面向社区媒体的用户分享行为预测

蒋 朦
清华大学计算机系
媒体所 1-512

总纲

- 背景及意义
- 预测方法设计
  - 数据准备
  - 媒体数据的话题模型
  - 用户的偏好模型
  - 用户的影响力模型
  - 用户的分享行为建模
- 预测结果验证
- 总结及展望

好友新鲜事与用户分享行为

社区媒体

开心网

人人网 renren.com

youku.com

flickr

转帖

社交网络

分享

?

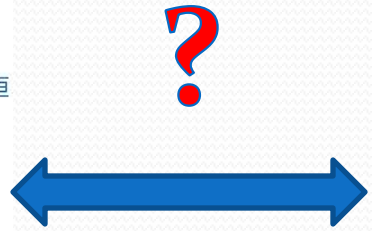
分享行为触发原因

用户注意到好友对媒体信息的分享推荐，浏览产生兴趣

人人网 renren.com

用户因兴趣而主动对媒体信息产生网络行为

基于内容的行为预测



1 内容相同的已评价信息

2 用户对已评价信息的评价

？用户对信息的评价

社交网络行为特征

用户影响力

兴趣偏好

?

话题层面

计算机

电影

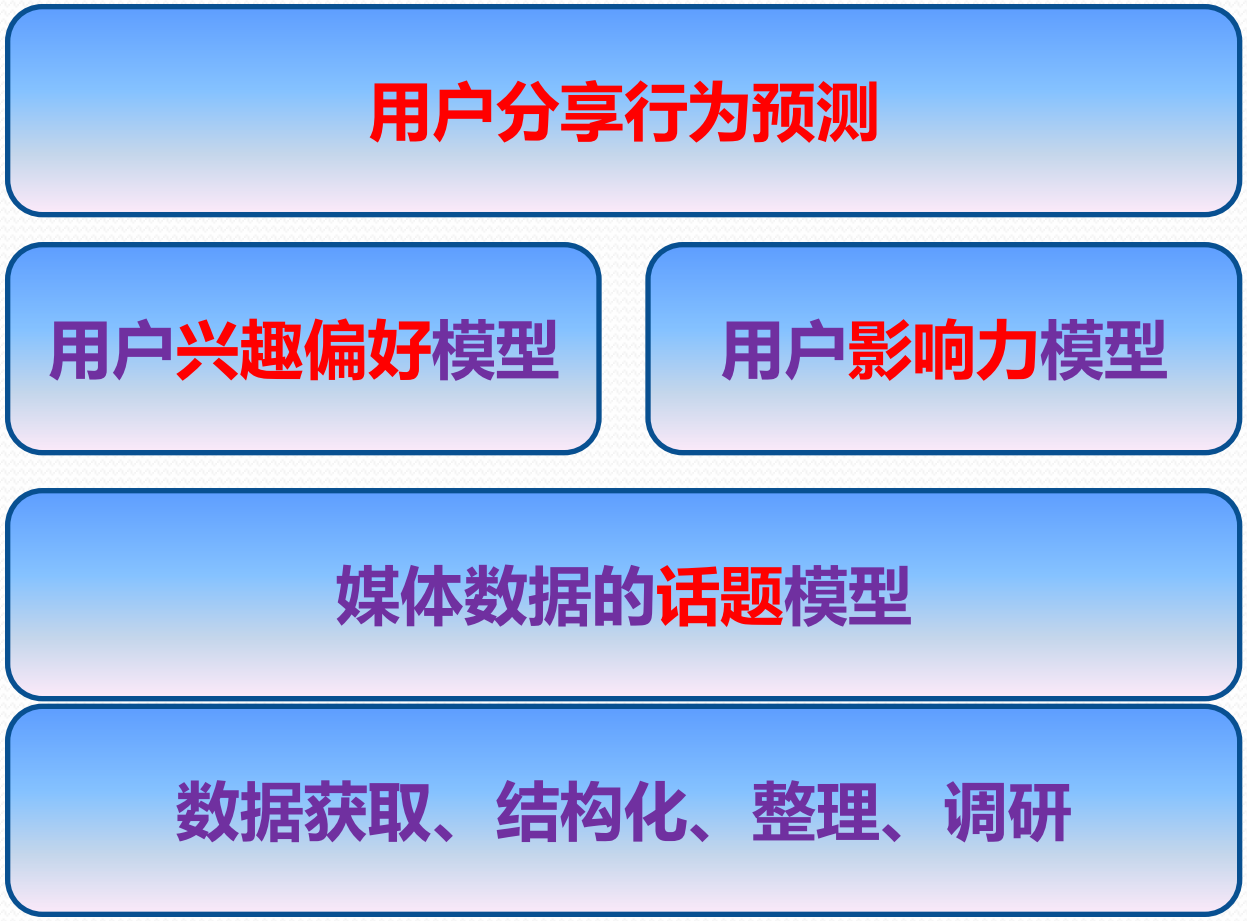
炒饭

笔记本

台式机

微机

分享行为预测框架

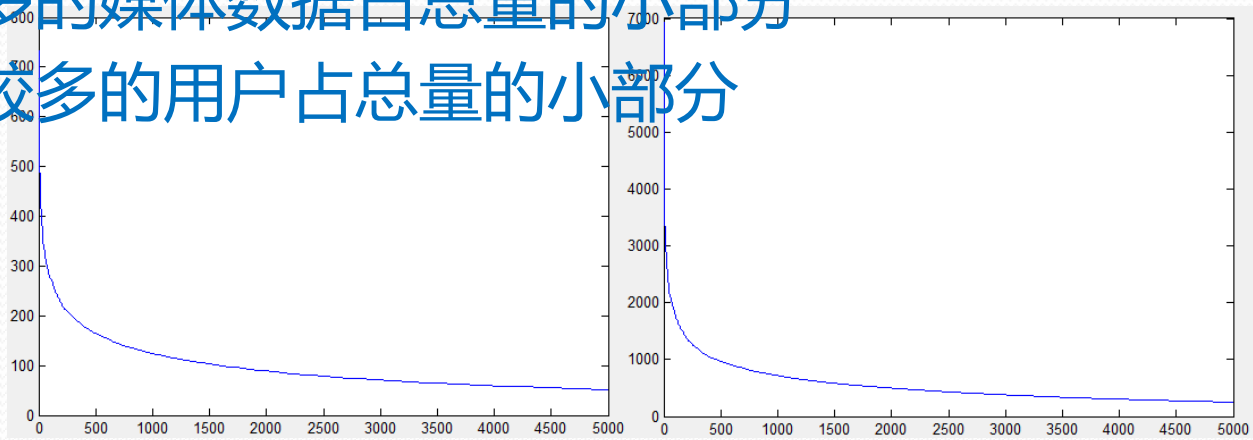


总纲

- 背景及意义
- 预测方法设计
  - 数据准备
  - 媒体数据的话题模型
  - 用户的偏好模型
  - 用户的影响力模型
  - 用户的分享行为建模
- 预测结果验证
- 总结及展望

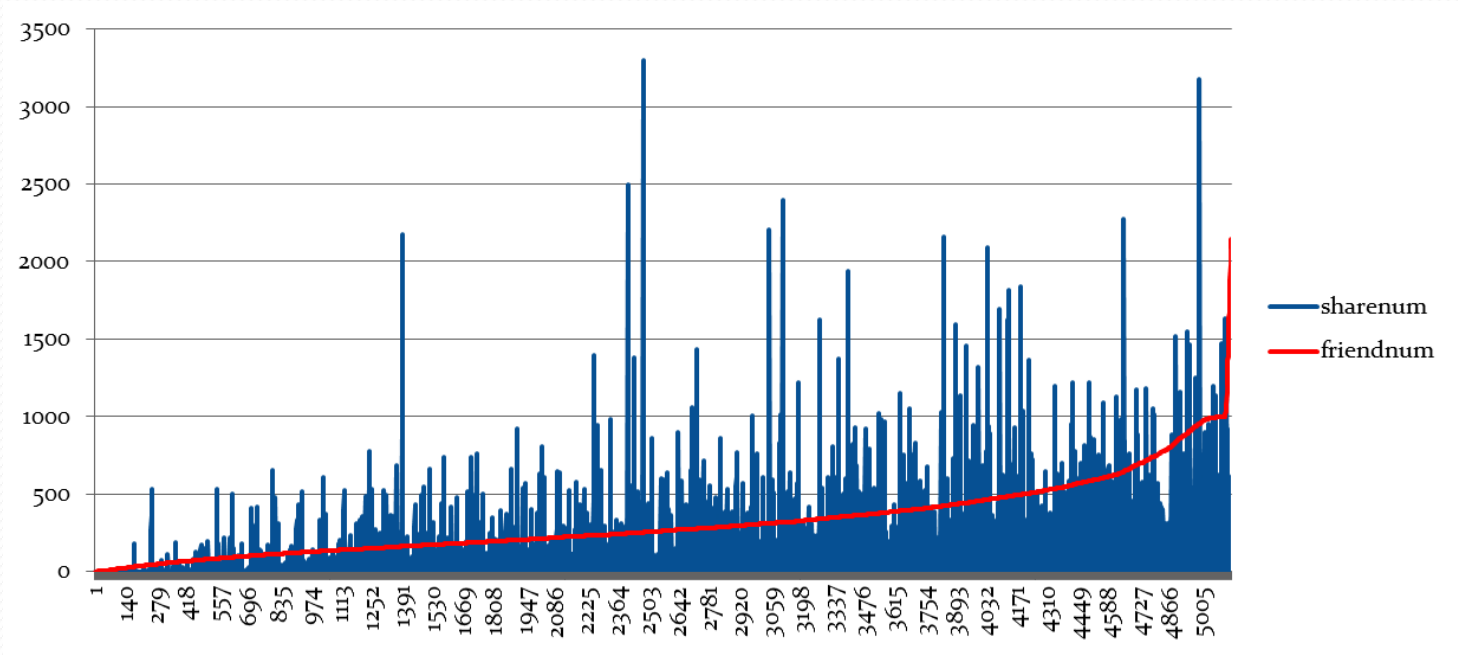
数据准备

- 数据来源：爬取近十万人人用户的千万条分享记录
- 数据结构
  - 网络结构数据：姓名，好友关系等
  - 媒体数据：分享记录类型，标题，分享者，时间等
- 长尾效应：贡献按贡献排行呈负指数分布
  - 分享者较多的媒体数据占总量的小部分
  - 分享记录较多的用户占总量的小部分



