持**10亿 x 1亿**的矩阵分解
  - 可以支持**100万 topics** 计算
  - 类似 Google Rephil 系统，挖掘长尾语义
- Q3: 如何调参优化模型质量
  - 每轮迭代对超参数做优化，**智能训练 topics 个数**



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



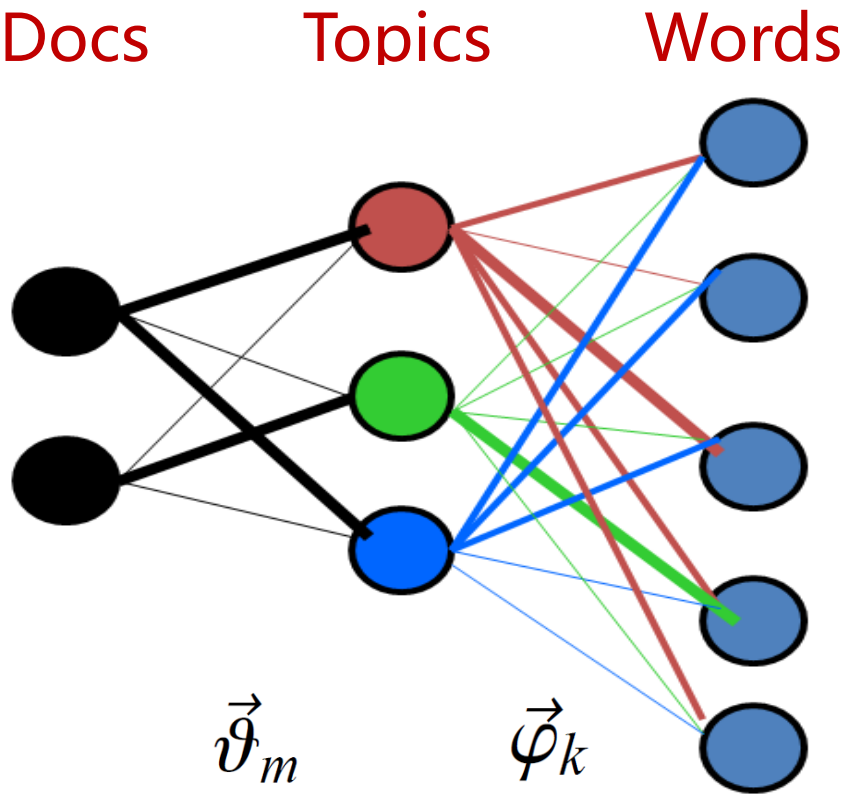


$$p(z_i{=}k|\vec{z}_{\neg i},\vec{w}) \propto \frac{n^{(t)}_{m,\neg i}+\alpha_k}{[\sum_{k=1}^{K} n^{(k)}_m+\alpha_k]-1} \cdot \frac{n^{(t)}_{k,\neg i}+\beta_t}{\sum_{t=1}^{V} n^{(t)}_{k,\neg i}+\beta_t}$$

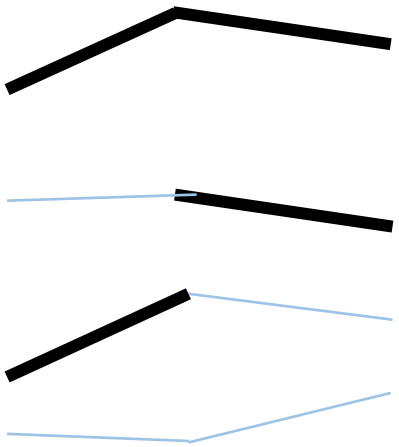
P(topic|doc)　　P(topic|word)

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

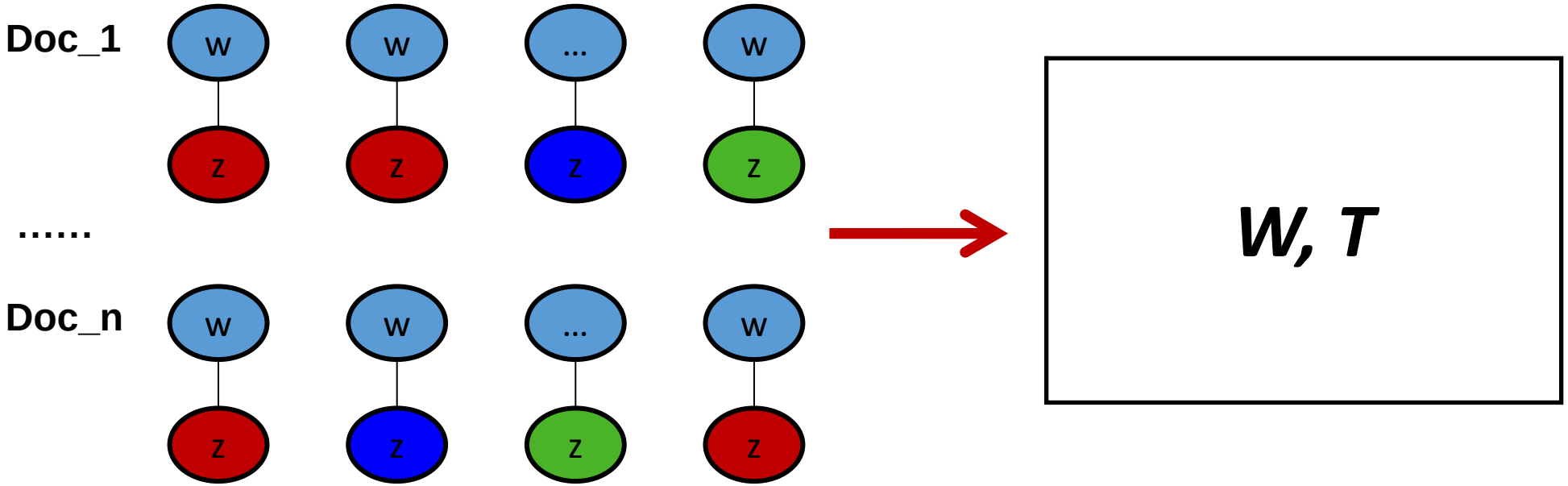


Limin Yao, David Mimno, and Andrew McCallum. ***Efficient Methods for Topic Model Inference on Streaming Document Collections.*** KDD 2009.

