促进云计算创新发展 培育信息产业新业态
The 7th China Cloud Computing Conference
第七届中国云计算大会

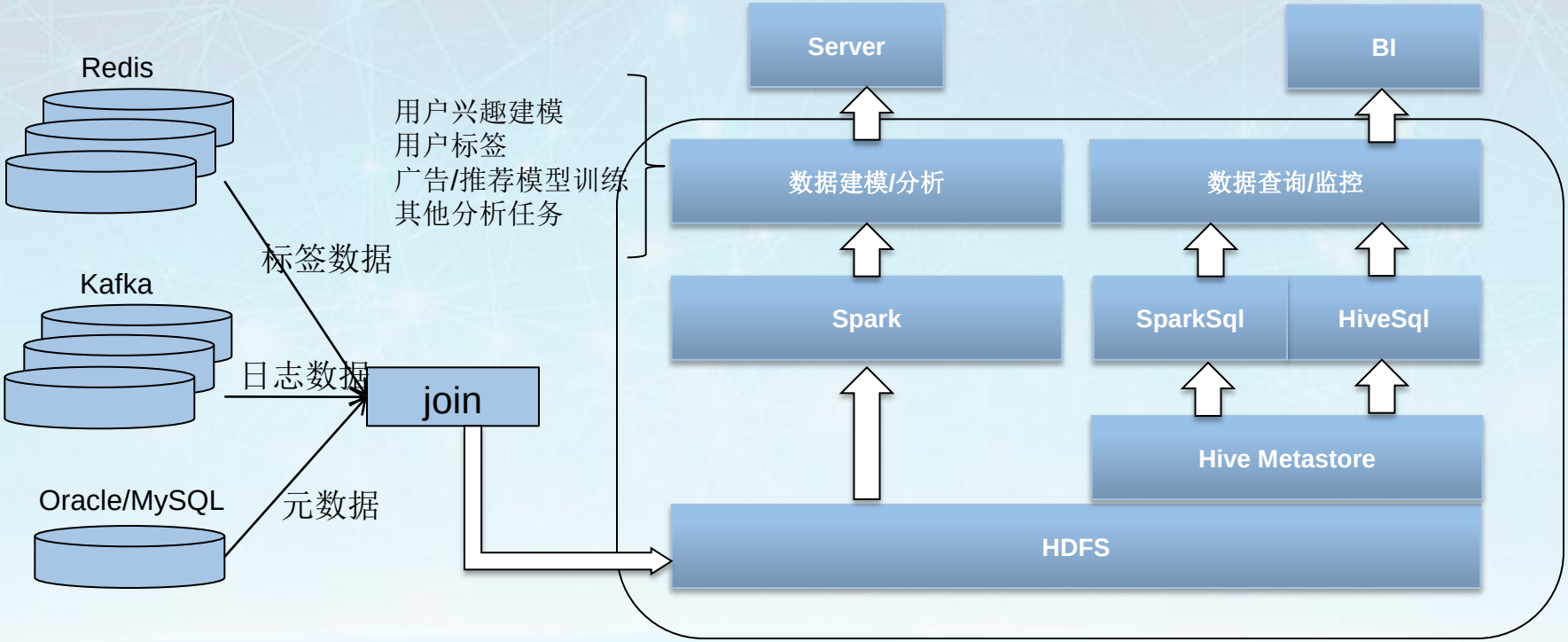
推荐算法和Spark实现

搜狐大数据中心 – 李滔

广告 vs. 新闻推荐

- 相同点
  - 都可看做一个点击率(ctr)估计的任务
  - 特征描述：用户、商品、上下文三个维度
  - 点击率是动态变化的
- 不同点
  - 点击率: 推荐点击率是广告的10~100倍
  - 特征描述: 广告细粒度特征 vs. 推荐相对粗粒度