数据准备

- 用户好友数量与分享数量的关系（相辅相成）
  - 用户活跃，分享增多，被关注更多，好友增多
  - 好友增多，接触信息增多，用户分享行为活跃



总纲

- 背景及意义
- 预测方法设计
  - 数据准备
  - 媒体数据的话题模型
  - 用户的偏好模型
  - 用户的影响力模型
  - 用户的分享行为建模
- 预测结果验证
- 总结及展望

媒体数据整理

- 分享标题：清华大学学生献血和募捐的一些最终数据
- 中文分词ICTCLAS，选用名词和动词
- 去除词长过短、词频不足和无用词：37,633
- 纯文本信息：清华大学 学生 献血 募捐 数据
- 聚类优于分类：词汇数量多，生活化，不规则
- LDA：根据文本词汇分布，聚类生成话题模型
- 10个话题，4000个单话题词汇，10000次迭代
- 语义归纳话题特征

词汇聚类结果

话题0	话题1	话题2	话题3	话题4	话题5	话题6	话题7	话题8	话题9
中国	男人	视频	英语	时尚	北京	大学	中国	语录	手机
美国	女人	电影	考试	广告	城市	学生	世界	老师	电脑
新闻	星座	音乐	学习	摄影	中国	学院	视频	高考	生活
日本	女生	歌曲	网站	世界	上海	毕业	北京	人生	减肥
总理	男生	明星	大学生	收藏	地震	活动	照片	同学	朋友
世界	女孩	专辑	大学	创意	家乡	中国	开幕式	作文	图片
韩国	爱情	娱乐	专业	love	南京	通知	nba	高中	游戏
国家	朋友	歌词	工作	品牌	小吃	同学	足球	笑话	照片

话题0	话题1	话题2	话题3	话题4
时政新闻	内心情感	娱乐媒体	考场职场	时尚潮流
话题5	话题6	话题7	话题8	话题9
地区特色	校园文化	体育赛事	网络流行	生活百态

媒体数据的话题模型

- 话题数量：M=10
- 索引库中第i条分享记录话题向量$S(i)=[s_{i,0} \ldots s_{i,M-1}]$

样例：清华大学学生献血和募捐的一些最终数据

m	0	1	2	3	4	5	6	7	8	9
$s_{i,m}$	0.03	0.03	0.03	0.03	0.03	0.03	0.70	0.03	0.03	0.03

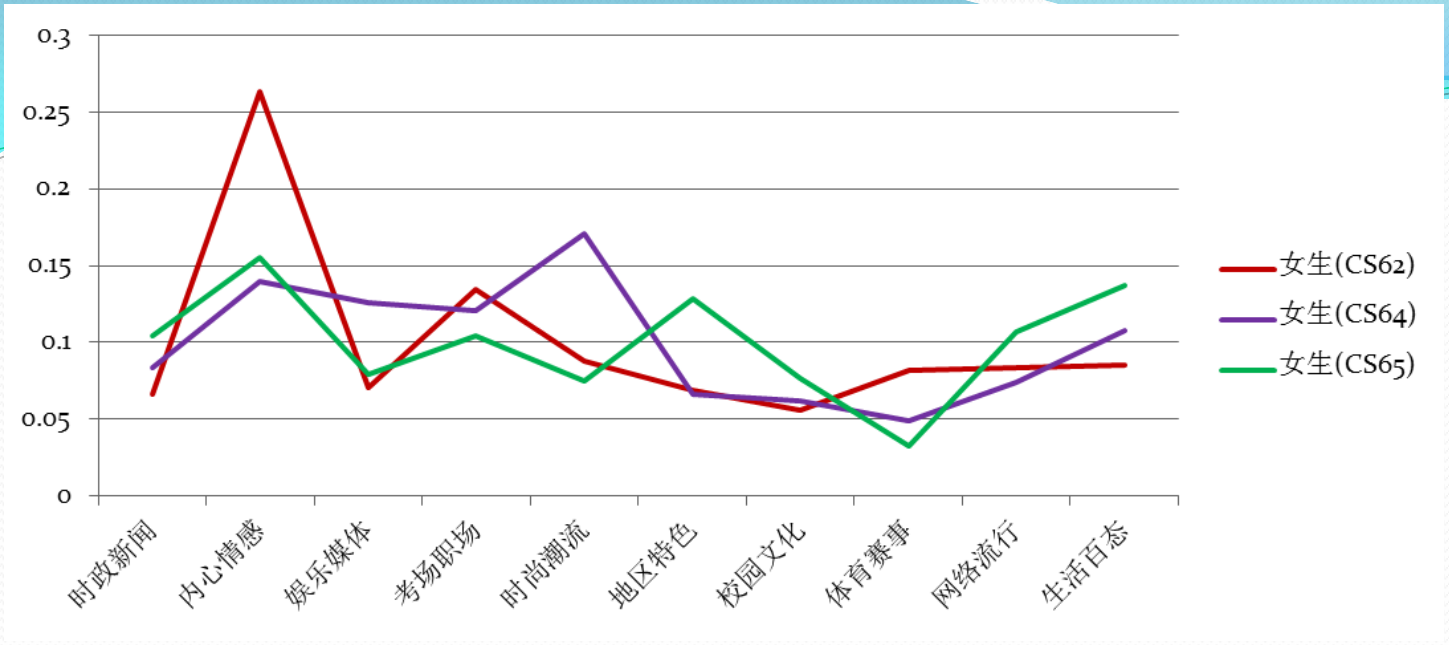
- 话题模型显示样例偏向话题6：**校园文化**
- 这是验证媒体数据话题模型的一种方法
  - 从索引库中随机选取若干条分享记录
  - 求解偏向话题特征，并与实际内容比对

