社会化推荐

张叶银

Yeyin.zhang@renren-inc.com

推荐系统

- 给用户推荐他可能感兴趣的东西
  - 好友推荐，商品推荐，文档推荐，广告推荐

- Amazon, Netflix, Google, Facebook, Youtube……

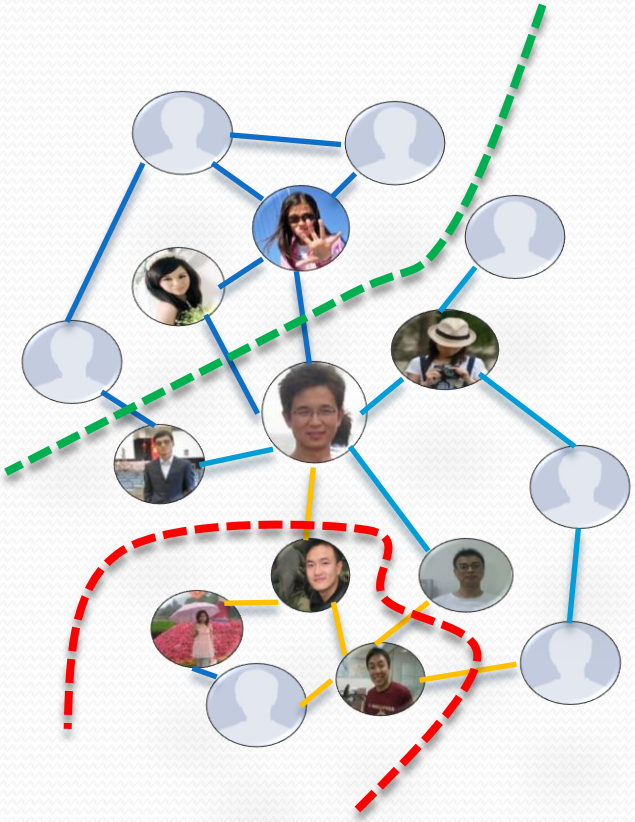
主流推荐算法

- 协同过滤
  - User-based, Item-based
- 内容过滤
  - 属性的相关性
- 社会网络
  - 图的方法

推荐系统评测

- 准确度
- 覆盖率
- 多样性
- 新颖性

Social Graph



好友推荐

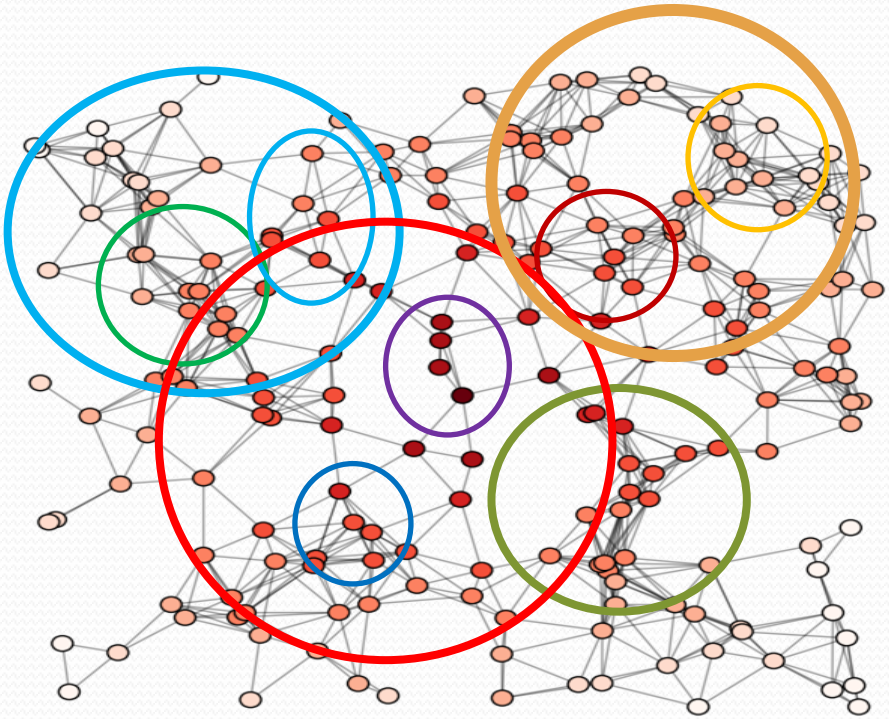
- 关键问题
  - 用户间的亲密度
    - 共同好友数目
    - 个人资料相似程度
    - 用户互动频度
    - 用户的兴趣