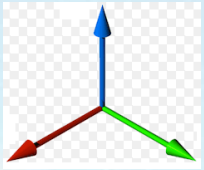
基于Hadoop EcoSystem的数据分析平台



CTR预估建模过程

上下文：
曝光时间、ip、useragent、广告栏位…

用户属性:
地域、人口属性（年龄、职业、性别）、**阅读兴趣**（体育、财经、政治…）、长期兴趣和短期兴趣、历史点击率…



item属性:
广告主、广告类别、关键词、历史点击率…
cms类别、主题、关键词、媒体…

日志处理 → 特征抽取 → 建模

Logistic Regression
Factorization Machines
GBDT
Radom Forest
FTRL
OBPR
…
另外需要考虑对新的item的探索问题(E&E)

用户兴趣建模

根据用户的阅读历史对用户兴趣建模

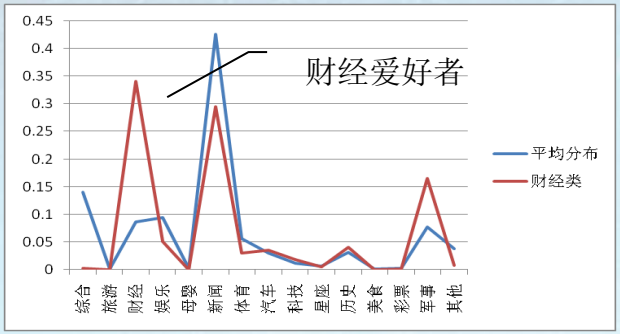
用户的新闻阅读分布式有偏的：
-- 热点新闻的巨大点击量
-- 新闻的曝光是有偏的

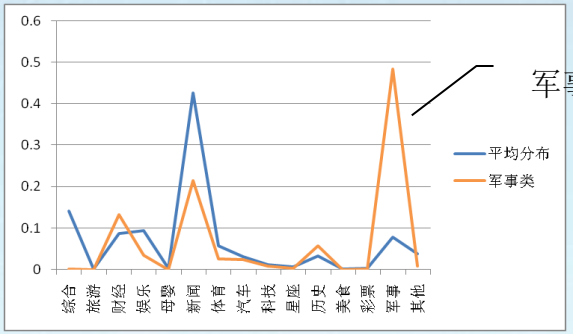
用户标签只能建立在用户相对于平均分布的偏离度上

手机搜狐网一个月的新闻点击分布

用户兴趣建模

用户兴趣是相对于整体用户阅读平均分布的偏离度的度量，常用的维度包括CMS分类，topic model，关键词等:





用户兴趣分为长期兴趣和短期兴趣；

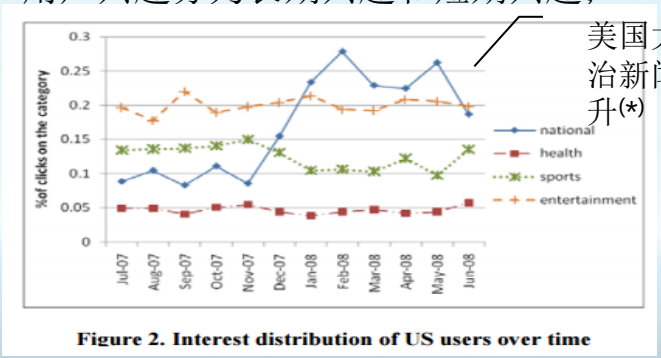
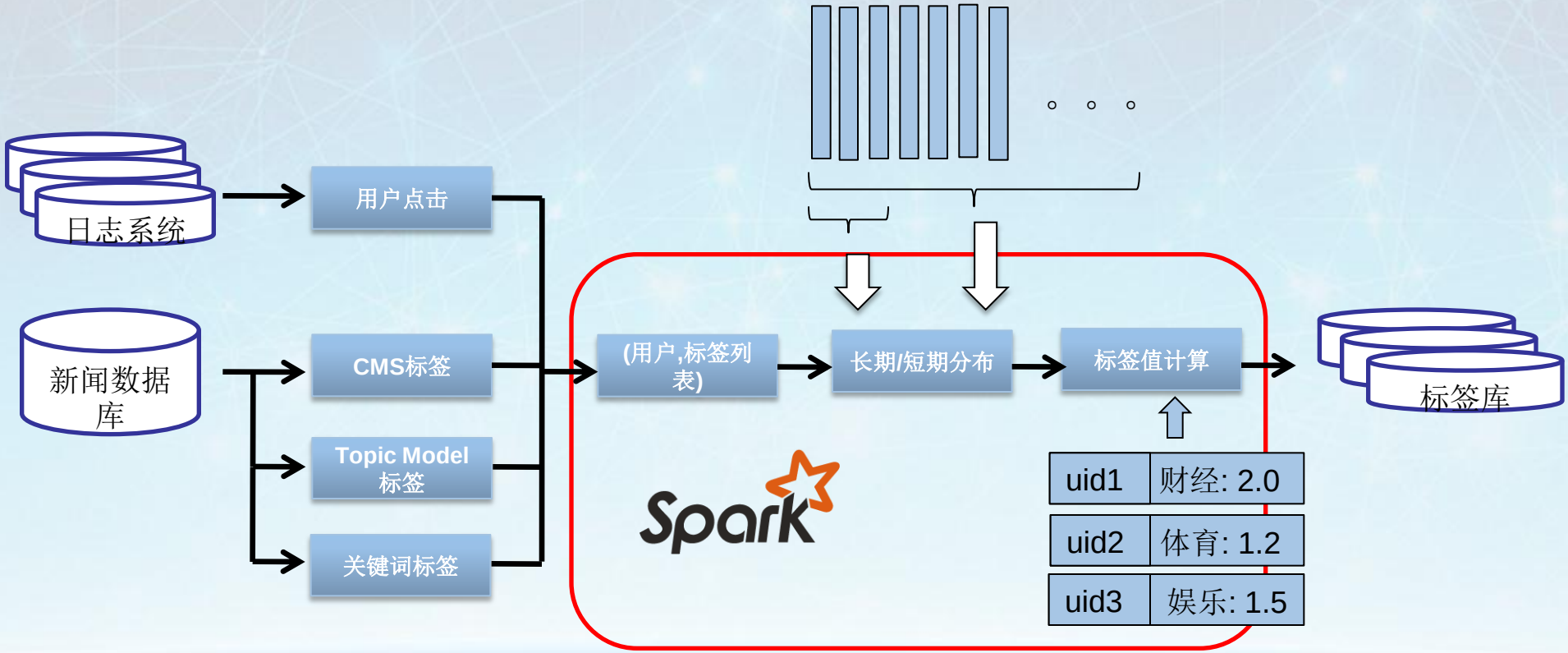


Figure 2. Interest distribution of US users over time

*Ref: J. Liu, Personalized News Recommendation Based on Click Behavior, 2010 International Conference on Intelligent User Interfaces

用户兴趣建模流程


日志系统
新闻数据库
用户点击
CMS标签
Topic Model 标签
关键词标签
(用户,标签列表)
长期/短期分布
标签值计算
Spark
uid1
财经: 2.0
uid2
体育: 1.2
uid3
娱乐: 1.5
标签库

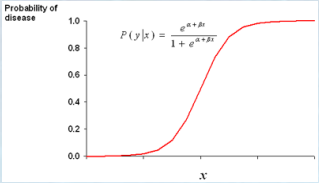
Spark带来的改善

- Spark丰富的数据处理操作
  - Map, reduce, filter, join, cogroup, …
- 处理时间相对于Hadoop缩短近10倍
- 代码量大为减少

广告和新闻推荐建模

- 多信息源的有效利用
  - 支持多维度的特征及特征组合
  - 避免过拟合 (Overfitting)
- 探索(Exploration)和利用 (Exploitation)之间的平衡
  - Exploration：获取信息；Exploitation：根据当前信息决策
  - Bandit方法: ε-Greedy, UCB
  - Bayesian方法: Thompson Sampling
- 在线训练