总纲

- 背景及意义
- 预测方法设计
  - 数据准备
  - 媒体数据的话题模型
  - 用户的偏好模型
  - 用户的影响力模型
  - 用户的分享行为建模
- 预测结果验证
- 总结及展望

用户的偏好模型

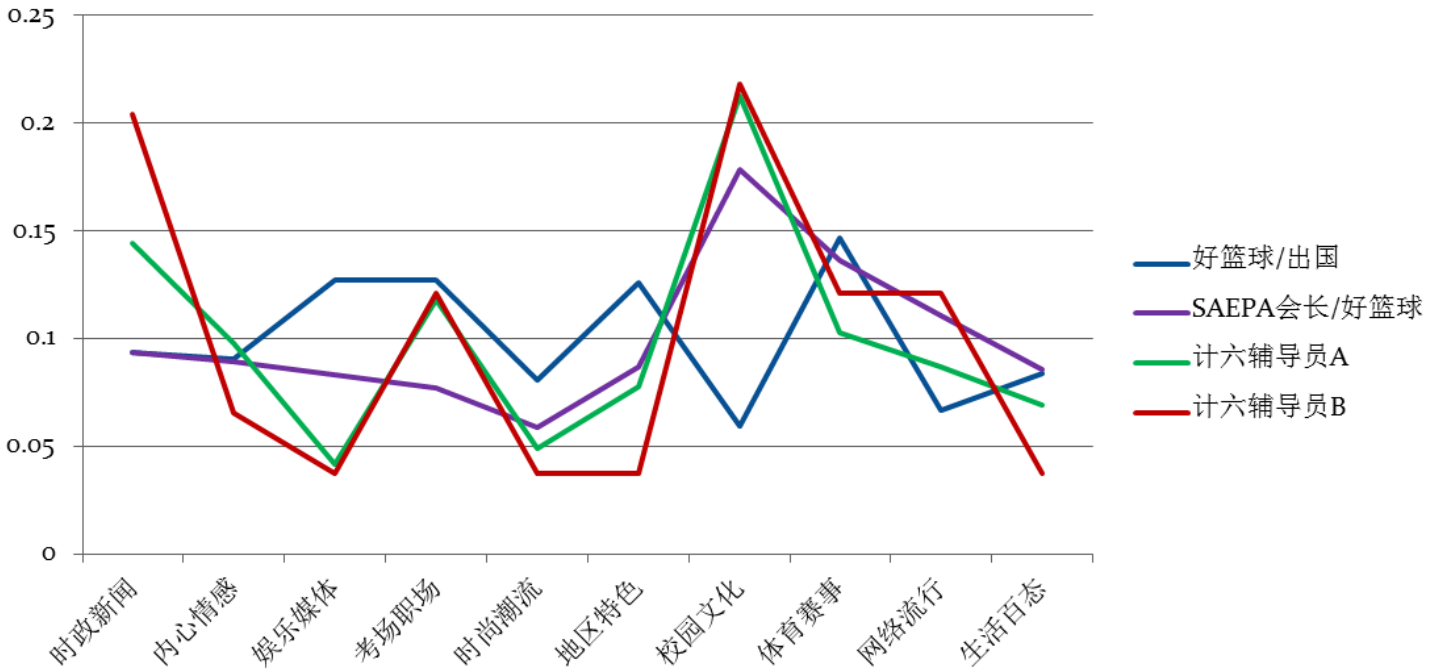
- 用户分享行为偏好主要体现在分享记录上
  - 用户偏好向量$P(p)=[p_{i,0} \cdots p_{i,M-1}]$
  - $p_j = sum(s_{i,j})/N$：记录编号i，话题j，分享列表长度N
- 偏好模型验证话题模型
  - 活跃用户在现实生活中的社会身份与兴趣爱好在分享行为上有较强烈反应
  - 调查问卷User Study
  - 克服语义测试的数据量大、判断难度大的缺点
  - 这里活跃用户指分享数量在300条以上



女生更为关注
内心情感
时尚潮流

好篮球
关注体育赛事

辅导员及协会会长
关注校园文化



总纲

- 背景及意义
- 预测方法设计
  - 数据准备
  - 媒体数据的话题模型
  - 用户的偏好模型
  - 用户的影响力模型
  - 用户的分享行为建模
- 预测结果验证
- 总结及展望

用户影响力模型

- 用户好友对用户分享行为的影响因子矩阵
  - $L = \begin{bmatrix} l_{0,0} & \cdots & l_{0,M-1} \\ \vdots & \ddots & \vdots \\ l_{N-1,0} & \cdots & l_{N-1,M-1} \end{bmatrix}$
  - $l_{i,j}$表示在话题j上用户好友u_i对用户的影响程度
- 用户好友的分享向量$F(p)=[f_0 \ldots f_{N-1}]$
  - $f_i = \begin{cases} 1 & \text{用户好友} u_i \text{分享过该媒体信息} \\ 0 & \text{用户好友} u_i \text{未分享该媒体信息} \end{cases}$