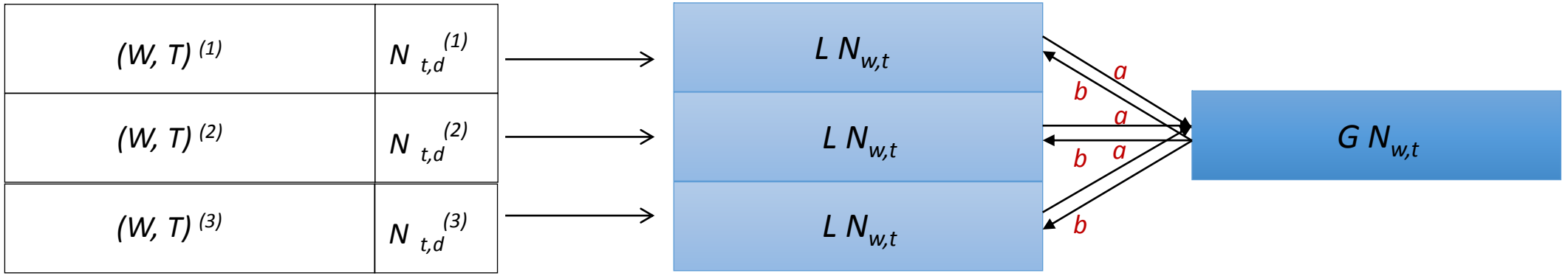


Path-ID	Probability
10	0.8
20	0.1
70	0.09
Sum of others	0.01

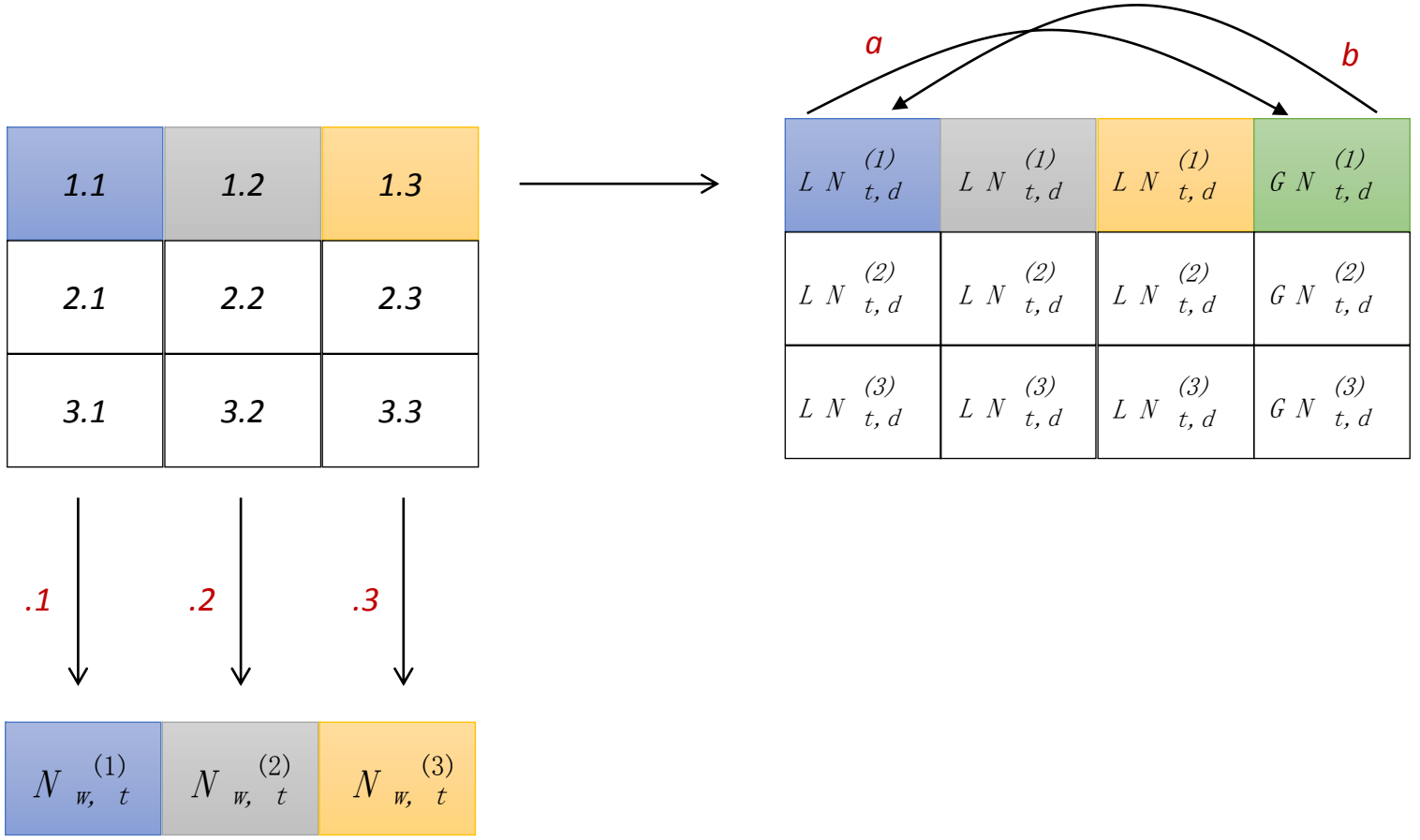
Q2: 十亿篇文档，百万词汇，百万 Topics



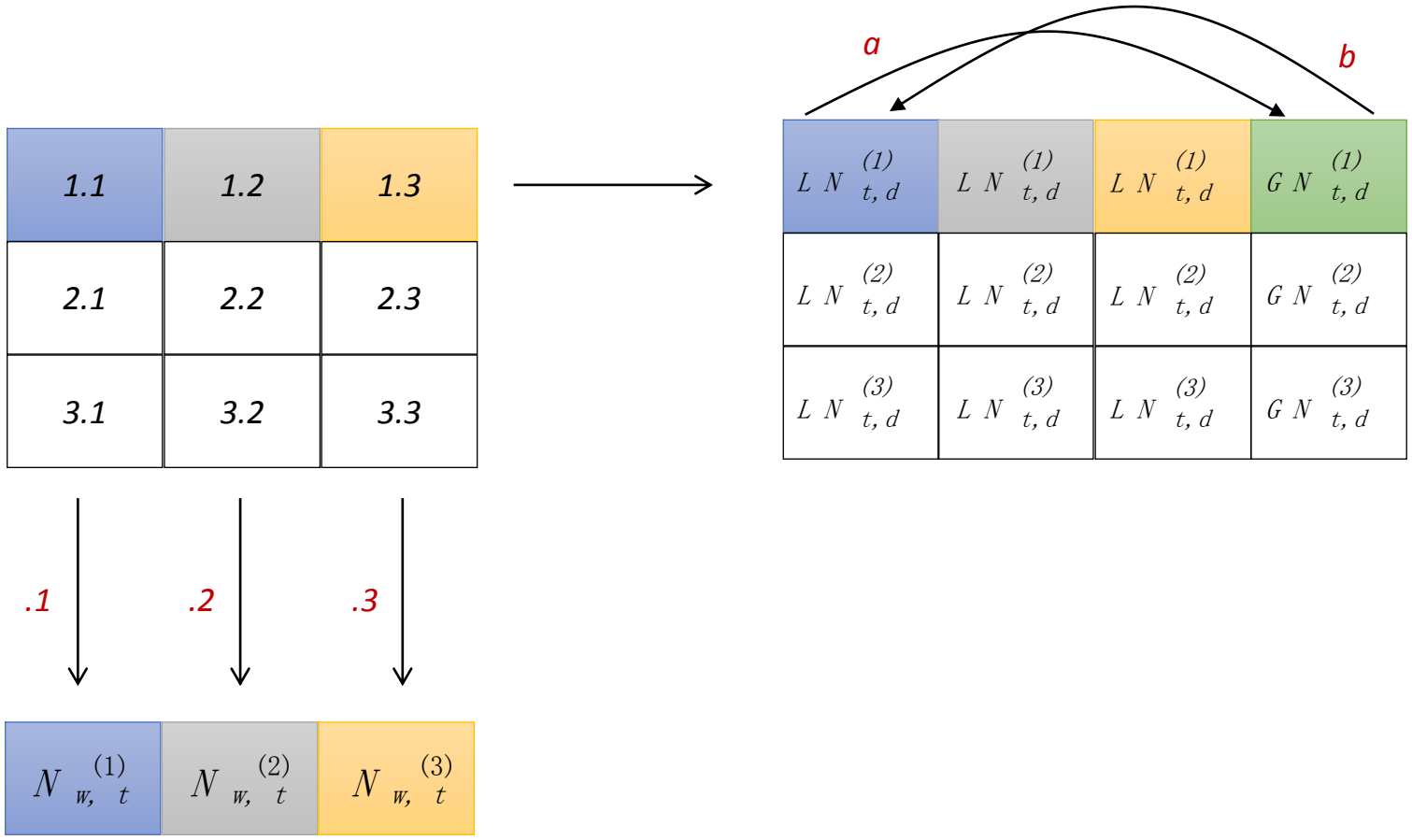
AD-LDA (Data Parallelism)