Logistic Regression



Logit函数

点击的概率P由特征的线性加权决定:

$P(y|x)=logit(w_1x_1+w_2x_2+\ldots+w_nx_n)$

常用的特征组合:
用户标签和item标签组合;
曝光时间和item标签组合;
用户地域和item标签组合;
。 。 。

特征组合

新闻标签=体育

用户标签=体育

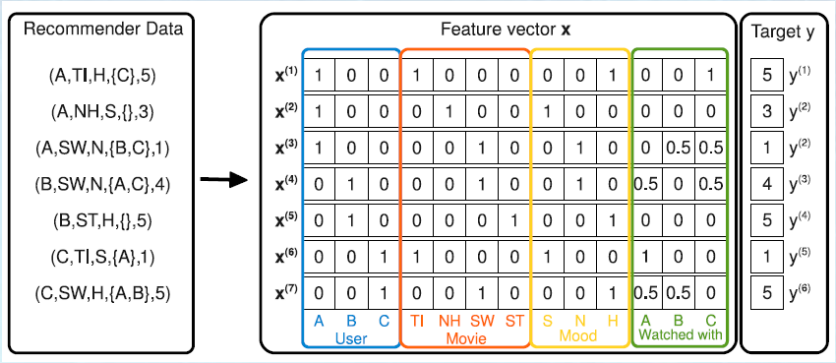
	x_1	x_2	x_3	x_4		x_n
x_1	w_{11}	w_{12}	w_{13}	w_{14}	…	w_{1n}
x_2	w_{21}	w_{22}	w_{23}	…	…	w_{2n}
x_3	w_{31}	w_{32}	w_{33}	…	…	w_{3n}
x_4	w_{41}	…	…	…	…	…
	…	…	…	…	…	…
x_n	w_{n1}	w_{n2}	w_{n3}	w_{n4}	…	w_{nn}

更好的模型: Factorization Machines

$$\hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

	x_1	x_2	x_3	x_4		x_n
x_1	$\langle v_1 \cdot v_1 \rangle$	$\langle v_1 \cdot v_2 \rangle$	$\langle v_1 \cdot v_3 \rangle$	$\langle v_1 \cdot v_4 \rangle$	...	$\langle v_1 \cdot v_n \rangle$
x_2	$\langle v_2 \cdot v_1 \rangle$	$\langle v_2 \cdot v_2 \rangle$	$\langle v_2 \cdot v_3 \rangle$	...	...	$\langle v_2 \cdot v_n \rangle$
x_3	$\langle v_3 \cdot v_1 \rangle$	$\langle v_3 \cdot v_2 \rangle$	$\langle v_3 \cdot v_3 \rangle$	...	...	$\langle v_3 \cdot v_n \rangle$
x_4	$\langle v_4 \cdot v_1 \rangle$	...	...	...	...	...
	...	...	...	...	...	...
x_n	$\langle v_n \cdot v_1 \rangle$	$\langle v_n \cdot v_2 \rangle$	$\langle v_n \cdot v_3 \rangle$	$\langle v_n \cdot v_4 \rangle$	...	$\langle v_n \cdot v_n \rangle$