People you may know

- 二度好友Friends of friends

$$similarity(user1, user2) = |friendset1 \cap friendset2|$$

Circles



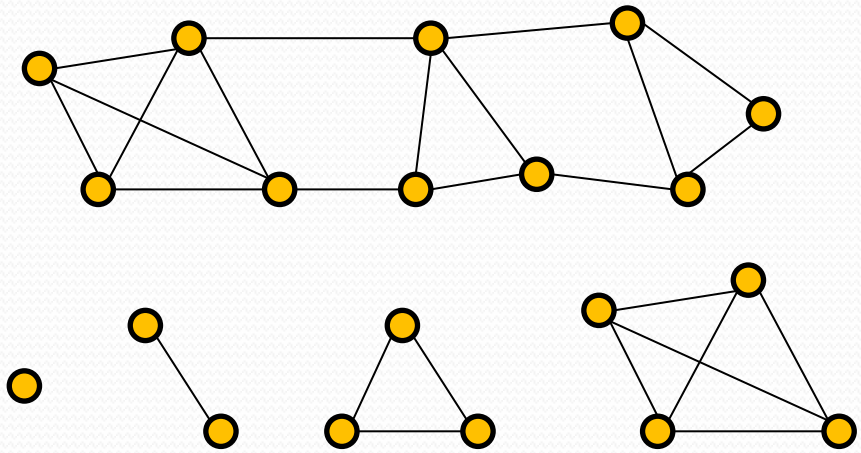
- 部门 »公司
- 班级 »学校
- 实验室 »班级
- …

好友簇

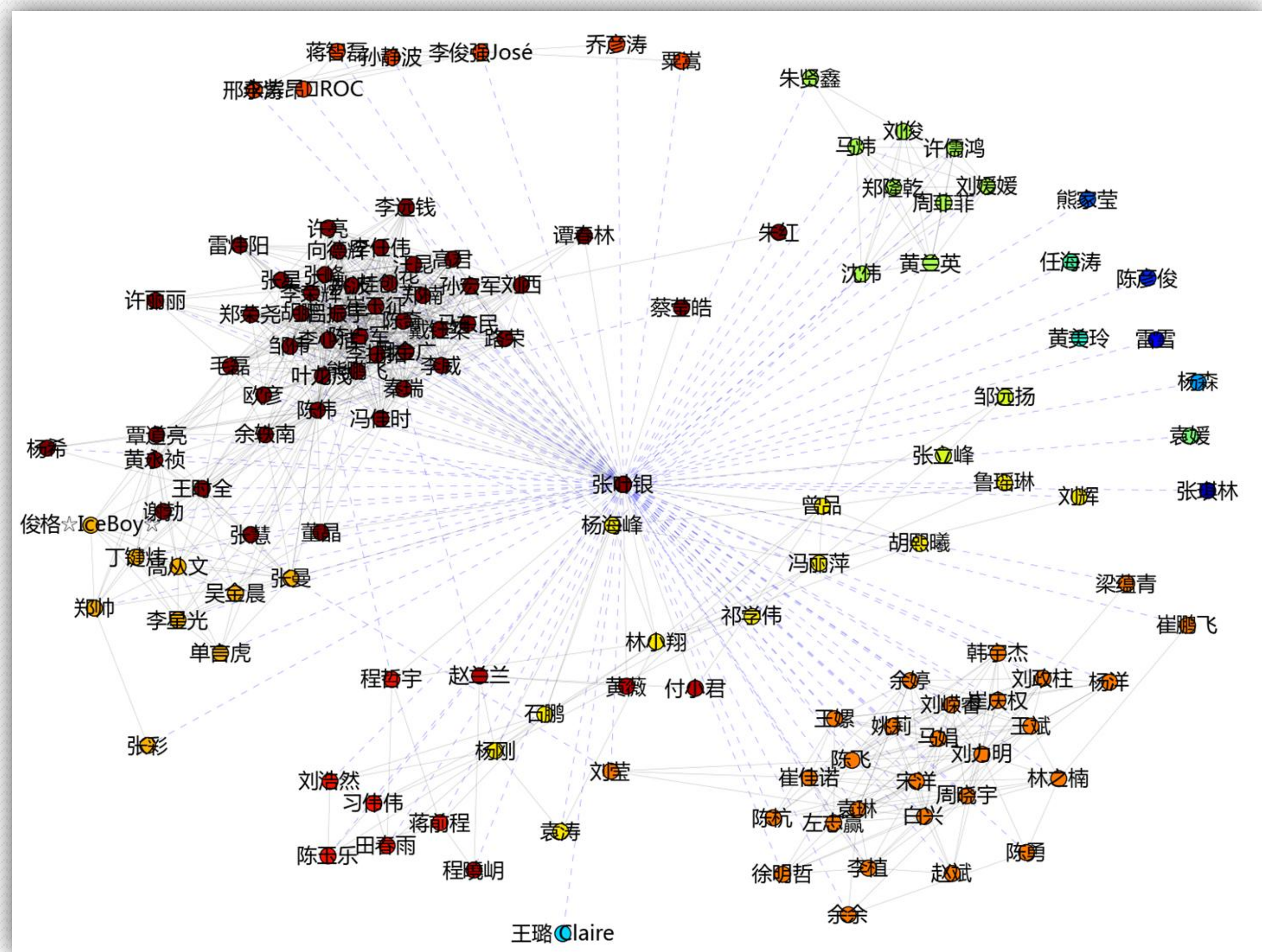
Strong ties → Weak ties

Strong ties

- Community Detection
  - 图的连通性
    - Clique:
    - 例如： $V' \subseteq V$, 对于任意的u, v $\in V', u \neq v, uv \in E$



圈子

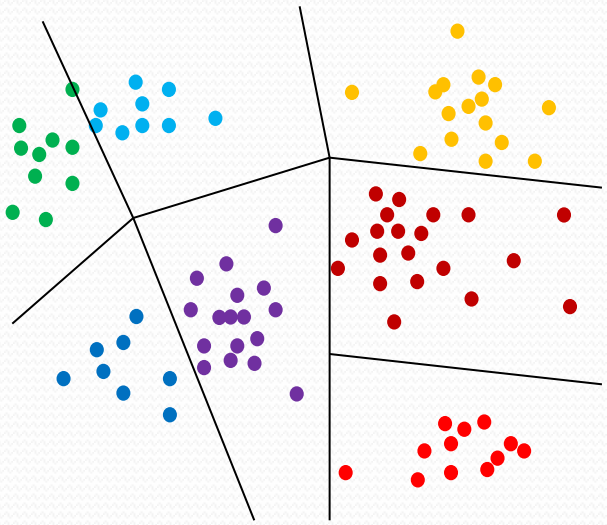


乔彦涛
刘俊
许儒鸿
郑隆乾
刘媛媛
任海涛
黄美玲
邹远扬
张立峰
刘辉
冯丽萍
崔鹏飞
祁学伟
林小翔
杨海峰
谭春林
俊格☆IceBoy☆
吴金晨
郑帅
张彩
黄薇
石鹏
杨刚
刘浩然
田春雨
袁涛
王璐Claire
林之楠
崔佳诺
周晓宇
徐明哲
赵斌

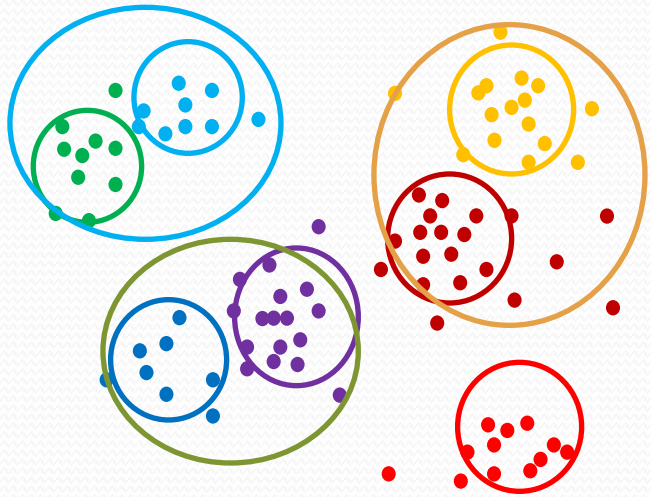
应用

- 推荐
- 隐私控制
- 新鲜事定制

To weak ties : unsupervised learning

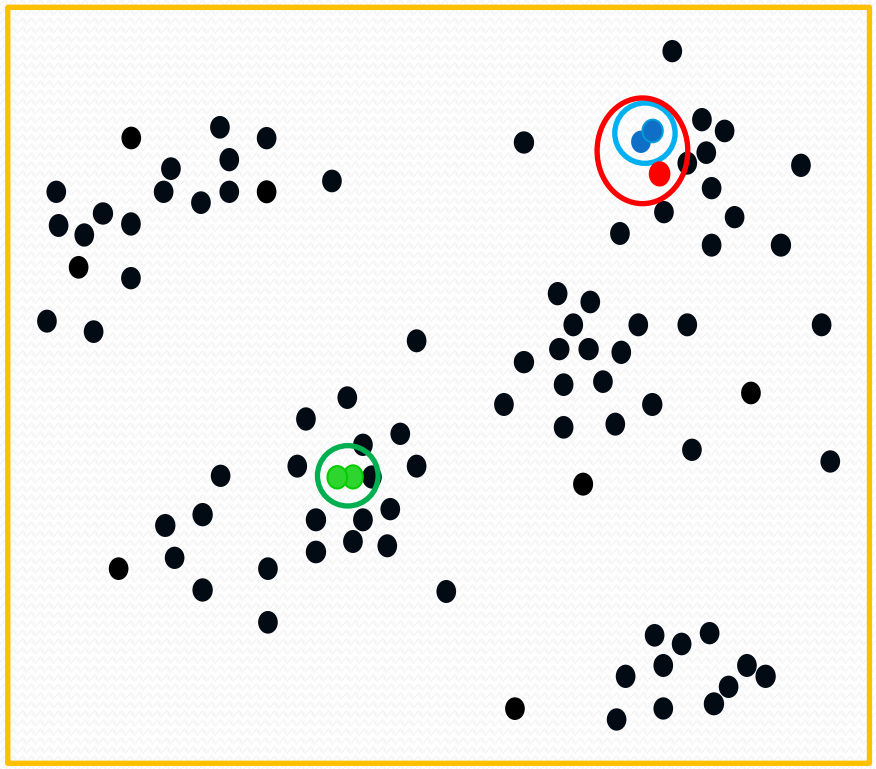


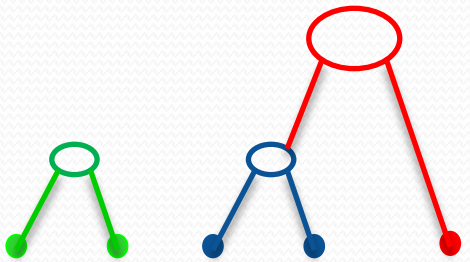
flat clustering



hierarchical clustering

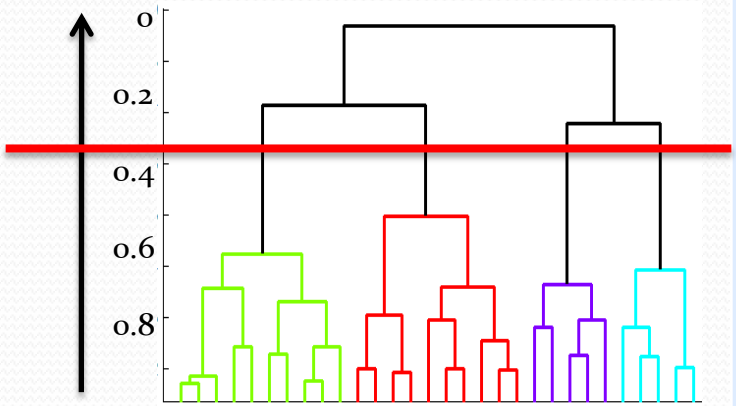
Hierarchical agglomerative Clustering





Hierarchical agglomerative clustering

Monotonic



Method: **Merge the nearest clusters until a single cluster is left**

Procedure HAC (N points, stop criterion)
{
(1) Initialize n points as n cluster centers;
(2) Iterate over centers until stop criterion is satisfied:

a. Compute pair-wise similarity between any two centers $sim(c_i, c_j)$

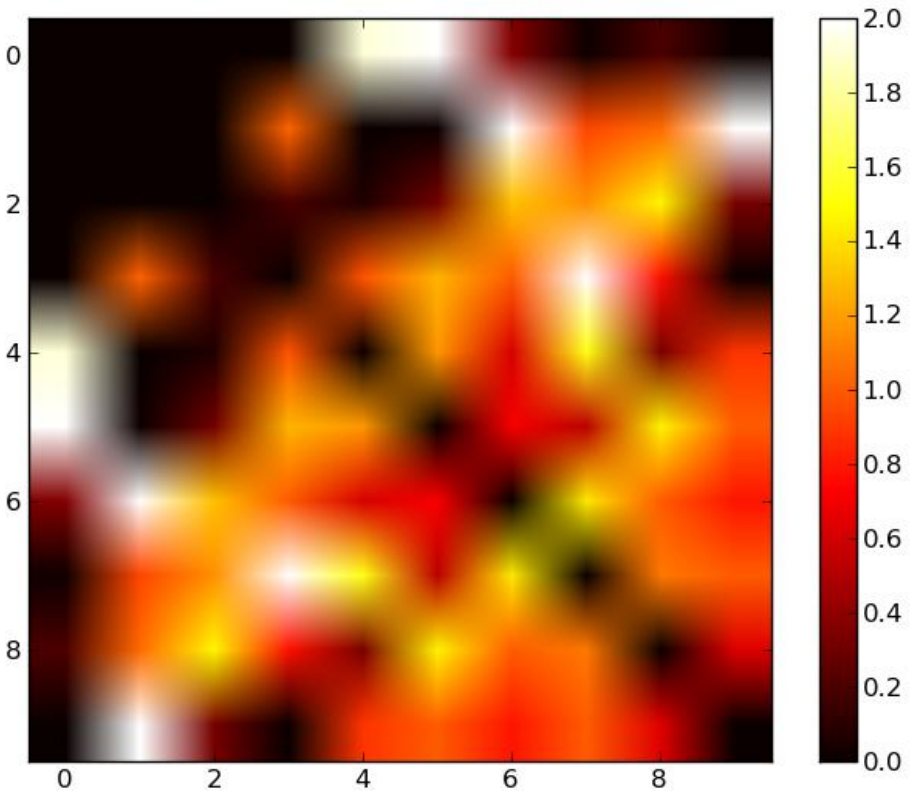
b. Find the nearest pair of centers

c. Merge the two centers

$$< i, j > \leftarrow \arg\max_{i,j} sim(c_i, c_j)$$

(3) Output the hierarchical clusters.
}

Pair-wise distances



Symmetric similarity matrix

To weak ties

- 相似性度量

$$similarity_{ij} = \left\| v_j \bigcap v_i \right\| \cdot \sum_k \sum_l R_{kl}$$

To weak ties

- 选择
  - 特征：
    - 用户交互：　$r_{ij} = \sum_{k \in A} a^k \cdot n_{ij}^k$
    - 用户关系：

$$c_i^o = \arg\max_{c_i^o} \{ \sum r_{jk} \mid c_i^o \in C_i^o, j,k \in v_i \}$$

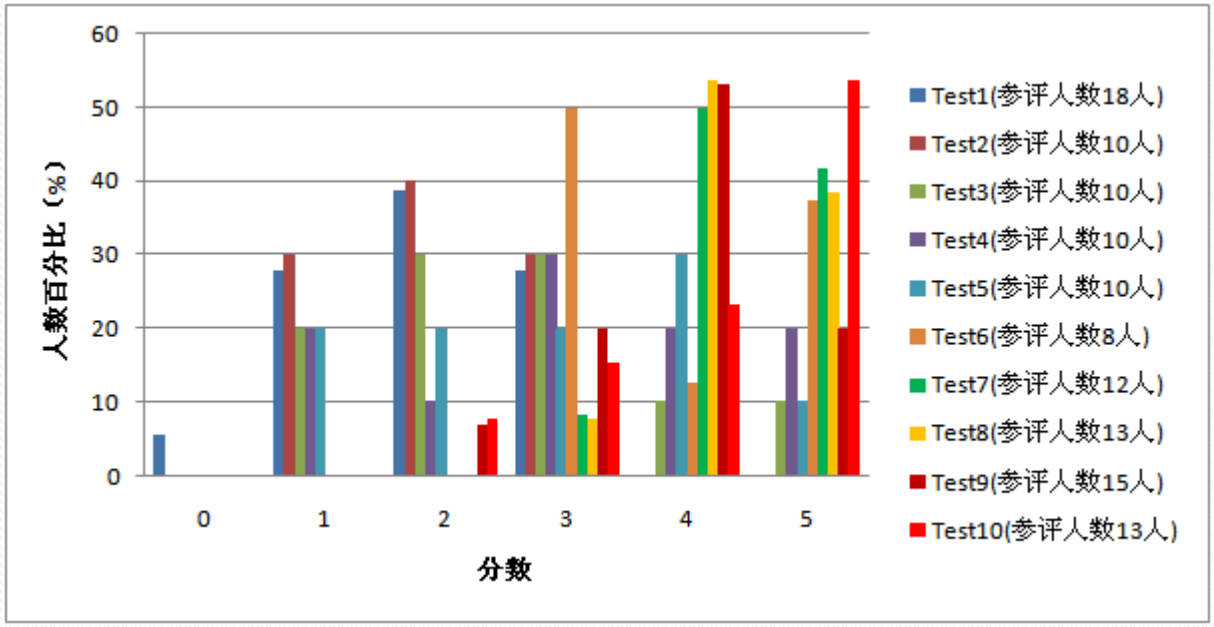
合并

$$o_i^r(U_i^r, C_i^r) = o_j^{r-1}(U_j^{r-1}, C_j^{r-1}) \bigcup o_k^{r-1}(U_k^{r-1}, C_k^{r-1}),$$