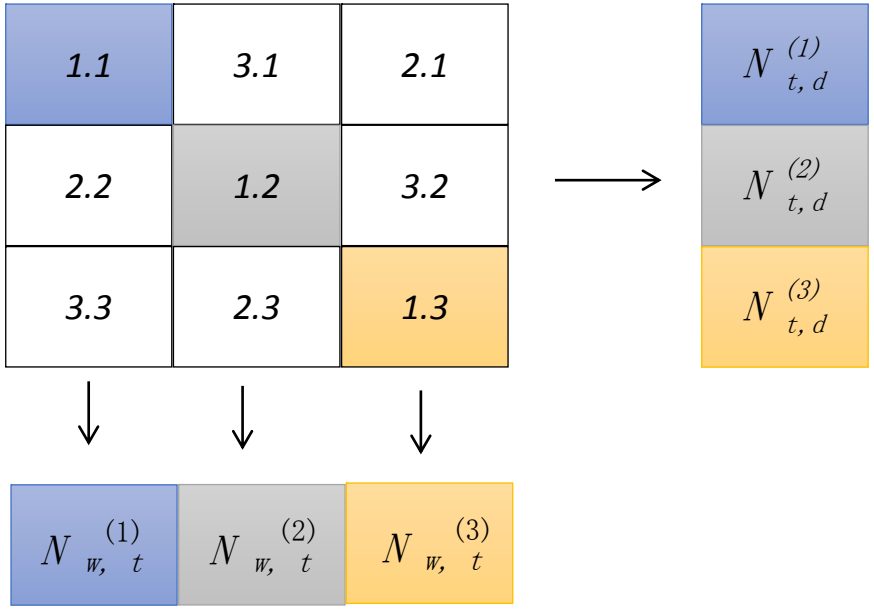
Model Parallelism - 1

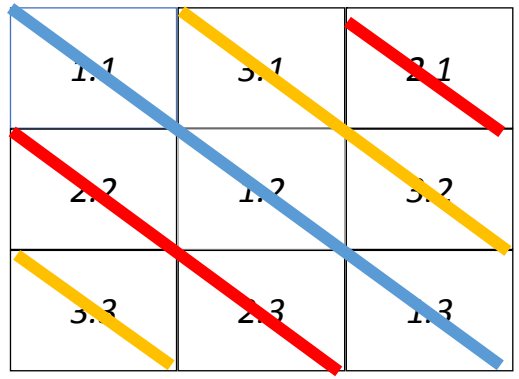


Model Parallelism - 1

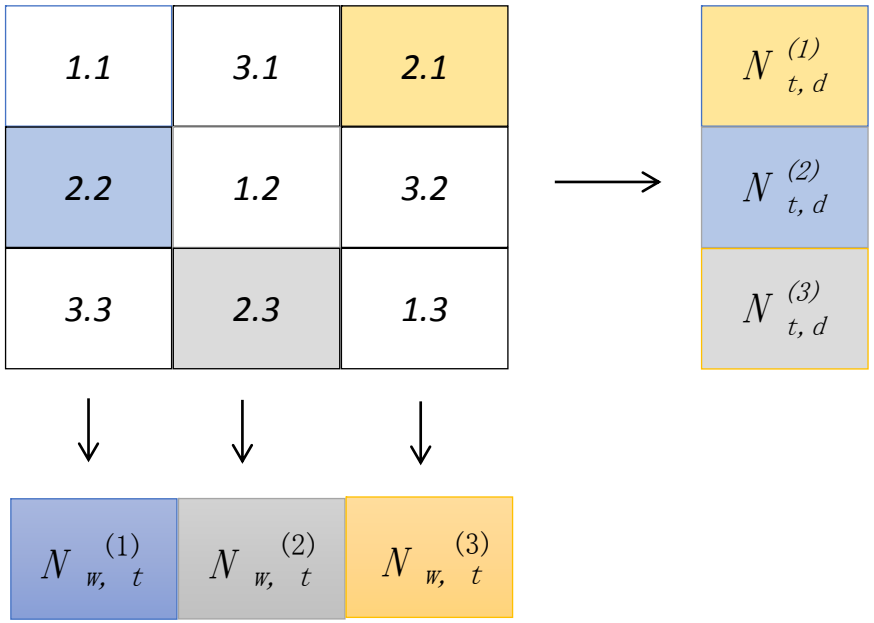


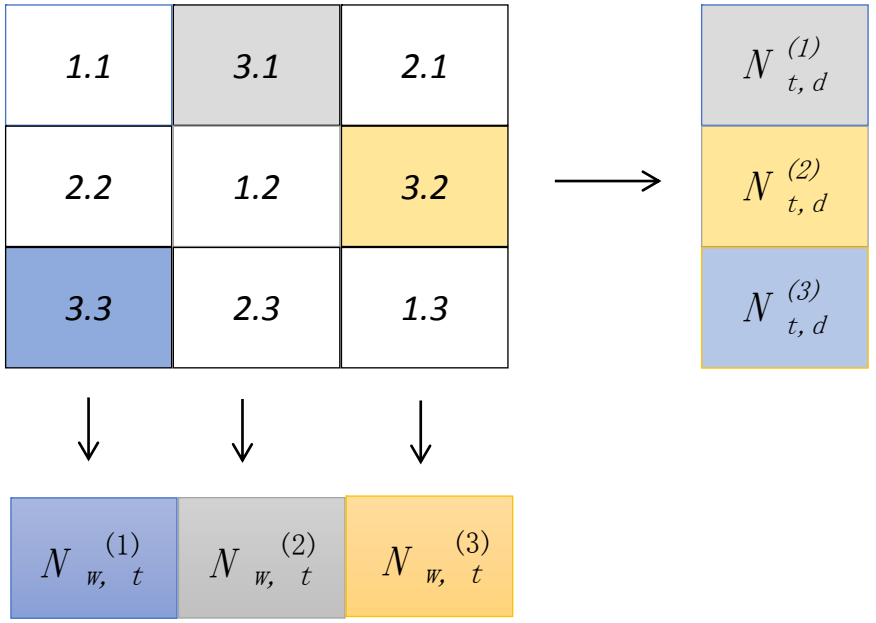
Lock-free Synchronization



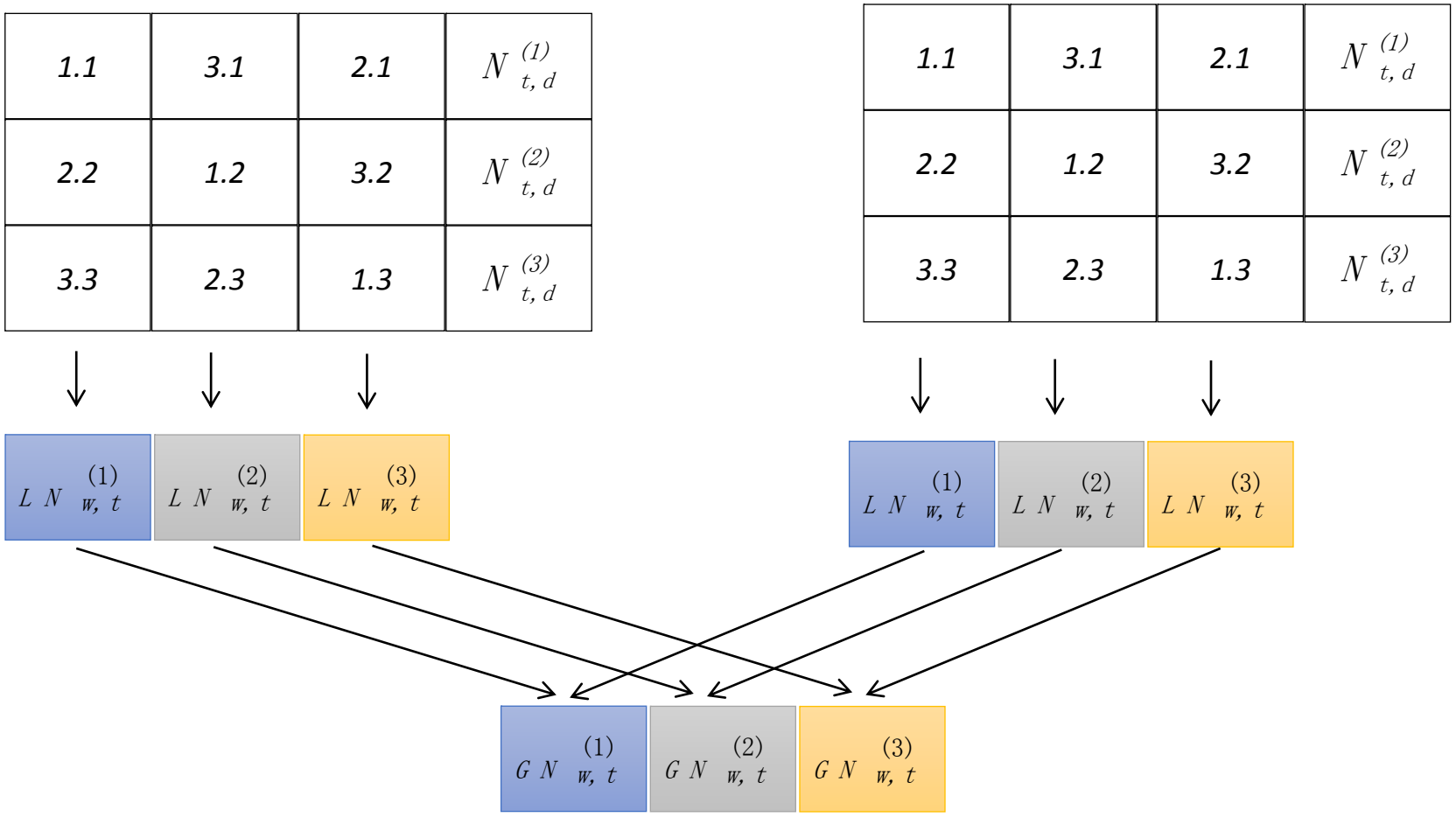


Lock-free Synchronization





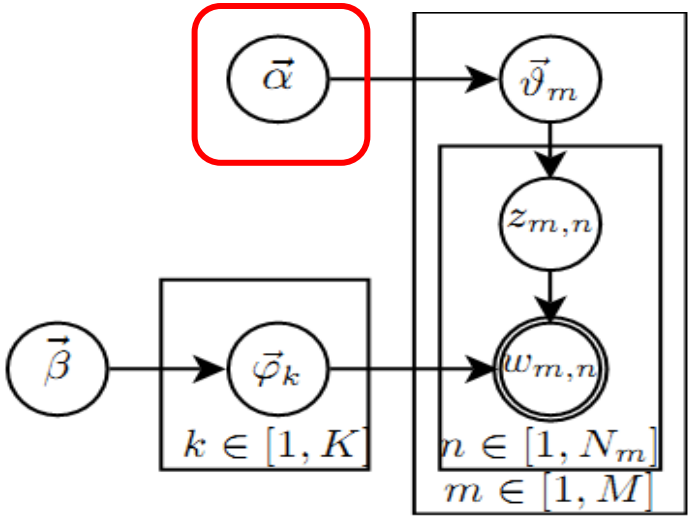
Model Parallelism + Data Parallelism

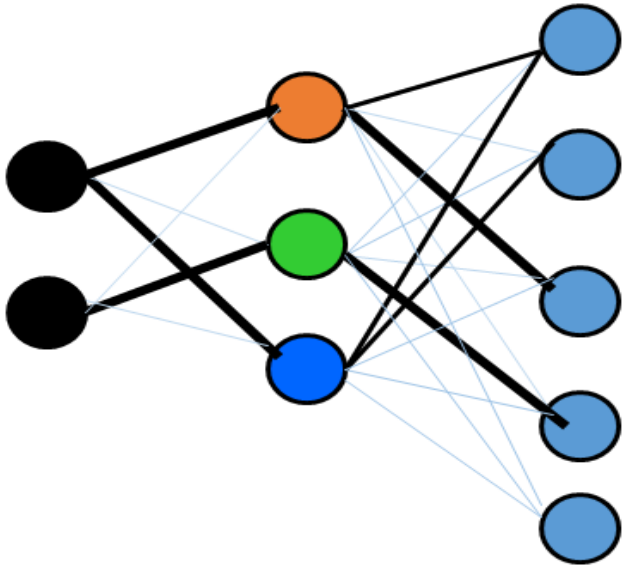


Q3: 优化模型质量

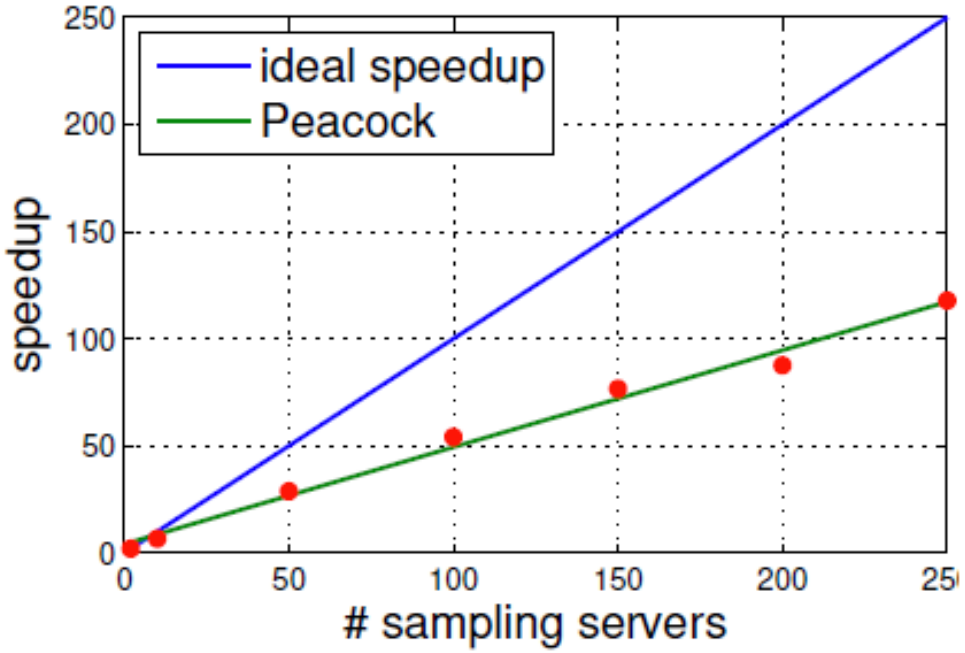
Rethinking LDA: Why Priors Matter Hanna M. Wallach
David Mimno Andrew McCallum NIPS 2009