每维特征对应一个p维权重向量，参数数量为n*(p+1)+1



优点:
General purpose模型，覆盖率像MF, SVD++等多种传统模型；
方便加入多种特征；
优化过程简单

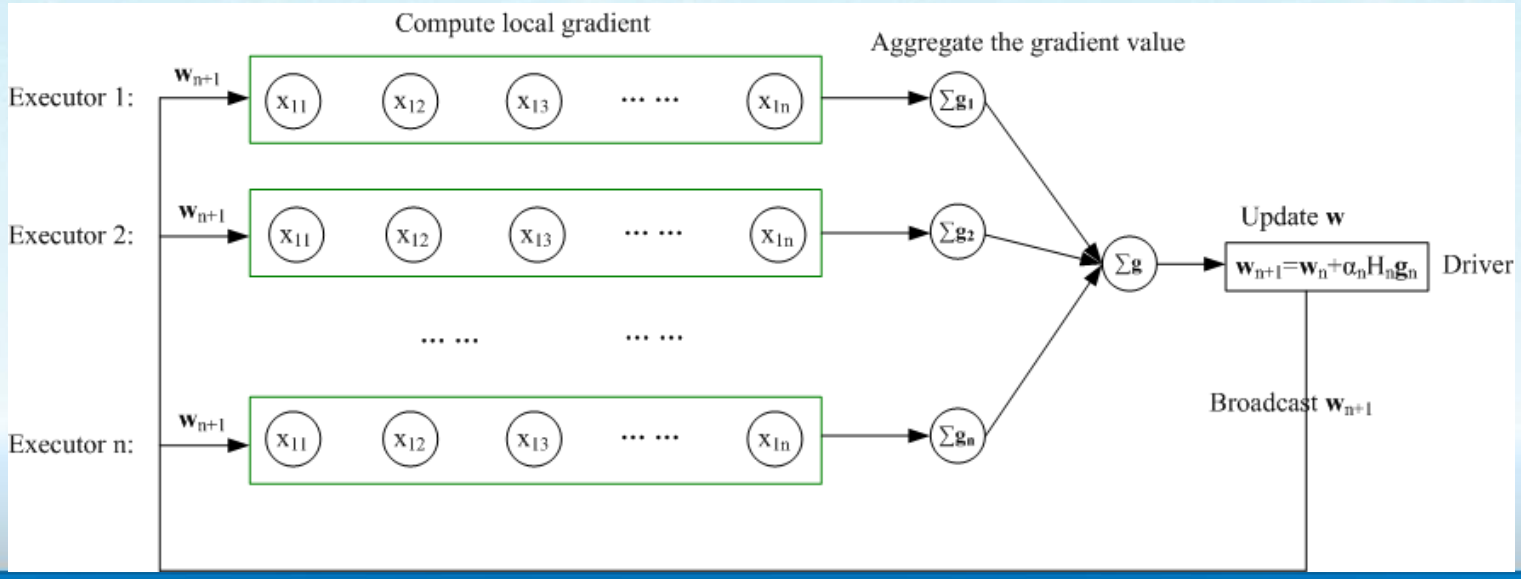
Steffen Rendle, Factorization Machines, in Proceedings of the 10th IEEE International Conference on Data Mining

Factorization Machines – 基于Spark的实现

实现了一个基于OWLQN的算法，以期得到稀疏解：

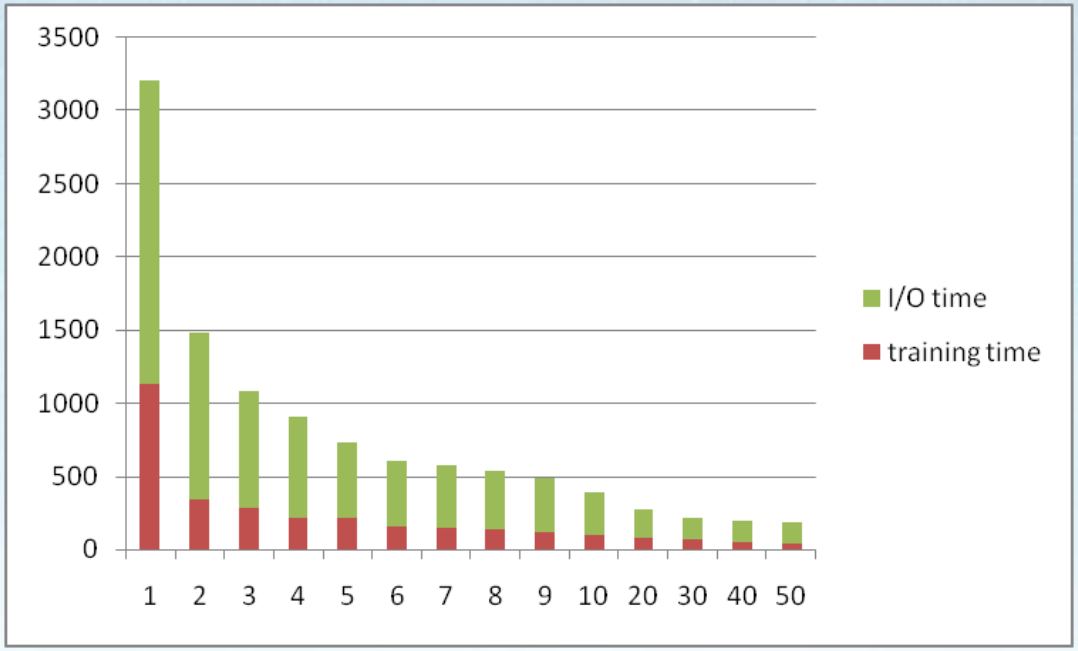
$$\hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j + \lambda \cdot \left(\sum_{i=0}^{n} |w_i| + \sum_{i=0}^{n} |\mathbf{v}_i| \right)$$

