个性化推荐系统的几个问题

2012.12.22 北京

演讲者

胖子@豆瓣　http://www.douban.com/people/1000037/

随缘放旷，任性逍遥，但尽凡心，别无胜解

- 工学学士、工学硕士，清华大学
- 3年时间，供应链管理研究、咨询
- 2年时间，商学院撰写运营管理案例
- 7年在豆瓣，算法工程师

两种基本算法再比较

User/Item Based CF

基本假设

归一化的评分矩阵 $U_{m\times n}$

假设其非零元素个数为 C $C \ll m \times n$

假设非零元素个数为均匀分布

平均每行为 $L = C / m$ 平均每列为 $L^{'} = C / n$

$$\max(L, L^{'}) \ll \min(m, n)$$

相似矩阵计算复杂度

$$\left|S_{m\times m}\right| = m^2\left[1-\left(1-\frac{L^2}{n^2}\right)^n\right] \qquad \left|S'_{n\times n}\right| = n^2\left[1-\left(1-\frac{L'^2}{m^2}\right)^m\right]$$

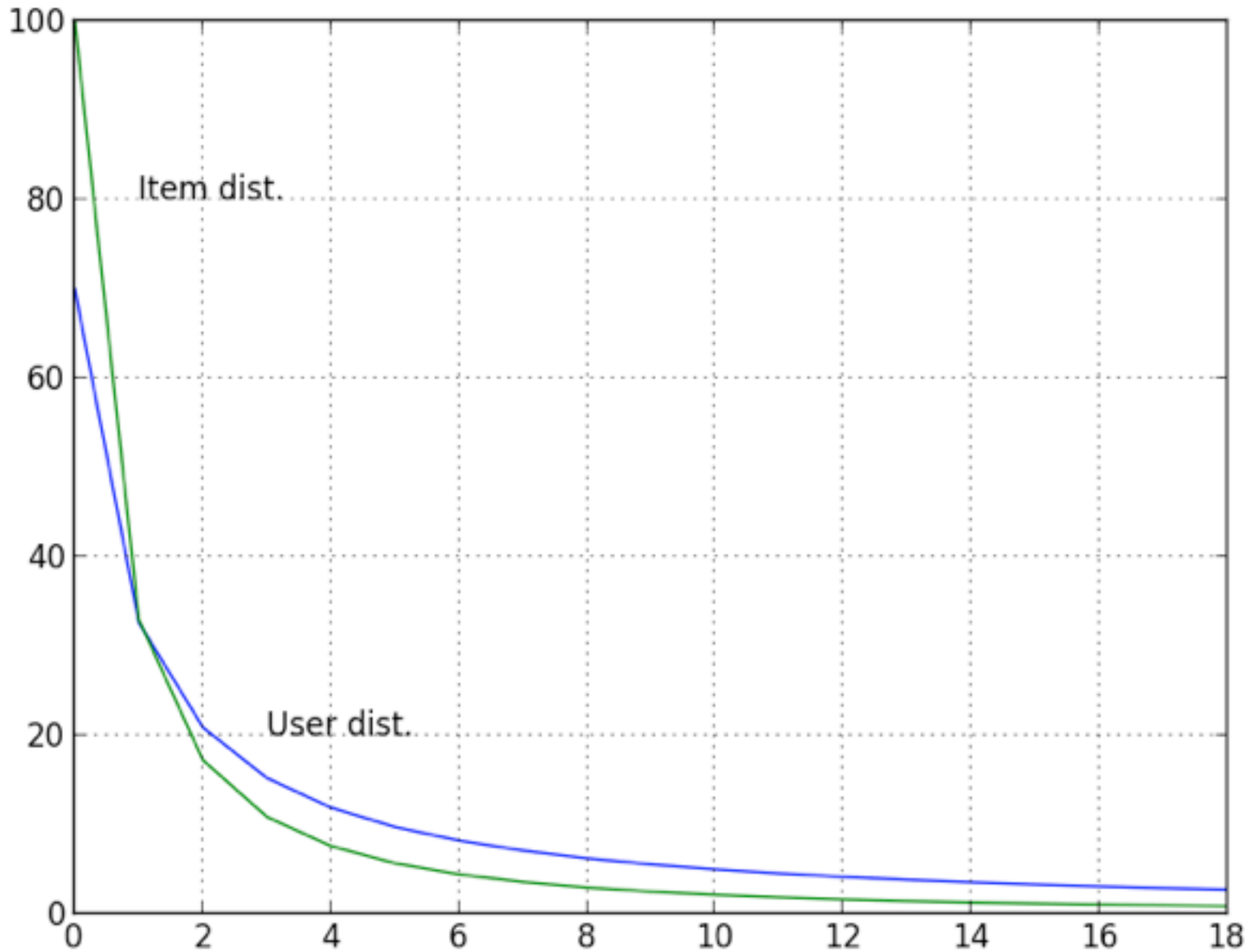
在不是非常严格的意义下，经展开忽略高阶项化简后，有:

$$\left|S_{m\times m}\right| \approx \frac{m^2L^2}{n} = \frac{C^2}{n}, \qquad \left|S'_{n\times n}\right| \approx \frac{n^2L'^2}{m} = \frac{C^2}{m}$$

讨论

- 近似线性的时间和空间复杂度
- 与非零元平方成正比，与用户和条目数成反比
- 增量更新

用户与条目收藏的幂律分布



实际的计算复杂度

$$\sum_{i=i}^{m} L_i(L_i - 1) \geq mL^2 \qquad \sum_{i=i}^{n} L_i^{'}(L_i^{'} - 1) \geq nL^{'2}$$

在幂律分布的假设下，相似度矩阵的计算量与收藏的分布直接相关，通常基于条目的CF要显著低于基于用户的CF算法

算法与产品的适配

- 不同的产品阶段
- 不同的用户群
- 不同的计算资源和框架

缺失值的处理

Missing Value

缺失值

- 收藏/评分矩阵非常稀疏
- 缺失代表什么?
- 怎样利用缺失值改善推荐?

缺失值作为负面反馈

- User-Oriented Negative Sampling
- Item-Oriented Negative Sampling

$$L(U,V) = \sum_{ij} W_{ij}(R_{ij} - U_{i.}V_{j.}^{T})^2 + \lambda(\|U\|_F^2 + \|V\|_F^2)$$

如何评价

- 负面反馈采样在给定数据集上能得到比较好的效果
- 改变了原有的信息结构
- 给用户带来了什么?

