推荐系统文章整理

Collaborative Topic Modeling for Recommending Scientific Articles

Wang C, Blei D M. Collaborative topic modeling for recommending scientific articles[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 448-456.

推荐系统的两类任务

- In-matrix 预测　Figure a 这种每篇文章至少被一个用户评价过的预测问题
- Out-matrix 预测　Figure b 像 4、5 这种每篇文章从未被别人评价过的预测问题，存在冷启动问题，通常的协同过滤算法无法处理

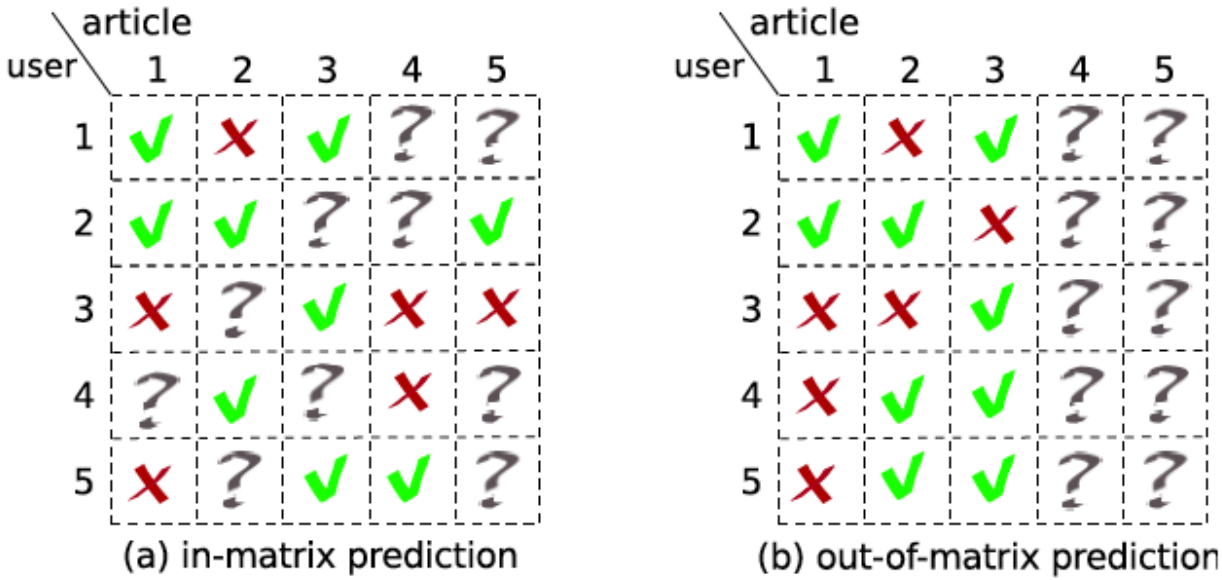


Figure: 两种推荐系统问题的图示，$\surd$ 表示喜欢，$\times$ 表示不喜欢，? 表示未被评分

矩阵分解方法做推荐

变量定义及损失函数

- 用低维空间来表示用户、物品隐向量 u_i, v_j，用户评分可以表示为 $\hat{r}_{ij} = u_i^T v_j$
- 于是可以转化为优化问题：最小化带正则的损失平方，如下式：

$$min_{U,V} \sum_{i,j} (r_{ij} - u_i^T v_j)^2 + \lambda_u \|u_i\|^2 + \lambda_v \|v_j\|^2$$

其中 $U = (u_i)_{i=1}^I$ 和 $V = (u_j)_{j=1}^J$

矩阵分解方法做推荐

probabilistic matrix factorization(PMF)

可以假设如下生成过程:

- 对于每个用户 i，用户 i 隐变量 $u_i \sim N(0, \lambda_u^{-1} I_K)$
- 对于每个物品 j，物品 j 隐变量 $v_j \sim N(0, \lambda_v^{-1} I_K)$
- 对于每个用户物品对 (i,j)，评分 $r_{ij} \sim N(u_i^T v_j, c_{ij}^{-1})$ 其中 c_{ij} 是针对 r_{ij} 的参数，定义如下:

$$c_{ij} = \begin{cases} a, & if \quad r_{ij} = 1 \\ b, & if \quad r_{ij} = 0 \end{cases}$$