Batch Gradient的算法非常适合在Spark上实现



Factorization Machines – 基于Spark的实现

对一个8GB的数据集（约40M样本），比较并行度和处理时间的关系。(设置样本RDD的partition数量为executor数量的3倍)



Executor 数量和处理时间(s)的关系

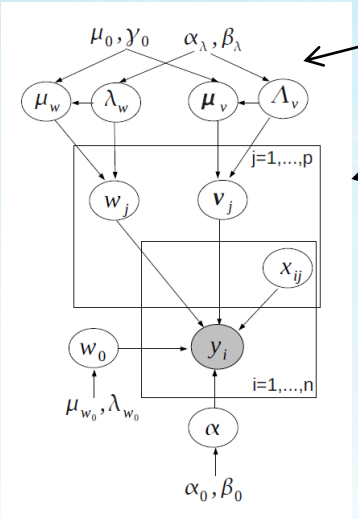
Bayesian Factorization Machines

- 为何要Bayes化:
  - 通过Bayes模型我们可以得到一个ctr的分布而非仅仅一个估计值
  - 有利于探索和利用问题:
  - 有利于控制投放速率

Bayesian Factorization Machines – 原理

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j \quad (1)$$

Graph Model



μ,λ的先验分布为Gamma分布

假设w,v服从N(μ,λ)的高斯分布

预测输出:

1) Sampling w,v according to:
   $w \sim N(\mu_w, \lambda_w)$
   $v \sim N(\mu_v, \lambda_v)$
2) Compute y based on equation (1)

这样我们得到的w,v是一个分布而非固定值，由此能得到预测值y的分布

C Freudenthaler, Bayesian factorization machines, 2011

Bayesian Factorization Machines – 基于Gibbs采样求解

设定参数 $\Theta_0 = \{\gamma_0, \mu_0, \alpha_0, \beta_0, \alpha_\lambda, \beta_\lambda\}$

初始化 $\theta \sim \mathcal{N}(0, \sigma)$

以指定迭代次数重复下列步骤:

Step1: Update μ_θ

$$p(\mu_\theta | \mathbf{y}, \mathbf{X}, \Theta, \Theta_H \setminus \{\mu_\theta\}, \Theta_0) = \mathcal{N}(\mu_{\mu_\theta}, \sigma^2_{\mu_\theta})$$

其中 $\mu_{\mu_\theta} = \sigma^2_{\mu_\theta} \lambda_\theta \left(\sum_{j=1}^{p} \theta_j + \gamma_0 \mu_0 \right)$, $\sigma^2_{\mu_\theta} = ((p + \gamma_0)\lambda_\theta)^{-1}$

Step2: Update λ_θ

$$p(\lambda_\theta | \mathbf{y}, \mathbf{X}, \Theta, \Theta_H \setminus \{\lambda_\theta\}, \Theta_0) = \Gamma(\alpha_\theta, \beta_\theta)$$