偏好相似度衡量影响力

专家效应

共同语言

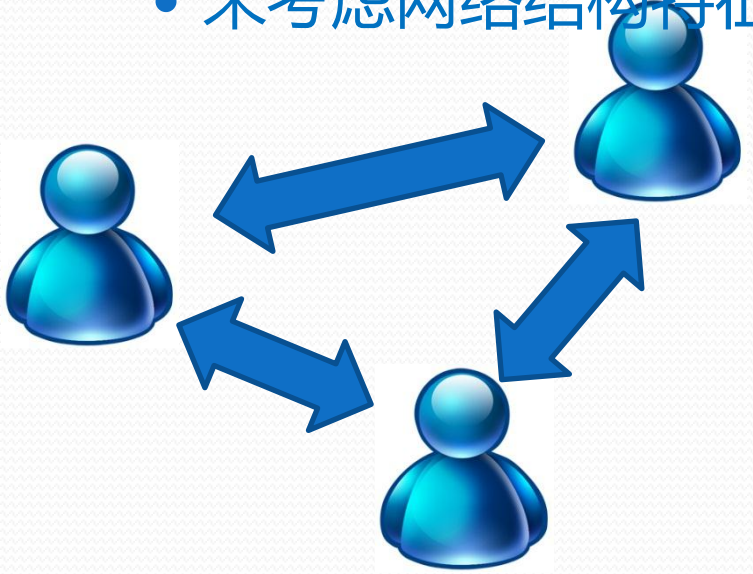
知音难寻

用户偏好相似度

- 用户p_1和p_2的偏好向量为
  - P(p_1)=[p_1(0)… p_1(M-1)]
  - P(p_2)=[p_2(0)… p_2(M-1)]
- 偏好相似度：向量夹角余弦值

$$l(p_1, p_2) = simi(p_1, p_2) = cos[\theta(p1, p2)] = \sum_{j=0}^{M-1} p_1(j) \cdot p_2(j)$$

- 优势
  - 用户兴趣特征
  - 媒体数据的语义特征
- 缺点
  - 未考虑网络结构特征



总纲

- 背景及意义
- 预测方法设计
  - 数据准备
  - 媒体数据的话题模型
  - 用户的偏好模型
  - 用户的影响力模型
  - 用户的分享行为建模
- 预测结果验证
- 总结及展望

用户分享行为概率

- 融合影响力与用户偏好的分享行为概率算法
- 话题向量S，偏好向量P，分享向量F，影响因子矩阵L
- 好友新鲜事触发行为概率

$$pref_{friend} = F \cdot L \cdot S^T = \sum_{i,j} f_i p_j p_{i,j} s_j$$

- 媒体信息兴趣触发行为概率

$$pref_{media} = P \cdot S^T = \sum_{j} p_j s_j$$

- 用户的分享行为概率

$$pref = \begin{cases} \alpha \cdot pref_{friend} & pref_{friend} > 0 \\ \beta \cdot pref_{media} & pref_{friend} = 0 \end{cases}$$

用户分享行为概率计算算法

输入：用户及好友关系列表，分享记录索引库，给定时刻t、用户user、媒体信息media

输出：在给定时刻t，给定用户user对给定媒体信息media的分享行为概率pref

1：初始化社交用户网络结构图

遍历用户及好友关系列表：更新用户的人人id，姓名，好友关系

遍历分享记录索引库：更新用户的分享记录列表，包含分享记录对应的话题向量

2：遍历用户网络结构图中每个用户：对所有分享记录的话题向量平均得到偏好向量

3：获取给定媒体信息的话题向量s[j]；获取给定用户的偏好向量p[j]及其好友u[i]的偏好向量p[i,j]；初始化好友分享向量f[i]，遍历用户的每个好友u[i]，判断其分享记录列表中是否存在时刻t之前分享的给定媒体信息media

4：pref_friend<-0，pref_media<-0

5：for i=0 to N-1 { for j=0 to M-1 : pref_friend<-pref_friend+f[i]*p[j]*p[i,j]*s[j] }

6：for j=0 to M-1 : pref_media<-pref_media+ p[j]*s[j]

7：if pref_friend > 0 用户对媒体信息的兴趣度pref=alpha*pref_friend

else pref=beta*pref_media

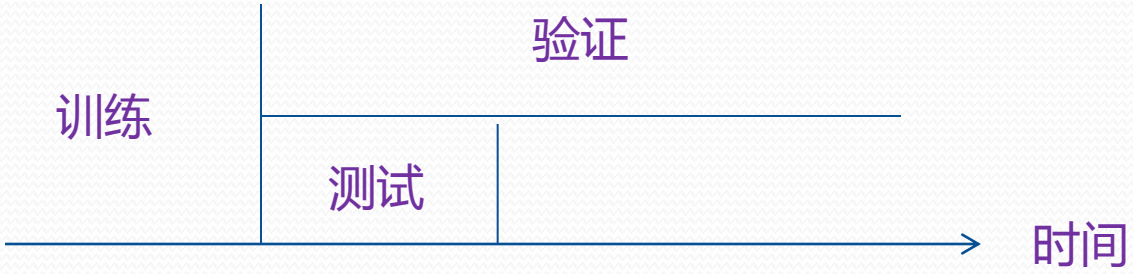
其中i∈[0,N)，j∈[0,M)，N为用户的好友数量，M为话题数量

总纲

- 背景及意义
- 预测方法设计
  - 数据准备
  - 媒体数据的话题模型
  - 用户的偏好模型
  - 用户的影响力模型
  - 用户的分享行为建模
- 预测结果验证
- 总结及展望