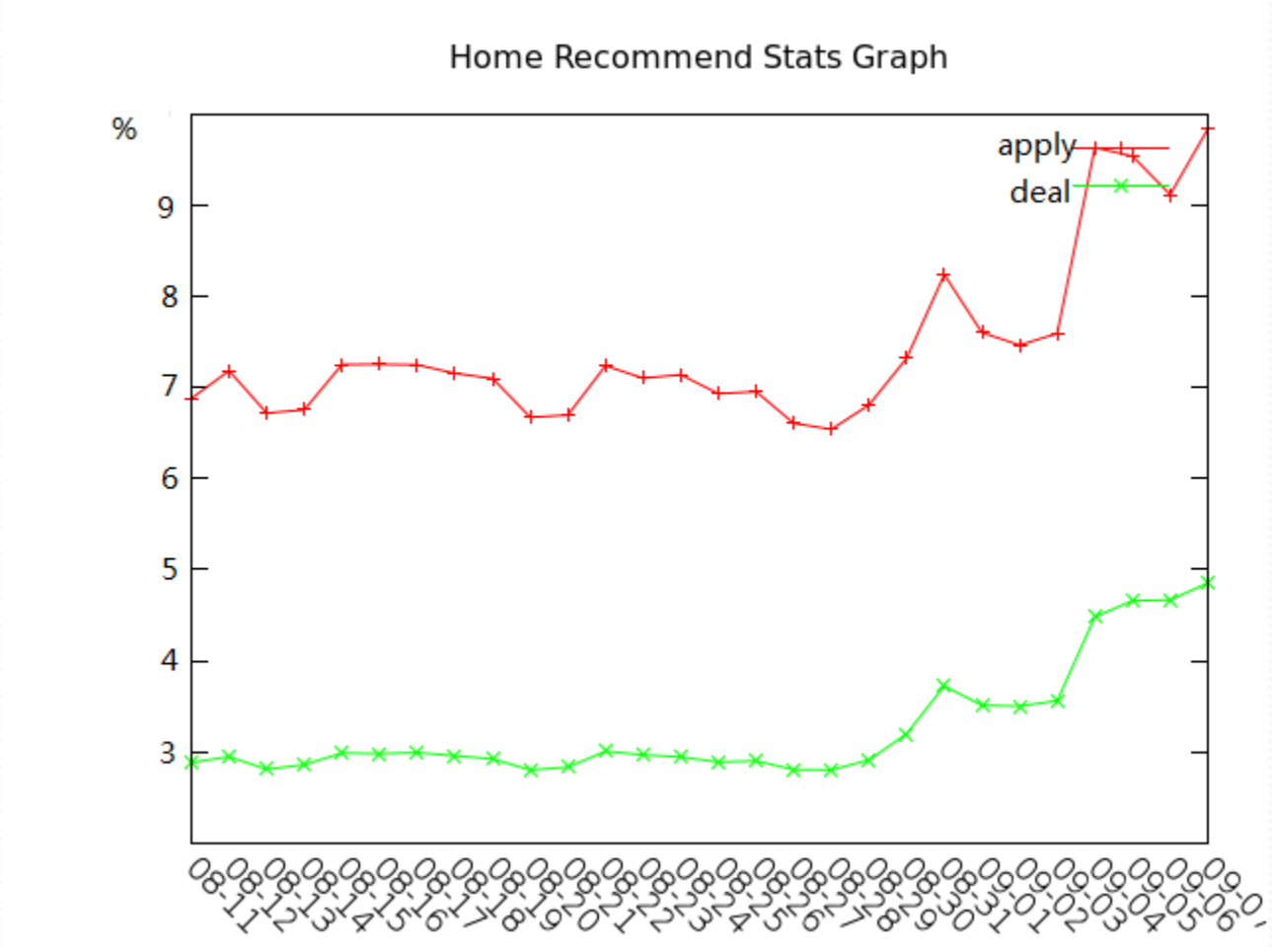
其中，$U_i^r = U_j^{r-1} \bigcup U_k^{r-1}, C_i^r = C_j^{r-1} \bigcap C_k^{r-1}$