其中 $\alpha_\theta = (\alpha_\lambda + p + 1)/2$, $\beta_\theta = \left(\sum_{j=1}^{p} (\theta_j - \mu_\theta)^2 + \gamma_0 (\mu_\theta - \mu_0)^2 + \beta_\lambda \right) /2$

Step3: Update α

$$p(\alpha | \mathbf{y}, \mathbf{X}, \Theta, \Theta_H \setminus \{\alpha\}, \Theta_0) = \Gamma(\alpha_n, \beta_n)$$

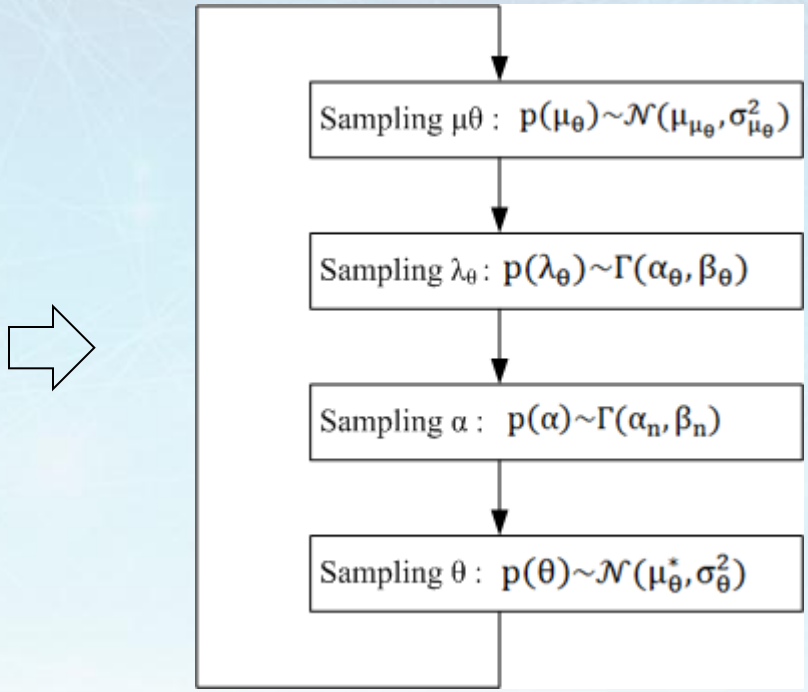
其中 $\alpha_n = (\alpha_0 + n)/2$, $\beta_n = \left(\sum_{i=1}^{n} (y_i - y(\mathbf{x}_i, \Theta))^2 + \beta_0 \right) /2$

Step4: Update θ

$$p(\theta | \mathbf{y}, \mathbf{X}, \Theta \setminus \{\theta\}, \Theta_H, \Theta_0) = \mathcal{N}(\mu^*_\theta, \sigma^2_\theta)$$