矩阵分解与生成模型

Matrix Factorization & Generative Model

矩阵分解

$$V = WH \qquad v_i = \sum_{j=1}^{N} h_{ji} w_j$$

$$F(W, H) = \|V - WH\|_F^2$$

不同损失函数可以引出不同的
矩阵分解形式和优化方法

生成模型

- 隐马模型，高斯混合模型
- 贝叶斯模型
- LDA，RBM

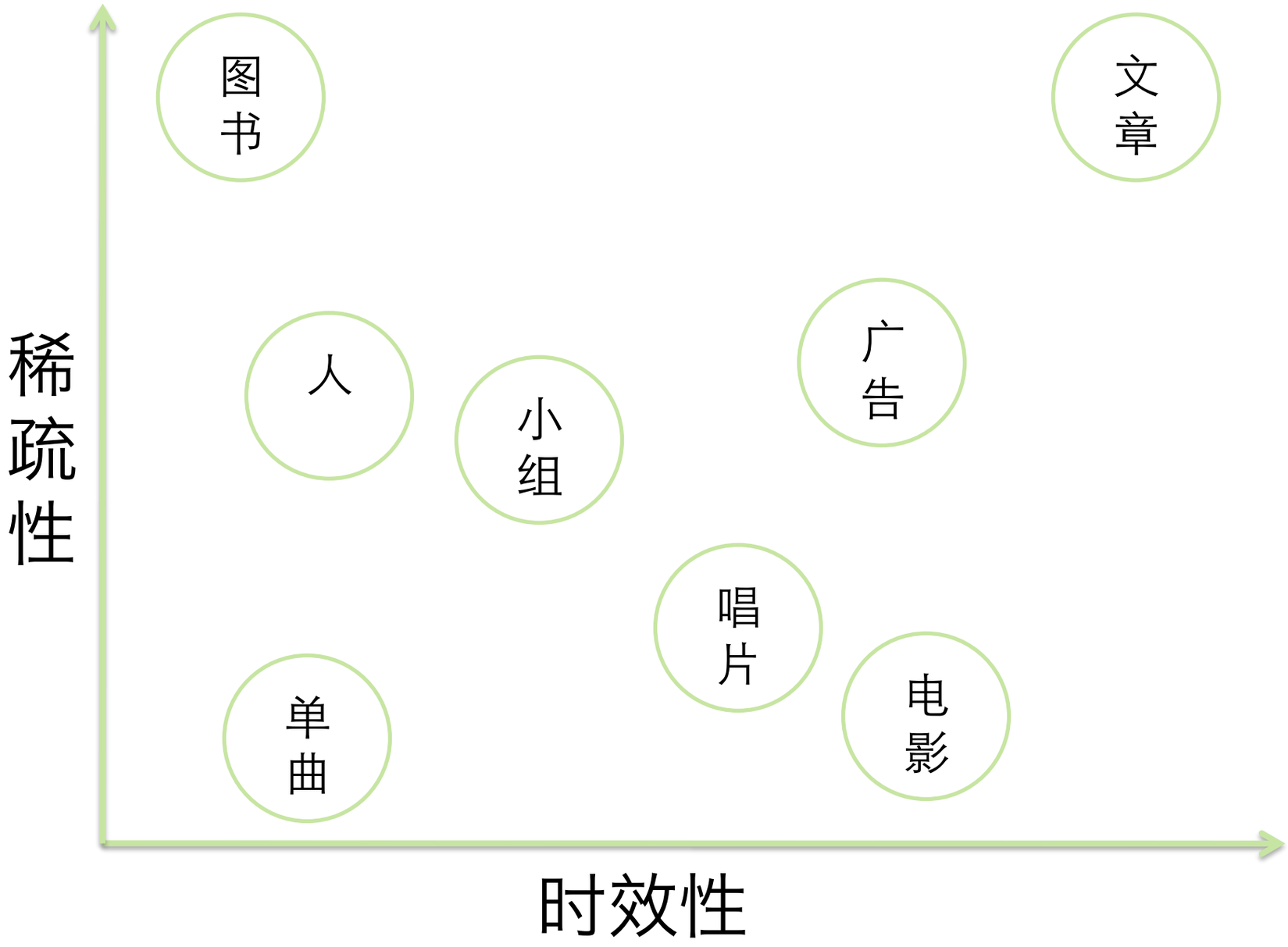
统一性

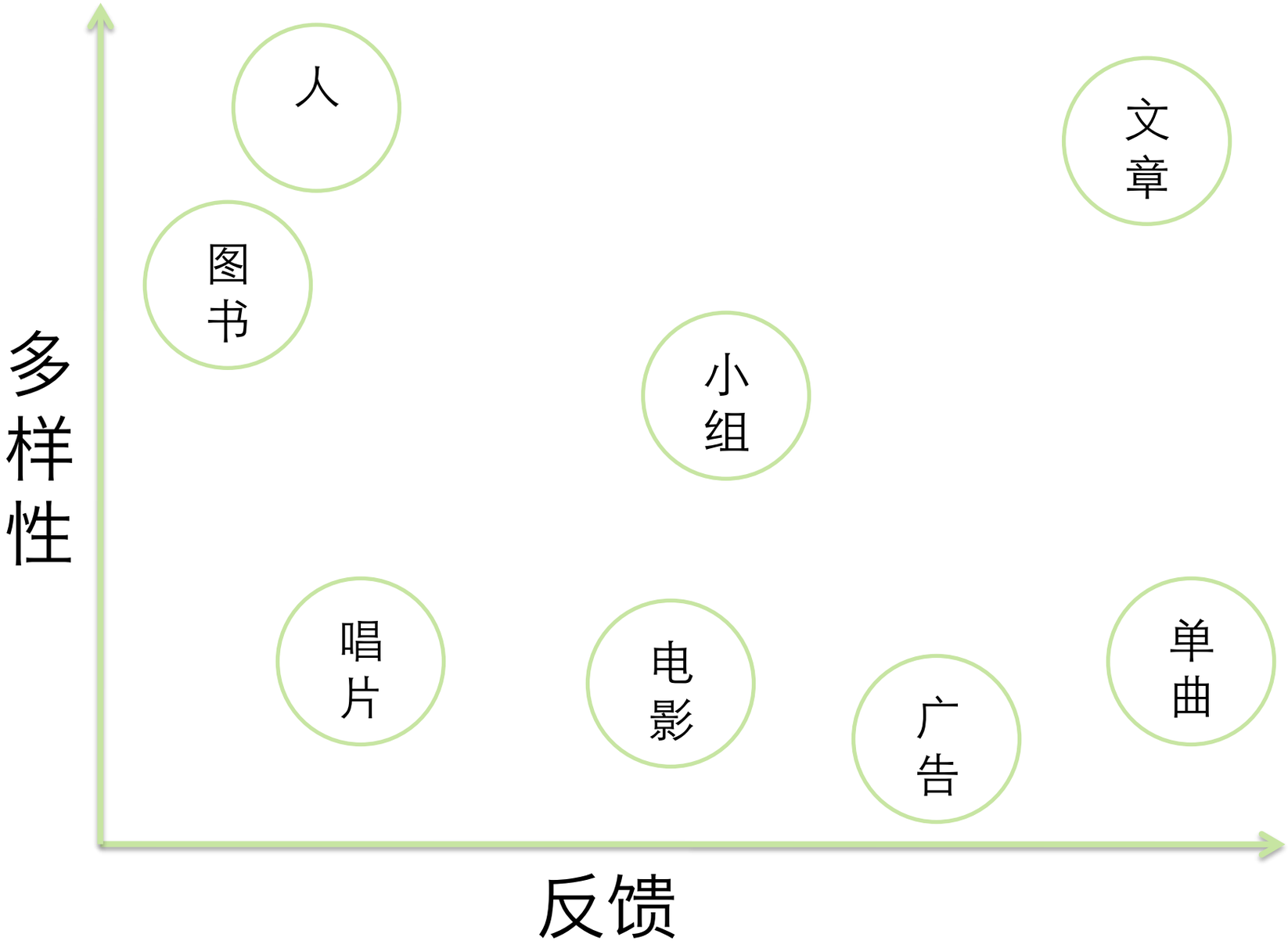
- 假设层面的统一
- 技术层面的统一

	冷启动	可解释性	惊喜	时效性	鲁棒性
协同过滤	★	★		★	★★
图模型		★	★		★
矩阵分解	★	★	★	★	★
Topic Model	★	★★		★	
增强学习	★	★		★★	
决策树		★	★	★	★
Boosting	★	★	★	★	★