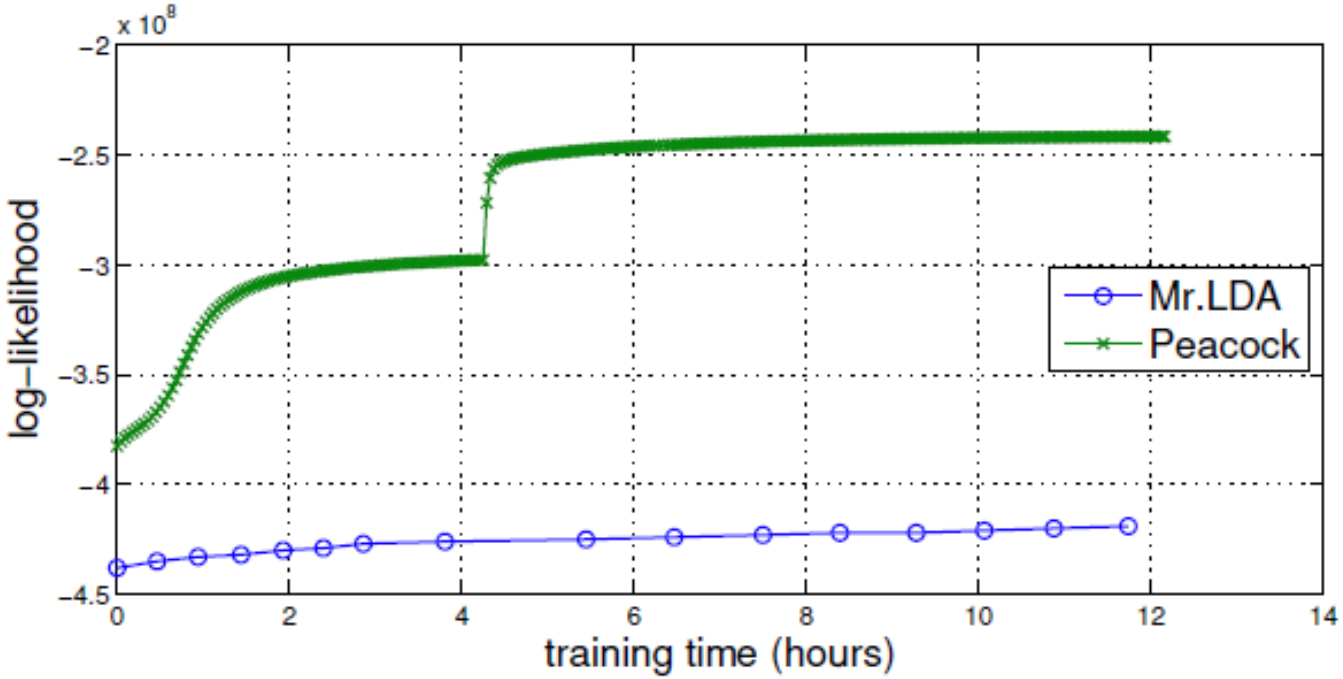
- 超参数 alpha 对模型质量有重要的影响
- 每轮迭代中，通过 MLE 估计优化 alpha



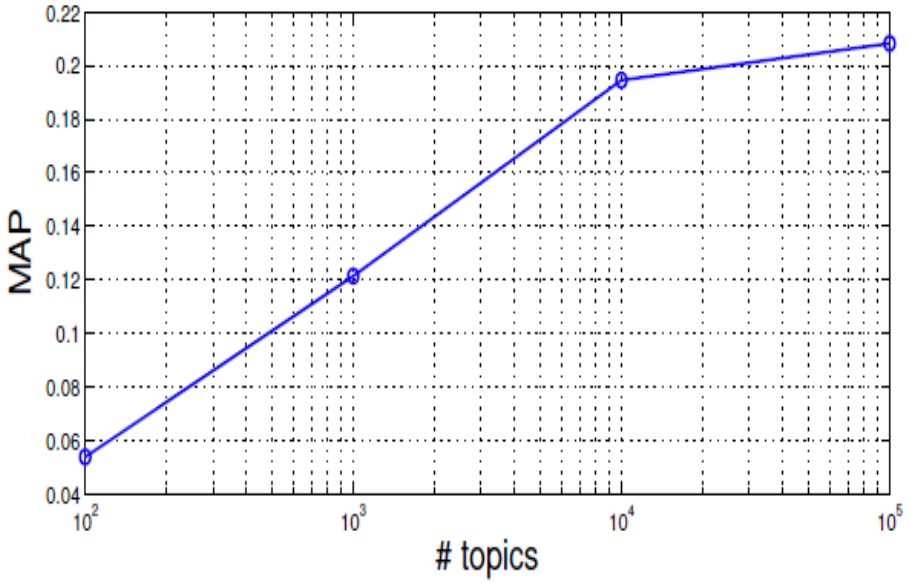


Peacock 性能

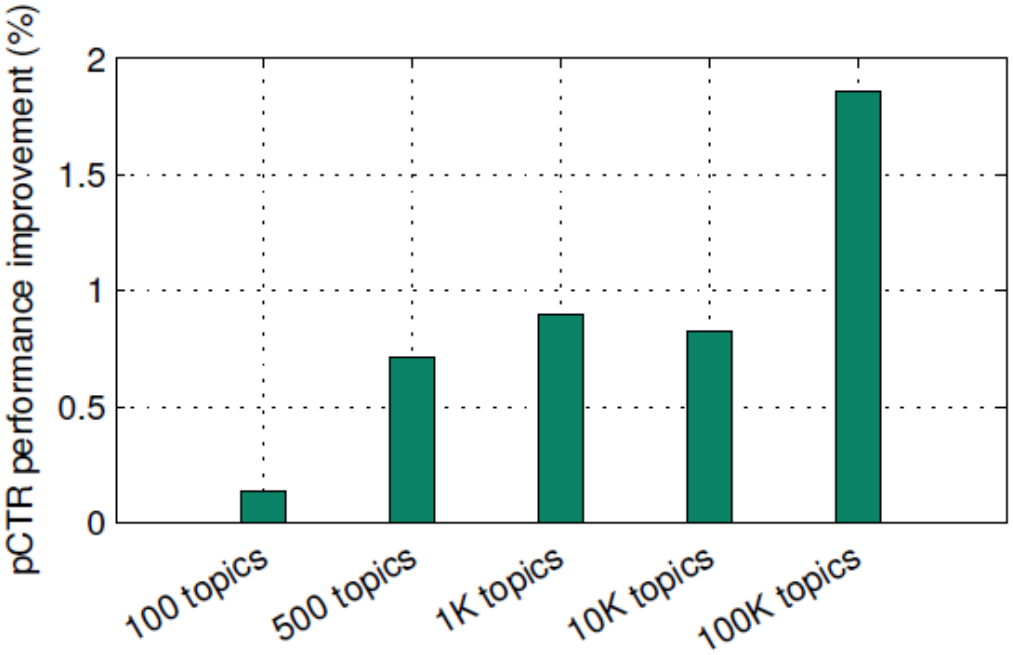

ideal speedup
Peacock
speedup
sampling servers
0
50
100
150
200
250


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

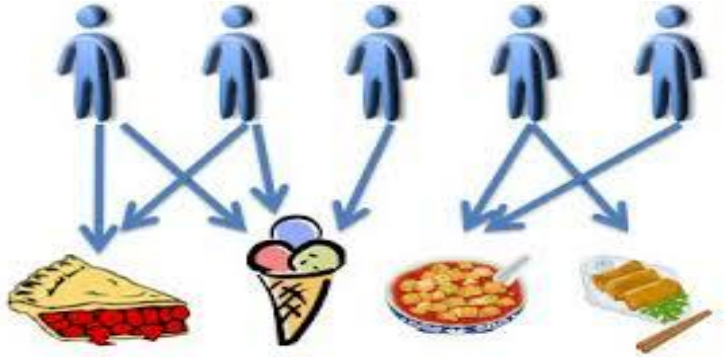
腾讯业务中的应用

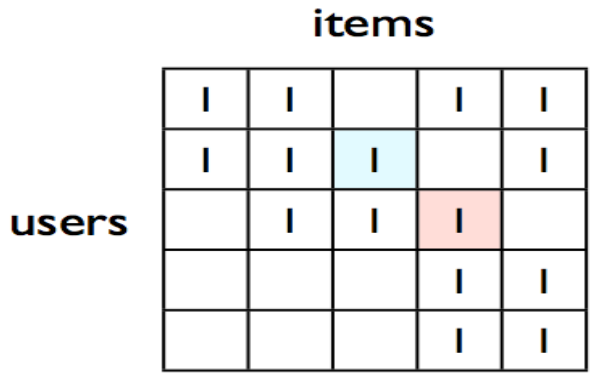
- 文本语义分析
- 广告相关性计算与 CTR 预估
- QQ群推荐
- QQ 群分类与广告定向