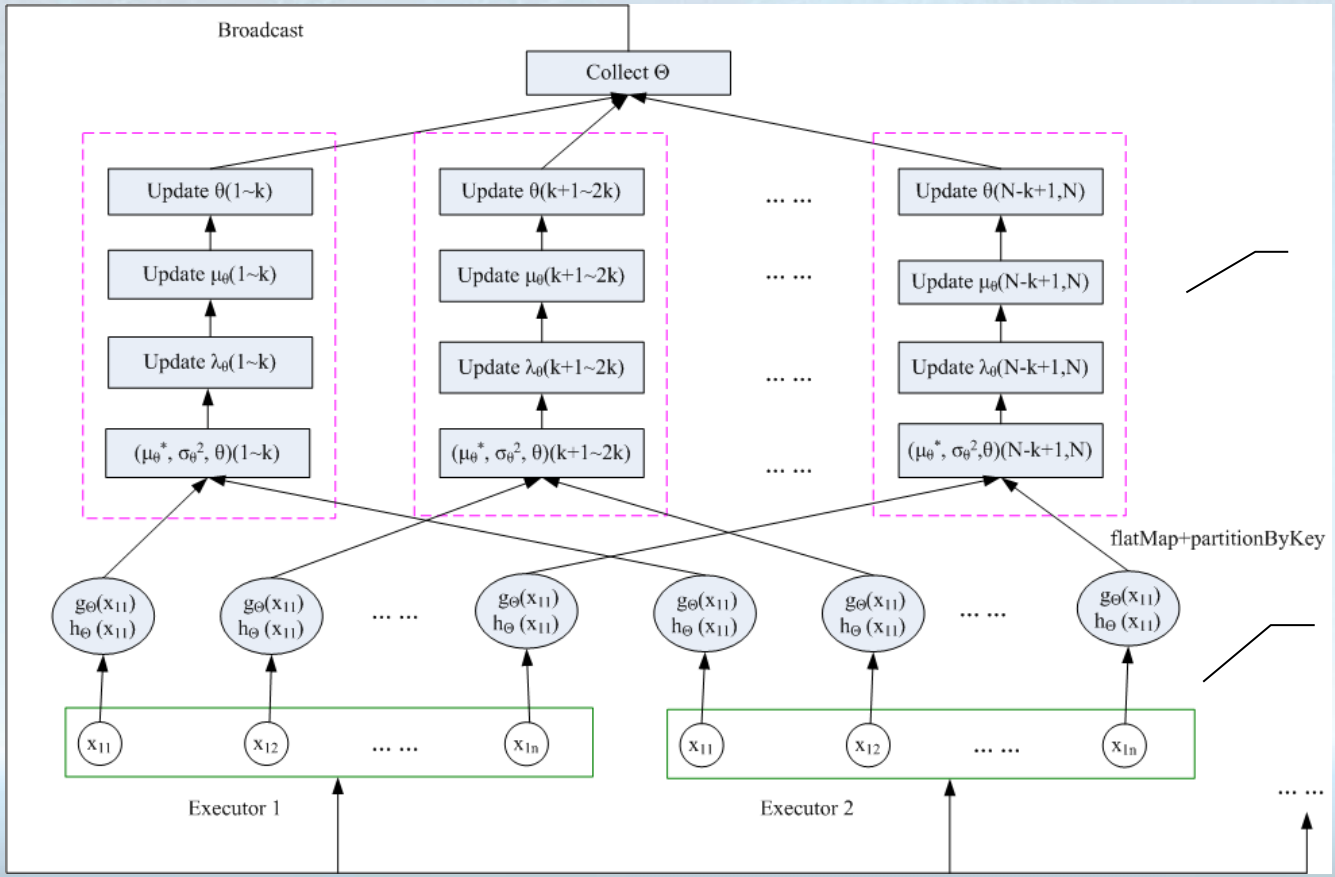
其中 $\mu^*_\theta = \sigma^2_\theta \left(\sum_{i=1}^{n} \alpha (y_i - g_\theta(\mathbf{x}_i)) h_\theta(\mathbf{x}_i) + \mu_\theta \lambda_\theta \right)$, $\sigma^2_\theta = \left(\alpha \sum_{i=1}^{n} h_\theta(\mathbf{x}_i)^2 + \lambda_\theta \right)^{-1}$

基于Gibbs采样的算法步骤



扫描整个数据集

Bayesian Factorization Machines – 基于Spark实现



每个executor更新一个参数子集

分布式计算预测值:

$$y(\mathrm{x}|\Theta) = g_\theta(\mathrm{x}) + \theta\, h_\theta(\mathrm{x})$$

时间复杂度: $O(kN_z)$

Summary

- 广告和推荐中的ctr估计是机器学习的典型应用；
  - 模型能有效地表示多样性的特征
  - 预测对象的动态变化需要模型有探索(Exploration)能力
- Spark可以作为ctr建模的有效的计算平台
  - 用户标签处理需要做大量日志数据的整合关联
  - ctr建模算法往往都是迭代运算

谢谢！