算法与产品

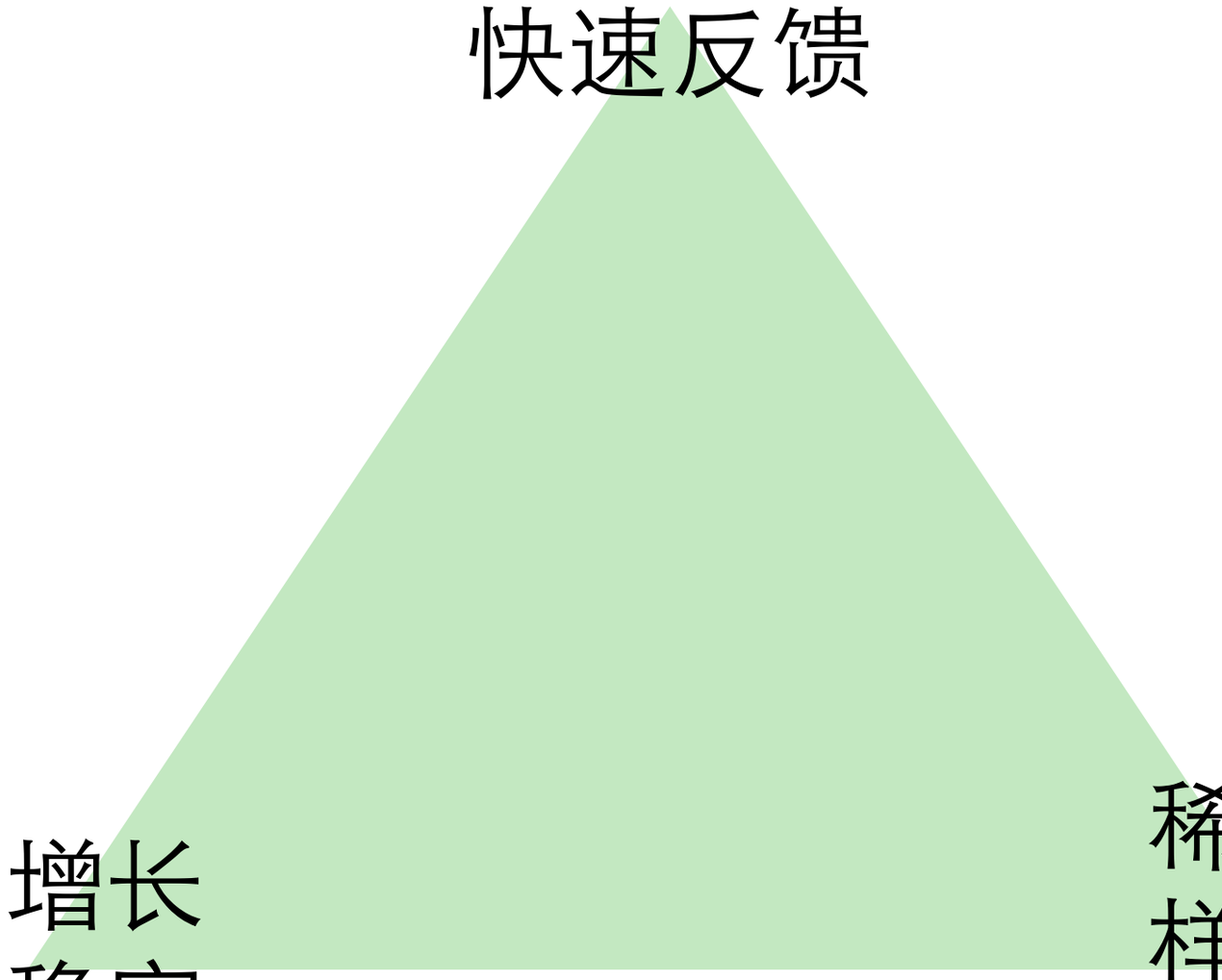
	用户数	条目数	稀疏性	多样性	时效性	反馈	推荐效果
图书	3,000,000	3,000,000	< 0.5%	高	低	慢	B
电影	5,000,000	100,000	1% ～5%	低	中	中	C
唱片	1,500,000	400,000	< 1%	低	低	中	C
小组	5,000,000	200,000	%1	中	中	中	B
人	5,000,000	5,000,000	< 0.5%	高	低	慢	D
文章	500,000	10,000,000	< 0.1%	高	高	快	C
单曲	5,000,000	1,000,000	5% ~ 10%	低	低	快	A
广告	30,000,000	50,000	1%	低	高	中	D


稀疏性
时效性
图书
人
小组
广告
文章
唱片
电影
单曲


人
图书
小组
文章
唱片
电影
广告
单曲
多样性
反馈

什么样的产品适合推荐?

能够获得
快速反馈



条目增长
相对稳定

稀疏性多
样性和时
效性平衡

其他

个性化推荐的历史

1992 ~ 2002

2002 ~ 2012

2012 ~

电子商务

新闻组

分类浏览

web 2.0

SNS

广告

兴趣网络

云计算

移动互联

网络融合

机器学习与人的学习

- 产品与人群
- 短期指标与长期指标
- 我们学到了什么

web面临的挑战

- 从自由与开放走向私有与封闭?
- 从第二人生走向第一人生
- 从信息经济走向体验经济

个性化推荐

- 前所未有的机会
- web 2.0, 云计算，成熟的技术准备
- 要么是平台，要么是平台的一部分

Algorithms should facilitate rather than replace social processes.

谢谢！

Q & A