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做文本语义理解

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)]

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

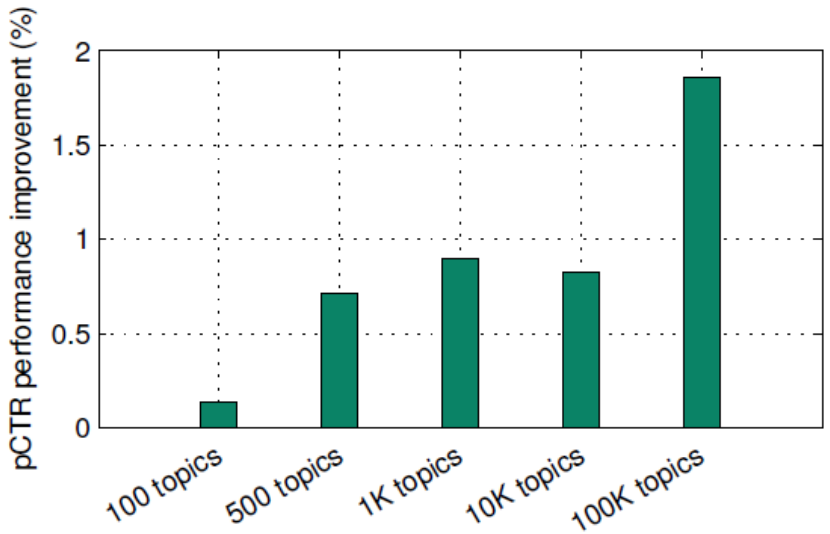
  离线 pCTR AUC 提升 **1.8%**

  在线实验 AdCTR 提升 **8.82%**

Peacock应用 - CTR

- 广告的Topic作为特征加入pCTR模型

Topic数量越多AUC提升越大



- 特征设计：P(w|d) = ∑P(w|z) P(z|d)，d-文档、z-主题、w-词
- 线上实验效果，CTR提升↑ 8.82%，RPM提升↑ 8.87%

TextMiner 语义分析平台

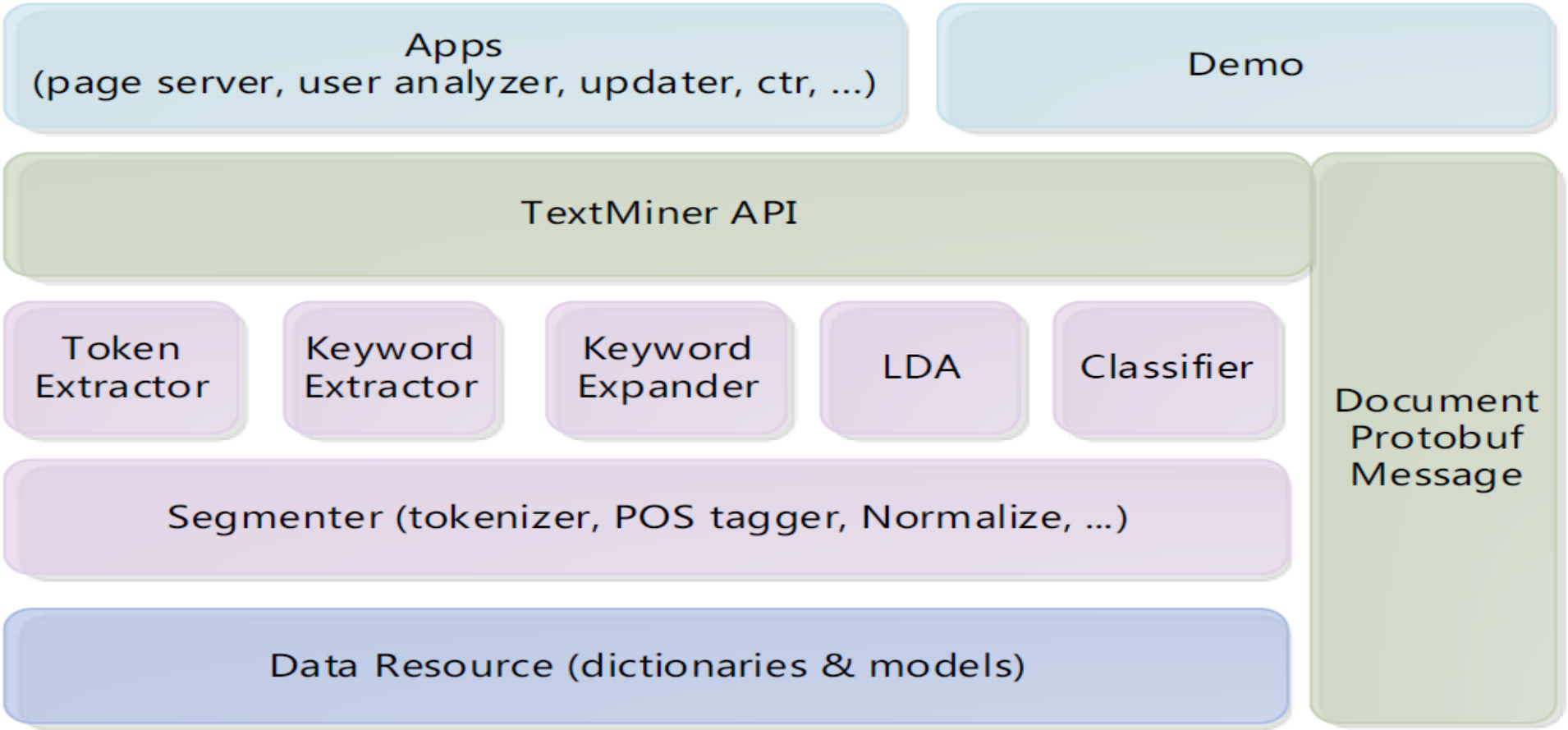
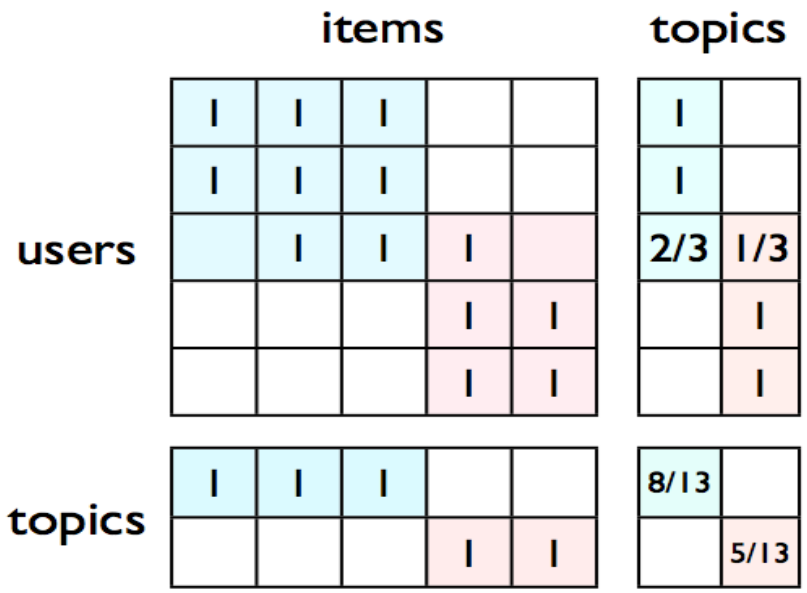
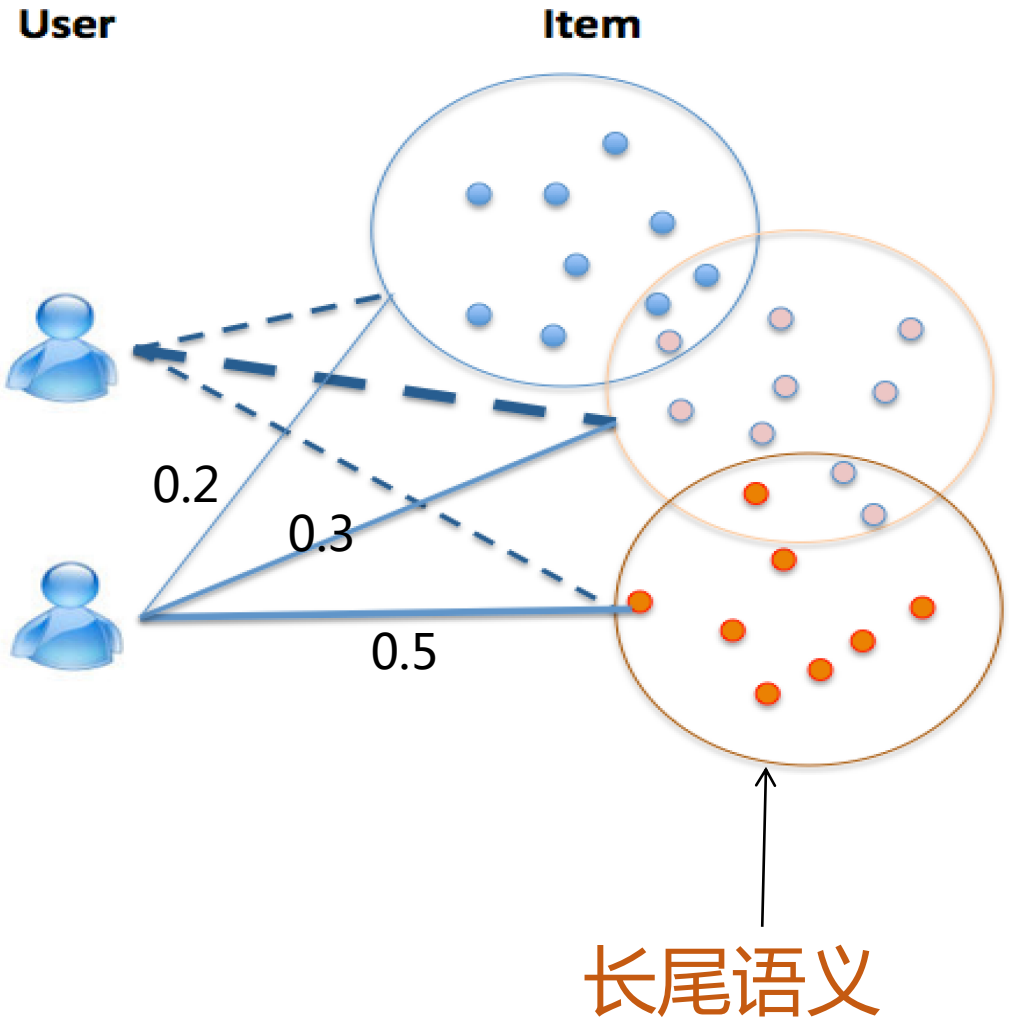