用户行为预测验证

- 验证方法（时间窗）
  - 训练时间：根据用户分享记录训练得出偏好模型
  - 测试时间：收集用户可能接触到的信息，计算行为概率
  - 验证时间：判断概率高的媒体信息是否被分享

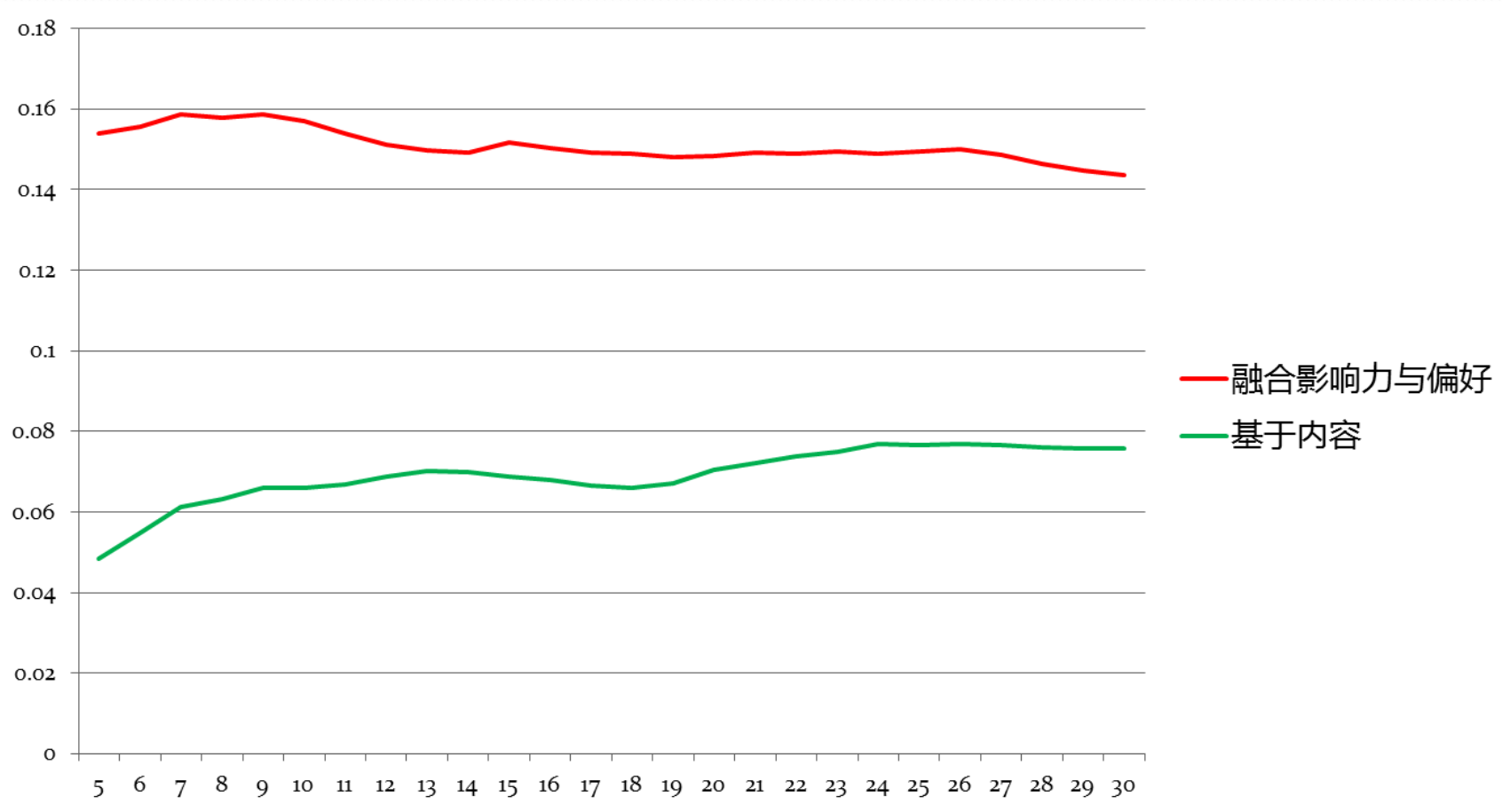


- 用户行为概率公式$pref=\gamma \cdot pref_{friend}+pref_{media}$
  - γ越大，用户兴趣度受好友影响较信息内容越强