评测

- 缺少标定的数据，主观评价为主



线上数据

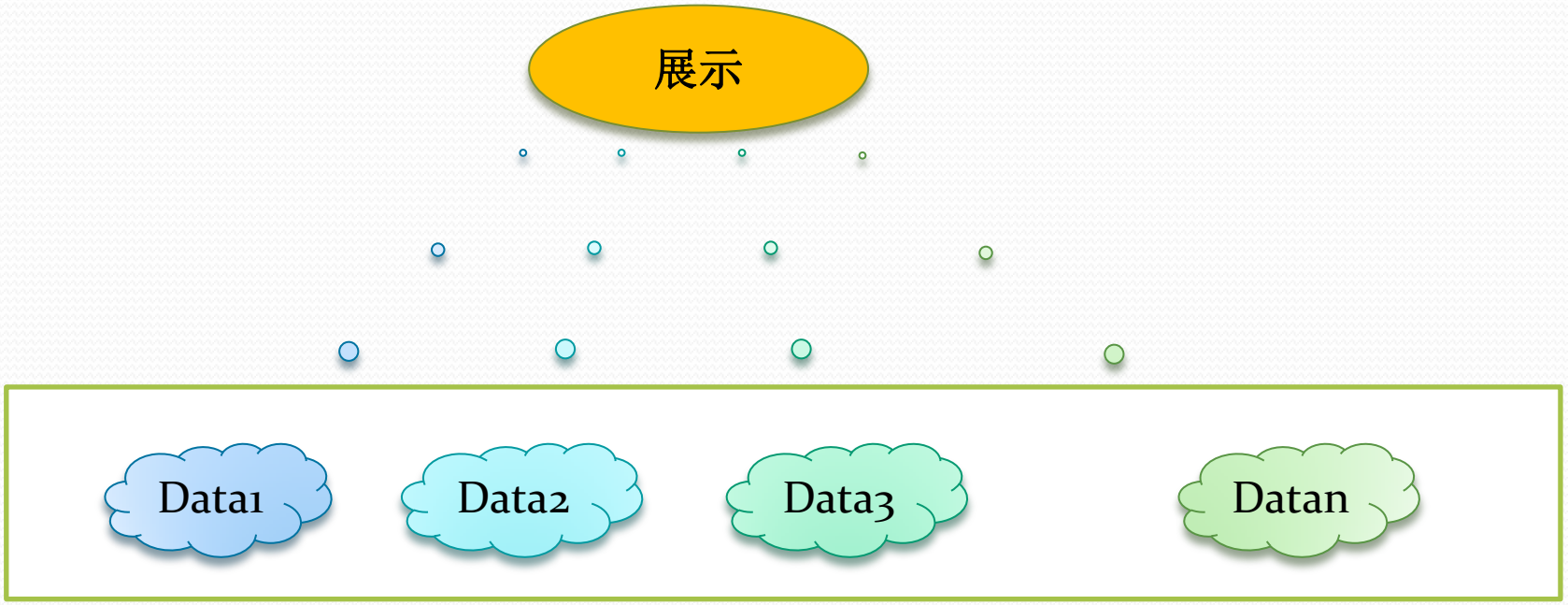


个性化推荐

- 用户偏好
- 用户兴趣
- 用户成长

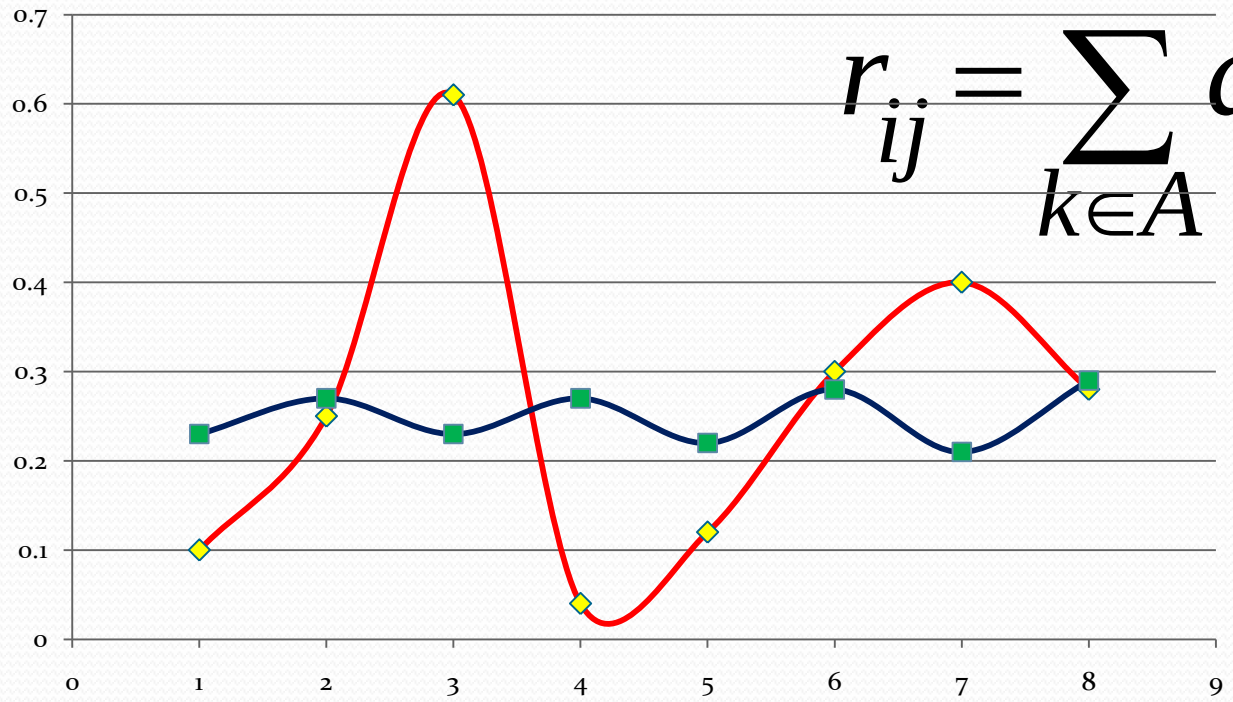
偏好

- 圈子
  - 年龄，学校，地域，性别.....



偏好的度量

- 用户行为
  - 访问个人主页、相册，分享，留言，评论等



$$r_{ij} = \sum_{k \in A} a^k \cdot n_{ij}^k$$

量化

- 信息熵
  - 年龄，学校，地域，性别
    - x =（0.1， 0.25， 0.61， 0.04）≈nonuniform
    - y =（0.23，0.27， 0.23，0.27）≈uniform

$$H[x] = -\sum_{x} p(x) \log p(x)$$

    - H(x)=-0.1*log0.1-0.25*log0.25-0.61*log0.61-0.04*log0.04=1.007
    - H(y)=-0.23*log0.23-0.27*log0.27-0.23*log0.23-0.27*log0.27=1.383

好友推荐

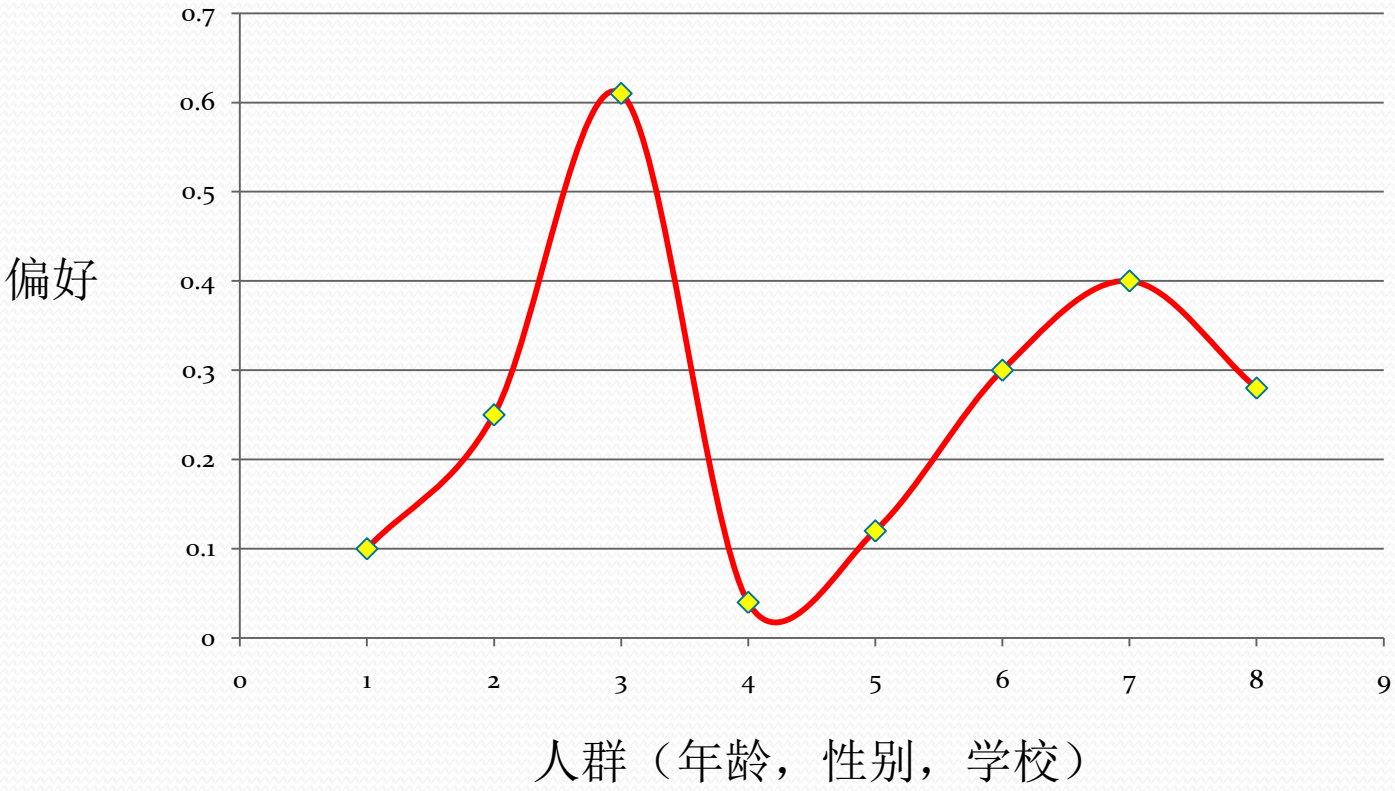
- 首页28个推荐位，从推荐源数据中随机选取

Ranking VS Sampling

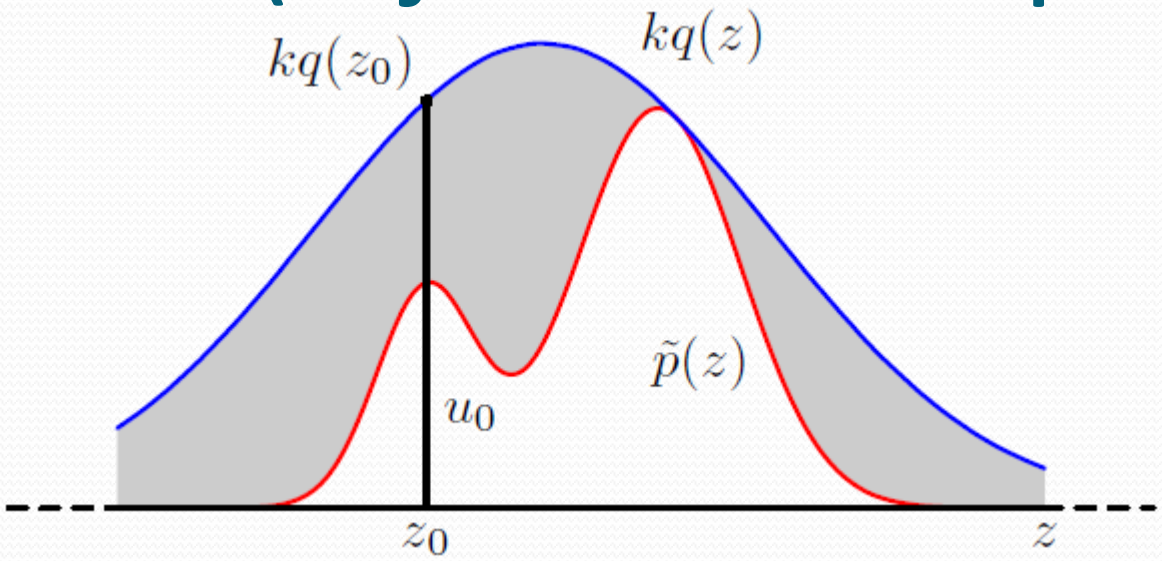
- Ranking
  - 亲密度
  - 影响力
- Sampling
  - 偏好
  - 推荐人群的多样化
  - 新颖性