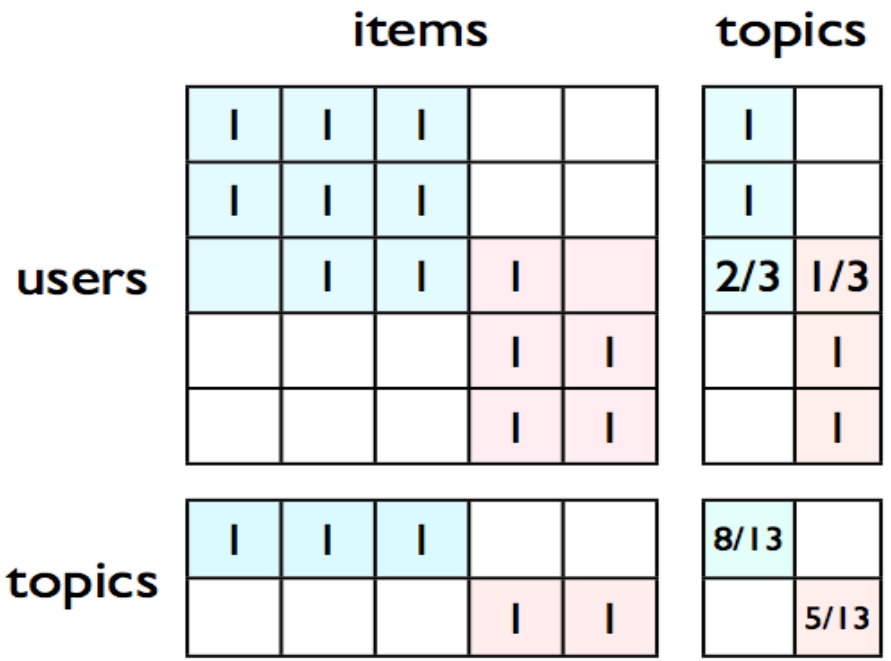
图 1 TextMiner 系统架构图

Recsys: 矩阵分解


items
topics
users
topics
1
2/3
1/3
8/13
5/13


User
Item
0.2
0.3
0.5
长尾语义

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

301655190:散户股票联盟_股票炒股黄金白银期货交流|融资融券信用卡贷款|短线牛股涨停黑马私募推荐|散户集中营|
204778270:散户股票联盟_股票炒股黄金白银期货交流|融资融券信用卡贷款|短线牛股涨停黑马私募推荐|散户集中营|
281643833:散户股票同盟_股票炒股黄金白银期货交流|融资融券信用卡贷款|短线牛股涨停黑马私募推荐|散户集中营|
291589134:散户股票联盟_股票炒股黄金白银期货|内幕信息私募合作操盘|短线暴涨牛股涨停黑马推荐|散户联合坐庄|
145682621:散户投资同盟_股票炒股黄金白银期货|内幕信息私募合作操盘|短线暴涨牛股涨停黑马推荐|散户联合坐庄|
181160252:散户部落PK私募_团结一切可以团结的散户力量，狂拉一个股票，达到互相盈利目的！股票炒股黄金白银期货
301994161:散户股票联盟_股票炒股黄金白银期货|内幕信息私募合作操盘|短线暴涨牛股涨停黑马推荐|散户联合坐庄|

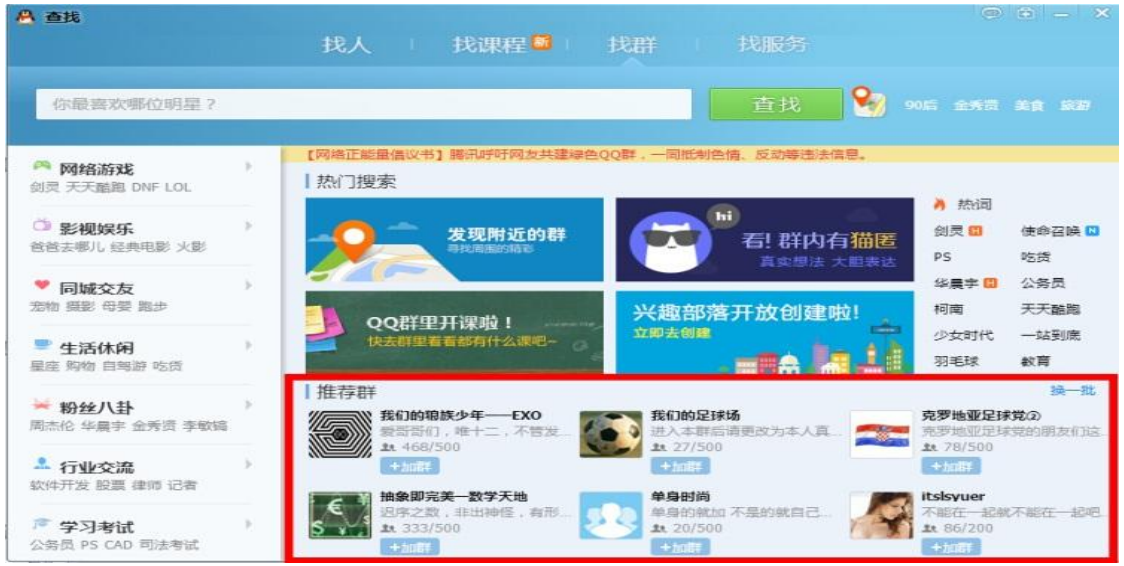
142471971:塔防三国S1老玩家军团_亲爱的：朋友兄弟们祝你们玩的开心！聊的舒心！本群独有的高级千人群\n互相交流
256443227:塔防小助手VIP群_
324615870:塔防三国伴侣交流群_塔三伴侣辅助专卖店：shop62657742.taobao.com
278413679:塔三吧【2群】_本群为塔防三国志文韬武略（S1）服务器！
164452487:塔防三国千人群_欢迎各路高手低手新手菜鸟老鸟加入，热爱三国！！！热爱塔防！！！拒绝广告，拒绝黄图
142164443:塔防三国Happy家族_长久开聊-技巧-攻略
314118916:塔防招募心得群_
135855153:塔防三国 千人售后3_塔防三国志 活跃10元 淘宝交易 广告绕道

109273480:济南孕妈妈_济南孕妈妈！！！
143256869:大眼猪千人济南妈妈群_济南 大眼猪亲子网，济南第一亲子服务平台。官方网址：www.dayanzhu.com 很好记的哦
134143694:济南妈妈团购交流群_济南妈妈团购交流 此群已满请加新群192338181
105739142:济南妈妈总群_
200422692:好妈妈济南群_一切为了孩子，为了孩子的一切，做个好妈妈，交流育儿心得，让宝宝更健康！
154901747:济南妈妈群_大家一起聊聊宝宝的事，进群首先修改群名片，不然直接清理
63437957:济南宝宝活动群_
192338181:济南妈妈总群_