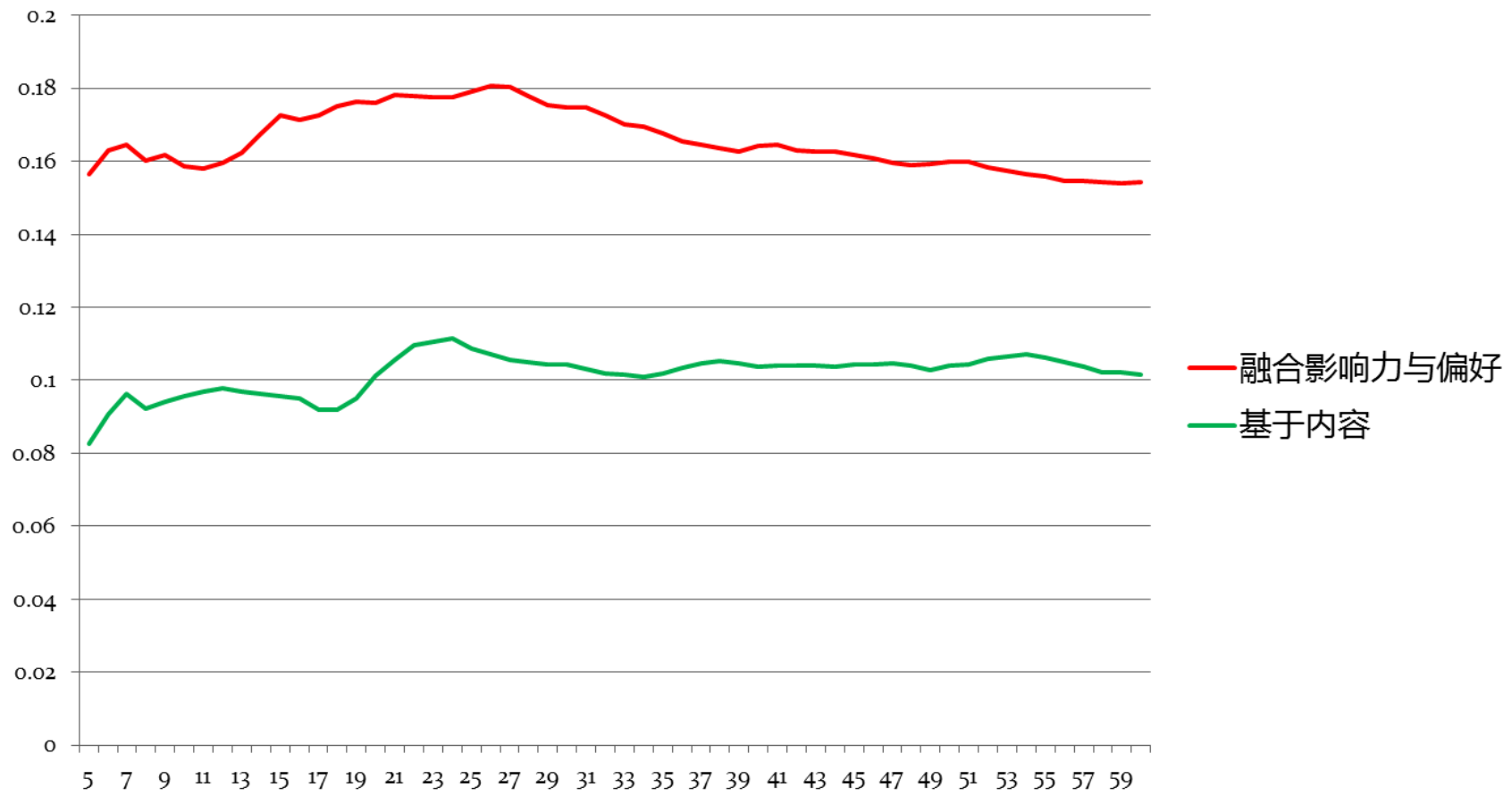
行为预测对比实验

- 融合用户影响力与偏好的行为预测算法
- 基于内容的行为预测算法
  - 根据用户已分享信息归纳用户兴趣偏好特征
  - 根据媒体信息内容与兴趣偏好匹配程度预测分享行为
- 验证指标：训练时长3个月，测试时间2009年9月
  - 设定分享数量的下限为min_k（30，60，90）
  - 随机挑选400名用户，分析分享数量达到要求的人
  - 参数k从5变化至min_k，系统推荐概率高的前k条信息
  - 用户实际分享其中x条，则准确率precision=x/k