采样

- 偏好分布密度函数P(x)未知



取舍抽样(rejection sampling)



$$p(accept) = \int \{\tilde{p}(z) / kq(z)\} \bullet q(z) dz$$

$$= \frac{1}{k} \int \tilde{p}(z) dz$$

用户兴趣

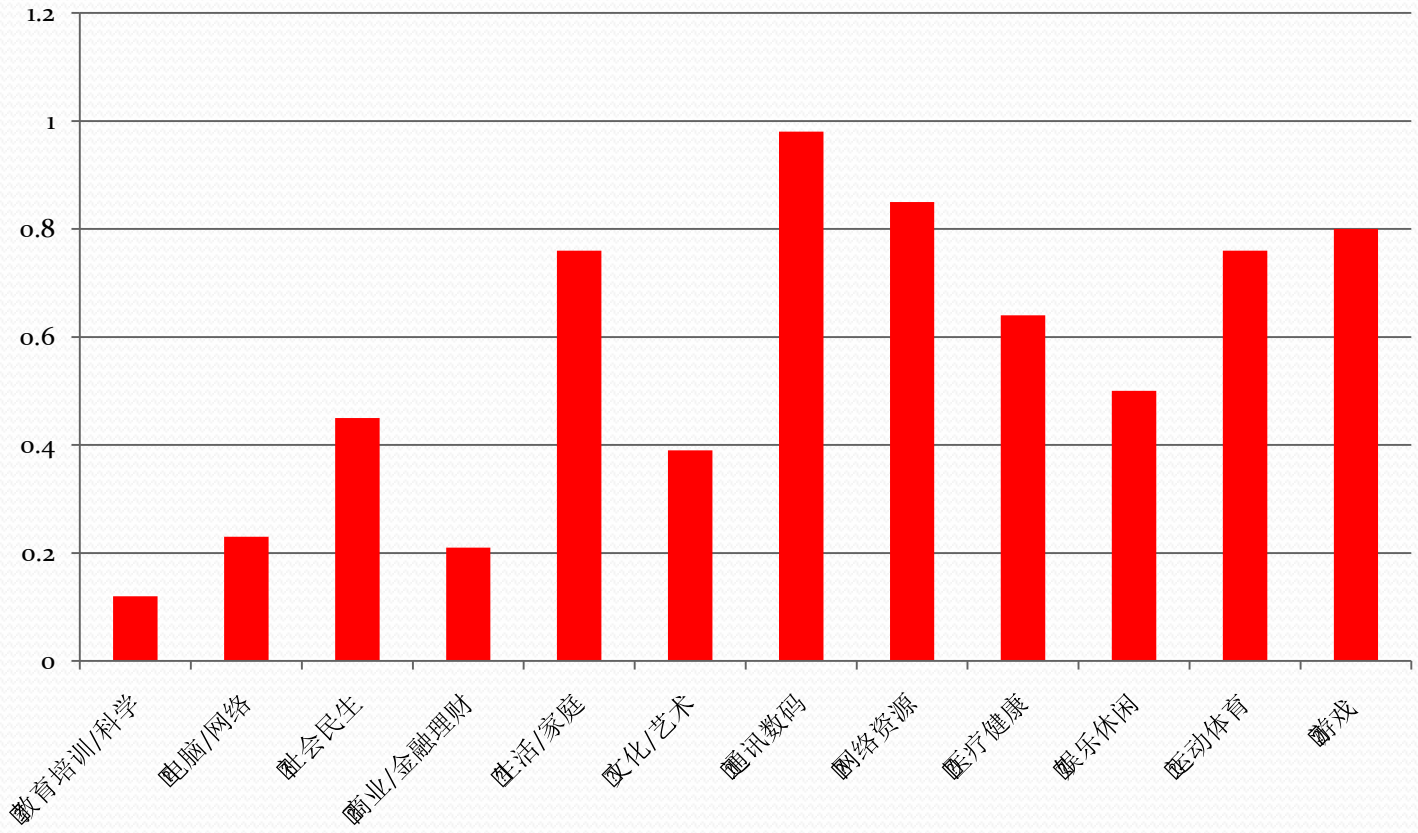
- 教育培训/科学
- 电脑/网络
- 社会民生
- 商业/金融理财
- 生活/家庭
- 文化/艺术
- 通讯数码
- 网络资源
- 医疗健康
- 娱乐休闲
- 运动体育
- 游戏
- ……

- 网站推荐
- 视频共享
- 音频/歌曲共享
- 软件共享
- 文档/报告共享
- 华语流行乐
- 日韩流行乐
- 欧美流行乐
- 电影
- 电视
- 动漫
- 华人明星
- ……

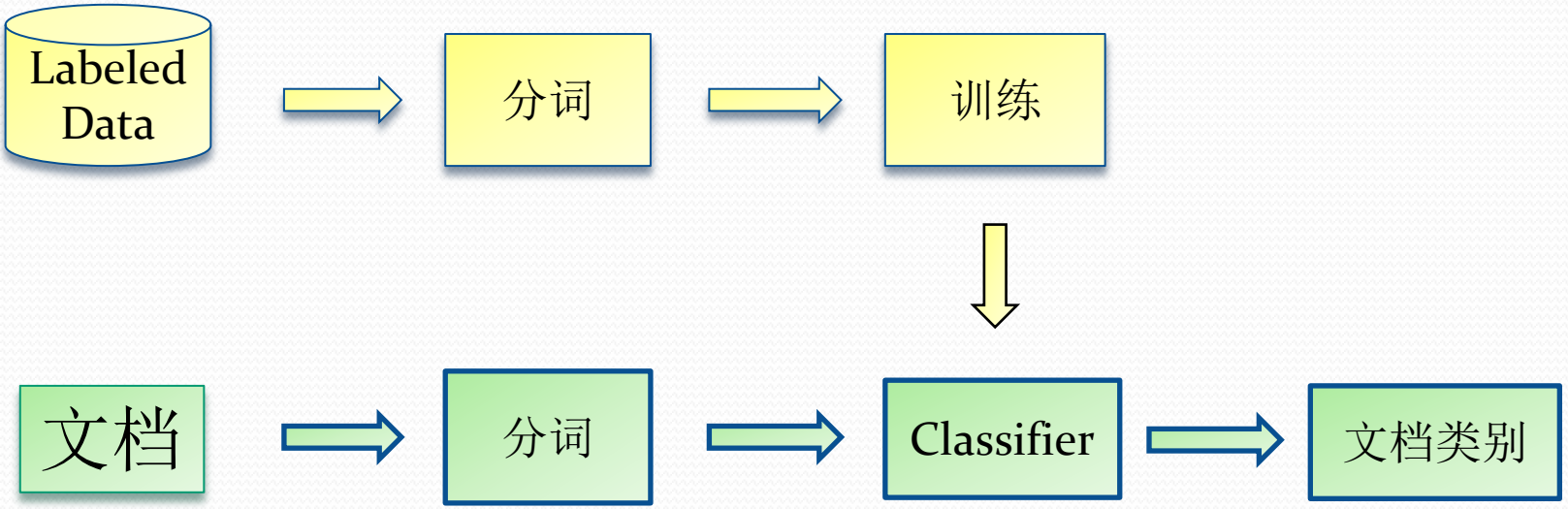
- 关键词1
- 关键词2
- 关键词3
- 关键词4
- 关键词5
- ……
- ……
- ……
- ……
- ……
- ……
- ……