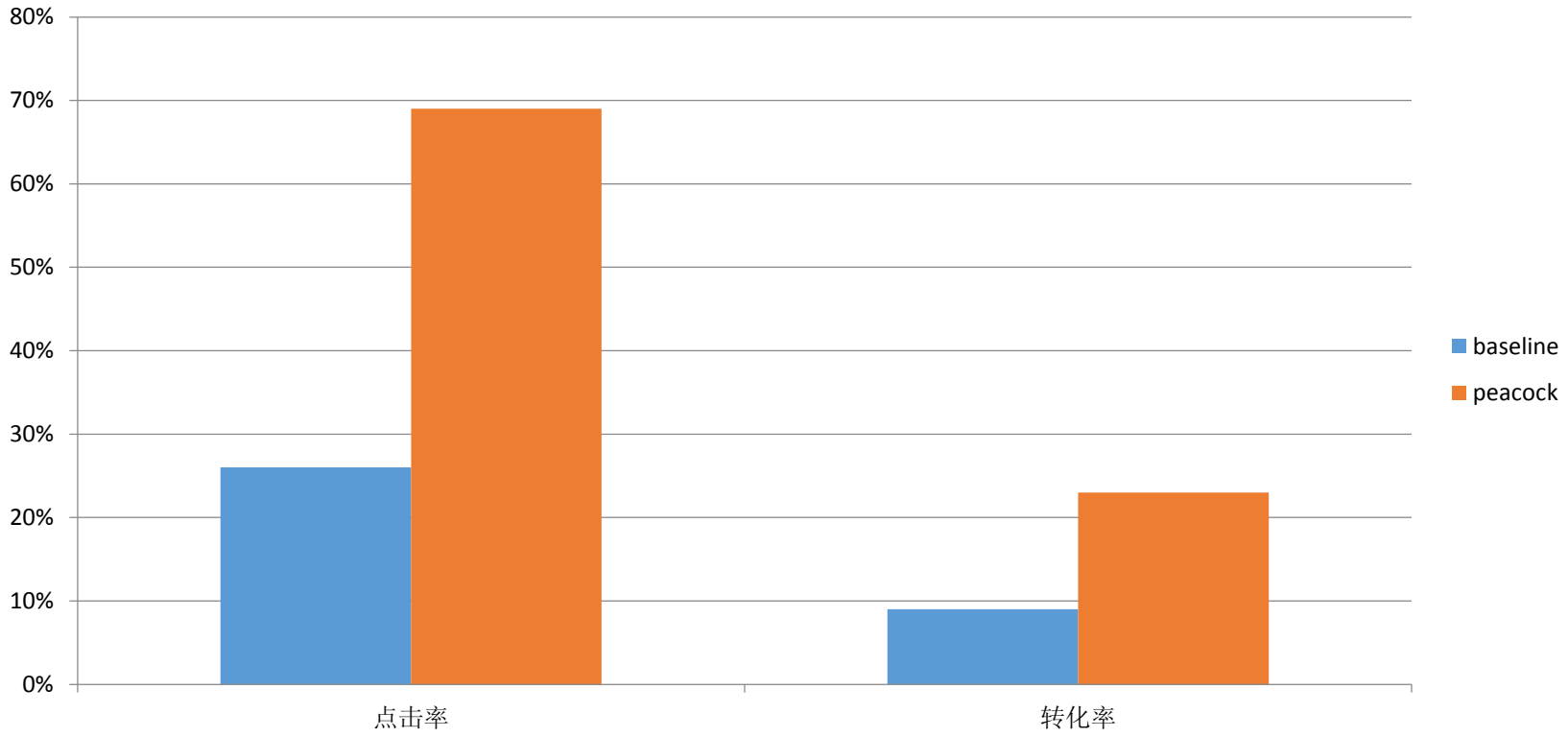
大家互相分享交流帮助！让游戏更快乐！
越级礼包；淘宝http://ttsm1983.taobao.com

QQ群推荐：online

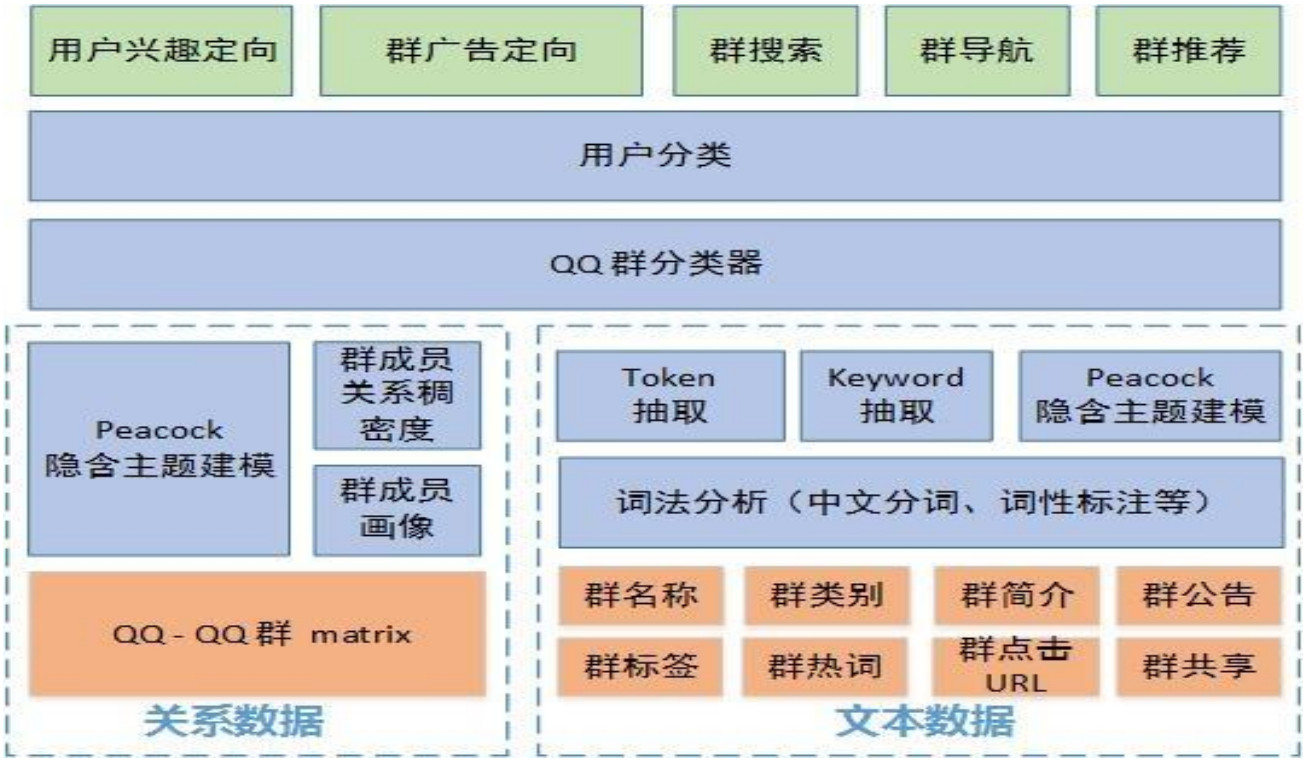
$$p(\text{QQ群}|user) \propto \sum_{topic} p(\text{QQ群}|topic)\, p(topic|user)$$