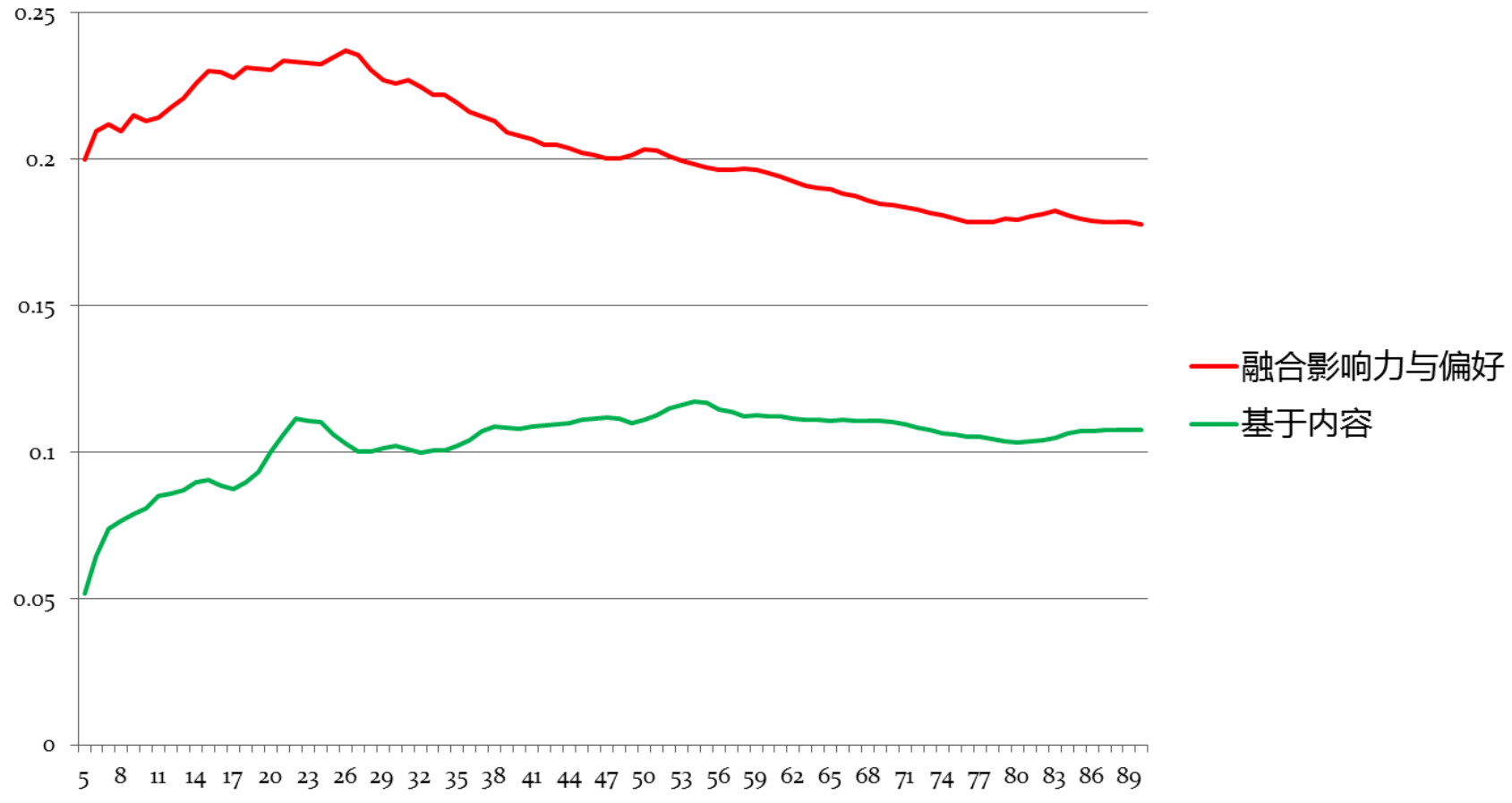
分享达到30条的用户


0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
融合影响力与偏好
基于内容

分享达到60条的用户


0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59
融合影响力与偏好
基于内容

分享达到90条的用户


0.25
0.2
0.15
0.1
0.05
0
5 8 11 14 17 20 23 26 29 32 35 38 41 44 47 50 53 56 59 62 65 68 71 74 77 80 83 86 89
融合影响力与偏好
基于内容

验证结果总结

分享数量下限	融合用户影响力与偏好的行为预测算法 较基于内容的行为预测算法的提高程度	
min_k	k:5~min_k	k:5_10
30	118.1%	162.0%
60	62.9%	74.9%
90	96.9%	195.6%

- 推荐给用户的媒体信息较少时优势更为明显
  - 推荐信息少，分享行为概率高，算法效果显著
  - 避免信息冗余，符合用户习惯
- 在保证信息新鲜度的同时可以对好友新鲜事依此重排

总纲

- 背景及意义
- 预测方法设计
  - 数据准备
  - 媒体数据的话题模型
  - 用户的偏好模型
  - 用户的影响力模型
  - 用户的分享行为建模
- 预测结果验证
- 总结及展望

工作总结

- 数据获取、整理、调研
- 标题内容语义分析聚类
- 媒体数据的话题模型
- 用户的偏好模型
- 用户偏好相似度及影响力模型
- 用户分享行为建模及概率算法
- 用户行为预测的验证
- 融合影响力与偏好的行为预测　基于内容的行为预测

展望—提高用户行为预测效果

- 改良媒体数据的**话题模型**
  - 当前仅考虑标题文本信息
  - 可使用短文本拓展信息、用户标注信息、图像视频特征
- 改良用户**偏好模型**
  - 用户偏好不仅体现在分享媒体信息上
  - 需要考虑用户创造媒体信息内容、实际浏览内容、对媒体信息的评价、自我标注的兴趣特征
- 改良用户**影响力模型**
  - 偏好相似度只是影响力的一个因素
  - 可以考虑用户交流频度和网络结构亲密程度，如用户访问页面频率和用户共同好友数量

感谢

- 感谢杨士强教授、孙立峰副教授在工作方向上的细心指导和谆谆教诲
- 感谢崔鹏、刘璐等实验室师兄师姐在研究思路、分析问题的角度等各个方面给予宝贵意见和建议
- 感谢老师和同学们在毕业设计开题、中期时给予的肯定和帮助

谢谢！