User access

- Access log



文档分类Supervised learning



Naïve Bayes

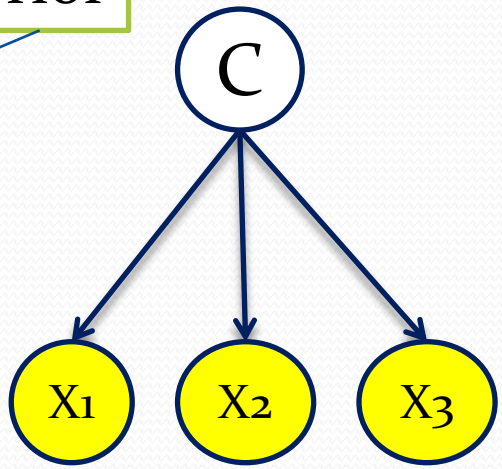
- Bayes

likelihood

prior

$$p(C \mid \vec{x}) = \frac{p(\vec{x} \mid C_k)\, p(C_k)}{p(\vec{x})}$$

posterior



- Conditional independent

$$p(C_k \mid x_1, x_2) \propto p(x_1, x_2 \mid C_k)\, p(C_k)$$

$$\propto p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(C_k)$$

内容过滤

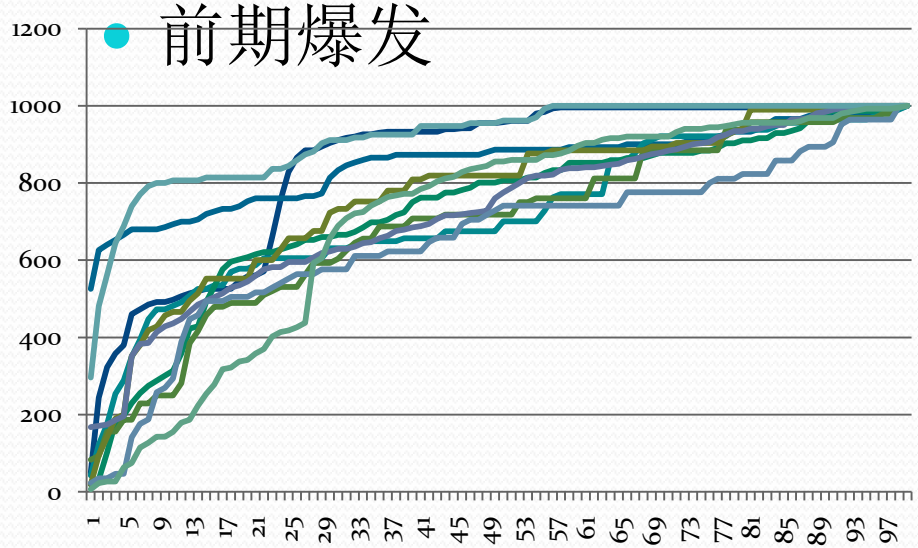
- 公共主页
  - 相关性

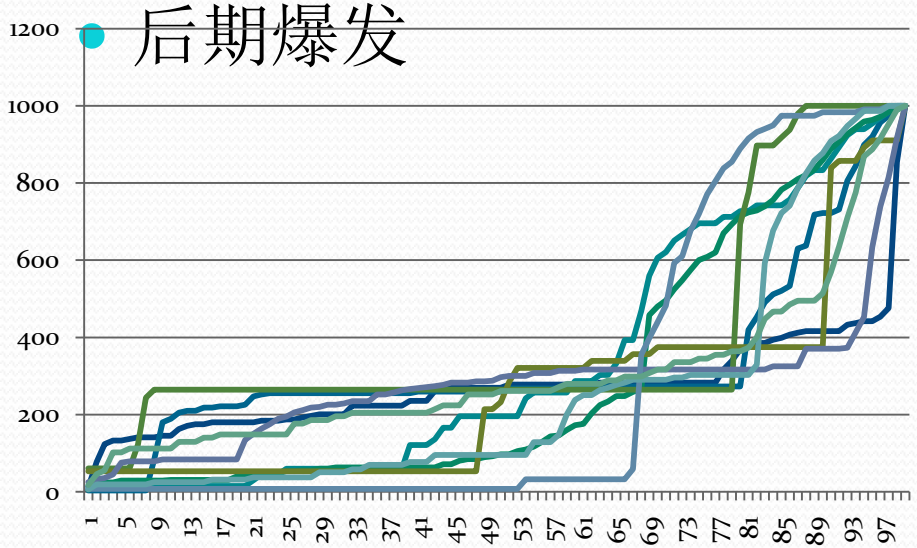
$$\cos(U,P)=(\sum_i U_i)^T \bullet (\sum_j P_j) / \left\|\sum_i U_i\right\| \left\|\sum_j P_j\right\|$$

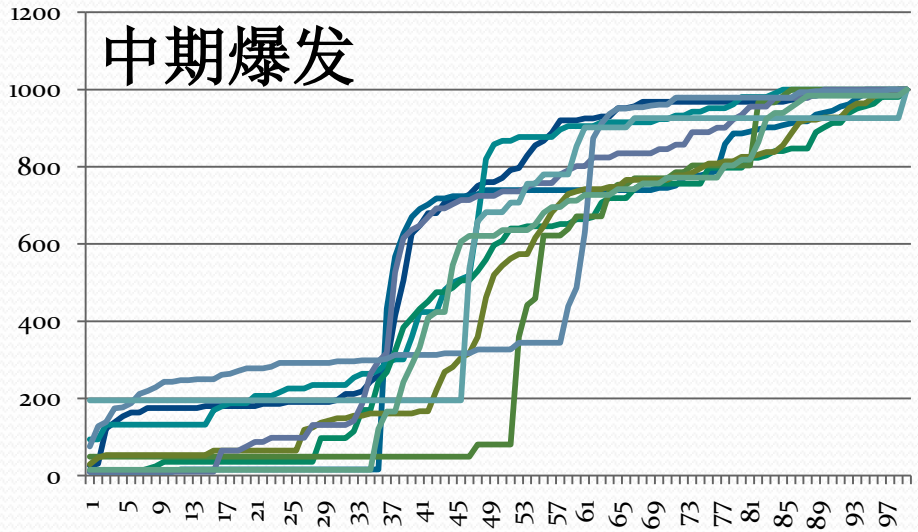
用户成长

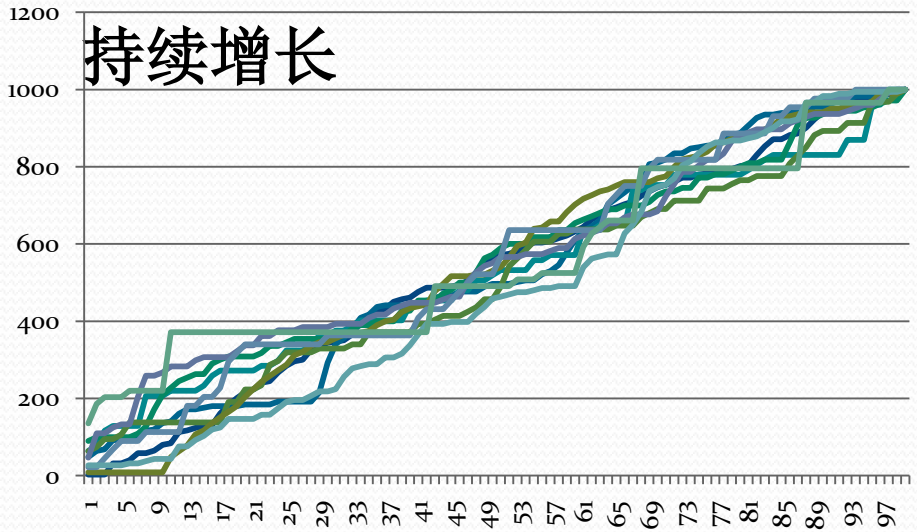
- 问题1
  - 用户在什么阶段加好友最多？
  - 爆发，还是持续增长？
- 问题2
  - 用户在什么阶段内，加什么样的好友？
  - 资料？二度好友？陌生人？