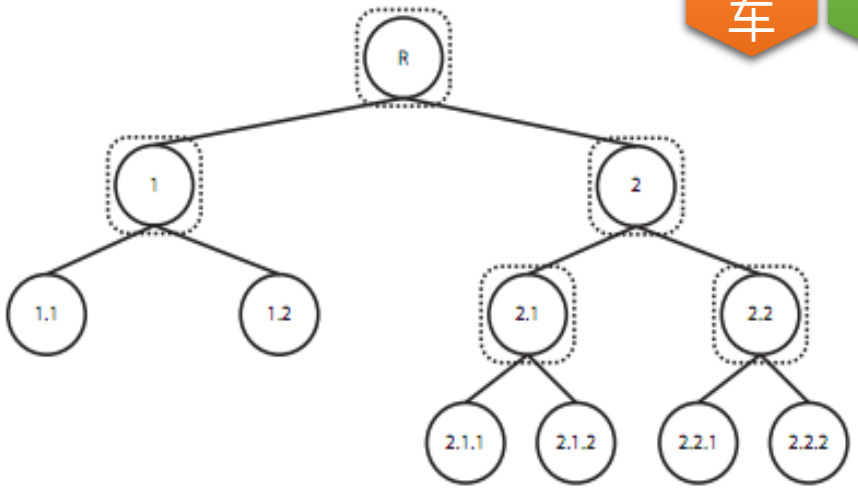


QQ群推荐：效果



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

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**

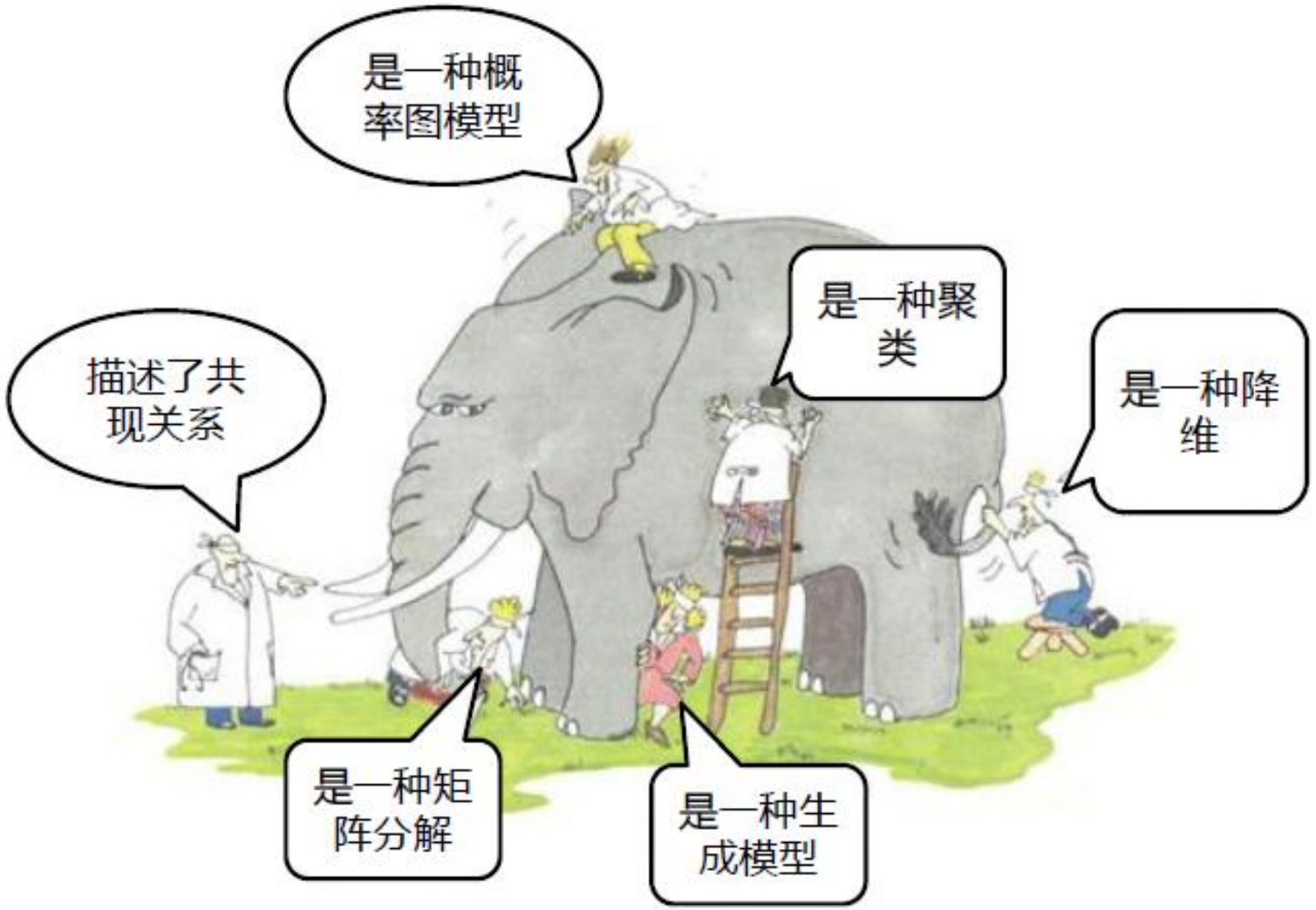
特征集	一级行业			二级行业		
	测试样本数	准确率	召回率	测试样本数	准确率	召回率
BOW(bag of words)	12987	82.33%	80.14%	12454	79.96%	79.96%
BOW+ peacock topics	12987	**86.82%**	**84.18%**	12454	**83.05%**	79.20%

- **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%↑**

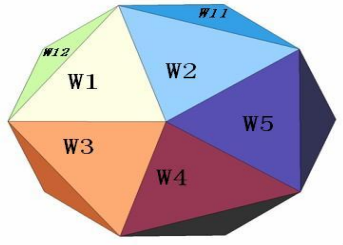
Thanks for your attentions!

LDA Topic Modeling



PLSA Text Modeling (Hofmann, 1999)

- 上帝有 K 个不同的topic-word 骰子
  每个骰子V 个面，每个面对应不同的 word

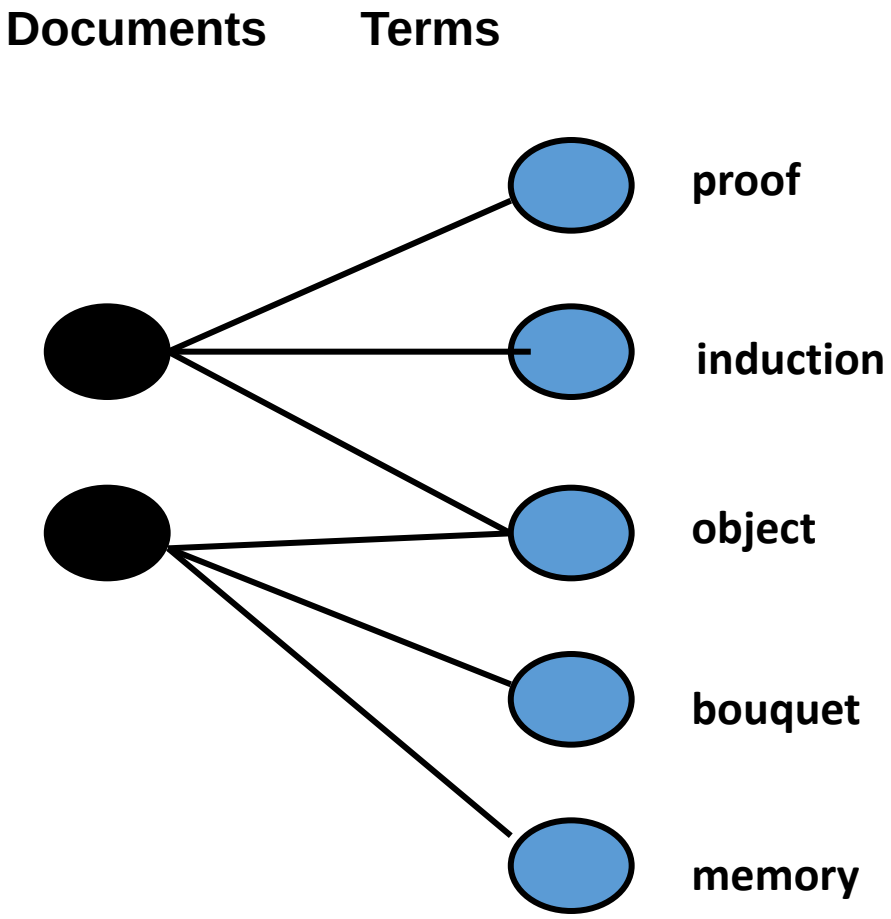


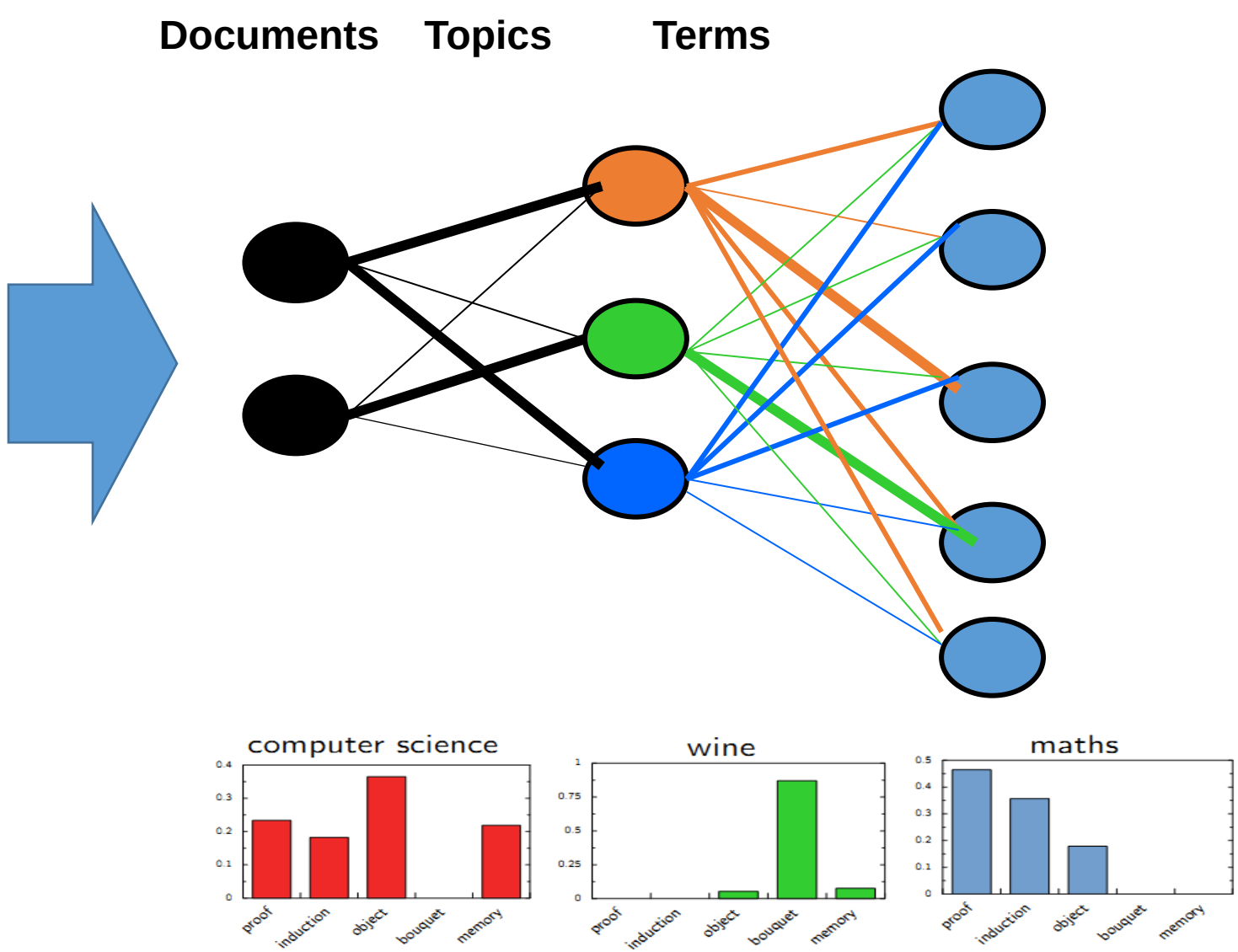
topic-word

- 对每篇 doc, 上帝都有一个doc-topic 骰子
  每个骰子 K 个面，每个面对应一个 topic 编号

doc-topic

Latent Topics





LDA Text Modeling