好友增长类型


前期爆发
1200
1000
800
600
400
200
0
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97

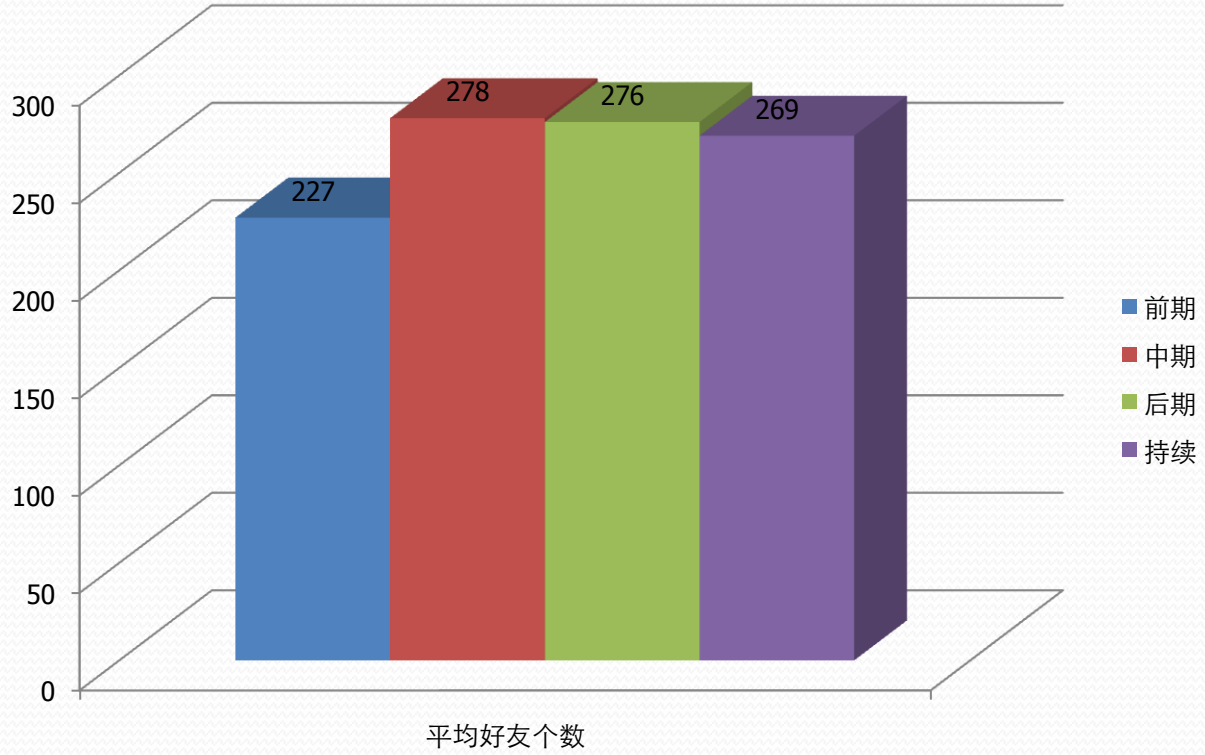

后期爆发
1200
1000
800
600
400
200
0
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97


中期爆发
1200
1000
800
600
400
200
0
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97


持续增长
1200
1000
800
600
400
200
0
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97

不同类型增长好友数

前期增长最终达到的好友数目较低



推荐

- 好友，公共主页，小站，音乐，日志，视频，小组.....
- 年龄a，好友数目n，兴趣i，活跃度d，推荐r

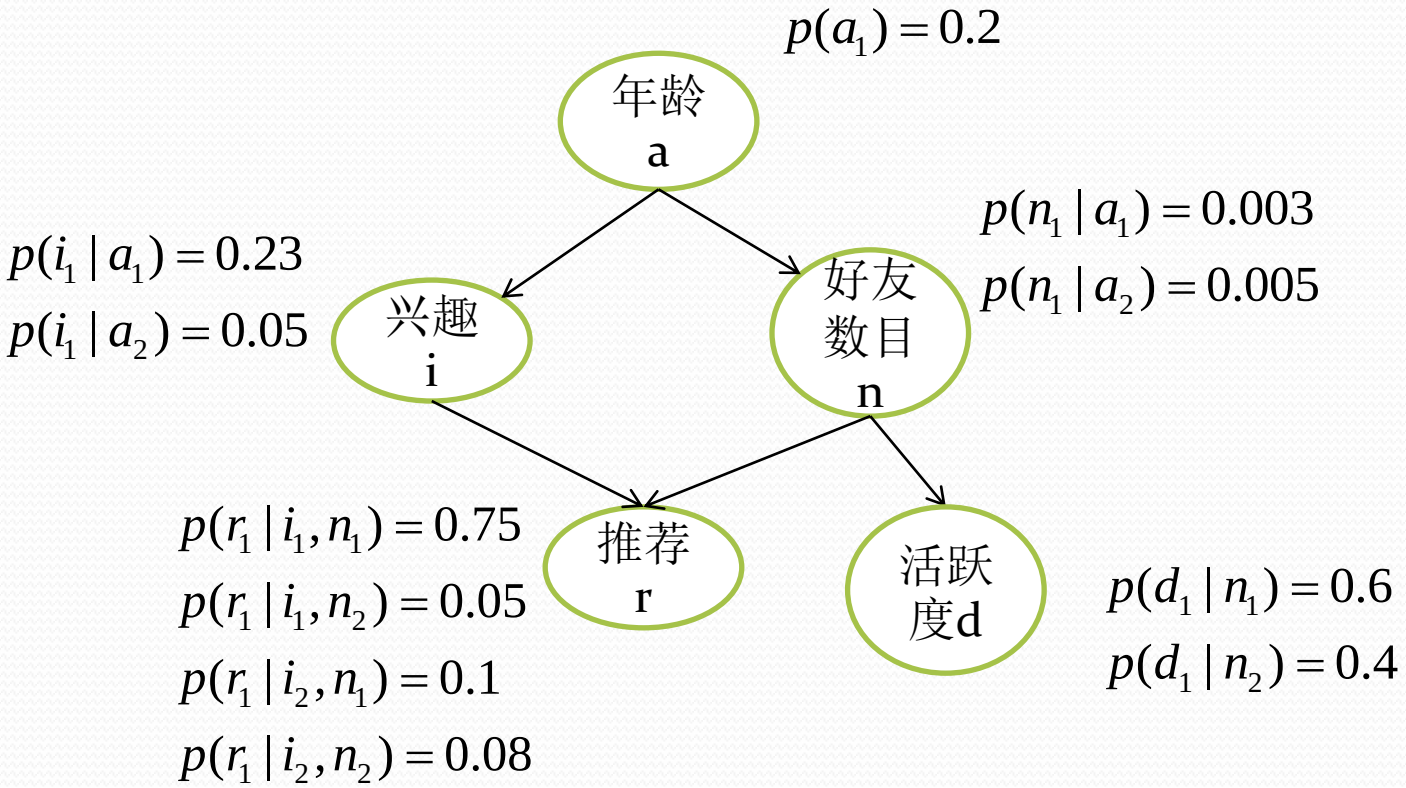
$$p(r \mid a,n,d)$$

$$= \frac{p(r,a,n,d)}{p(a,n,d)} = \frac{\sum_{i} p(r,a,i,n,d)}{\sum_{i,r} p(r,a,i,n,d)}$$

$$2^5$$

贝叶斯推理

- 先验知识：用户成长模型，用户兴趣，用户偏好



Factorization

$$p(a,i,n,r,d)$$
$$= p(a)\,p(i\mid a)\,p(n\mid a)\,p(r\mid i)\,p(r\mid n)\,p(d\mid n)$$

- 谢谢！