- 具体理解参考:
  http://blog.csdn.net/shenxiaolu1984/article/details/50372909

概率主题模型

隐狄利克雷模型（文档主题模型）

隐狄利克雷模型生成过程：（注意：下列参数均为向量）

- 1、对于每篇文章 w_j，从参数为 α（参数预先给定）的狄利克雷分布得到主题参数，即:
  $\theta_j \sim Dirichlet(\alpha)$

- 2、对于每个单词 n
  (a) 从参数为 θ_j 的多项分布得到词的主题，即:
  $z_{jn} \sim Mult(\theta_j)$

  (b) 从参数为 $\beta_{z_{jn}}$（β 参数对应刚得到的主题）的狄利克雷分布得到单词分布参数，并由该参数通过多项分布生成最终单词，即:
  $w_{jn} \sim Mult(Dirichlet(\beta_{z_{jn}}))$

- 3、具体理解参考:
  http://blog.csdn.net/yhao2014/article/details/51098037

协同主题回归

模型布局

- α 为主题先验参数，生成 θ 主题后验参数，生成对应主题 z，然后根据 β_z 作为单词先验参数生成单词后验参数，最终生成单词（分布对应 LDA）
- 由主题后验参数 θ 加入干扰项 ε 得到隐物品向量 v，和隐用户向量 u 得到评分 r

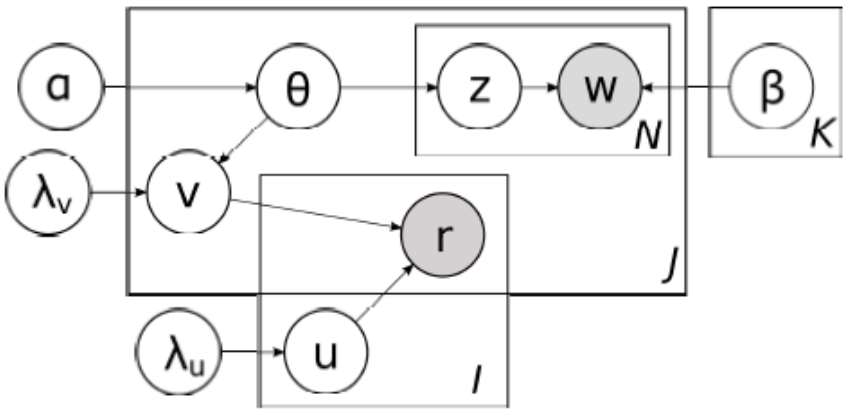


Figure: CTR 模型的图示

协同主题回归

模型生成过程

假设有 K 个主题，主题先验参数 $\beta=\beta_{1:K}$

- 1、对于每个用户 i，用户 i 隐变量 $u_i \sim N(0, \lambda_u^{-1} I_K)$
- 2、对于每个物品 j,
  (a) 主题后验参数 $\theta_j \sim Dirichlet(\alpha)$
  (b) 物品偏移量 $\varepsilon \sim N(0, \lambda_v^{-1} I_K)$，并得到隐物品向量
  $v_j = \varepsilon_j + \theta_j$，这个偏离量指文章内容以外的用户造成的影响
  (c) 对于每个单词 w_{jn}
  生成主题 $z_{jn} \sim Mult(\theta_j)$，单词 $w_{jn} \sim Mult(Dirichlet(\beta_{z_{jn}}))$
- 3、对于每个用户物品对 (i,j)，评分 $r_{ij} \sim N(u_i^T v_j, c_{ij}^{-1})$
  协同主题回归模型的解释
  $E[r_{ij}|u_i, \theta_j, \varepsilon_j] = u_i^T(\theta_j + \varepsilon_j)$

协同主题回归

参数学习-E 步

- 采用 EM 算法，E 步优化 u_i, v_j, θ_j，M 步优化 β
- 最大化似然函数，优化 U，V，$\theta_{1:J}$ 和 R，给定 $\lambda_u, \lambda_v, \beta$，损失函数如下
  $L = -\frac{\lambda_u}{2}\sum_i u_i^T u_i - \frac{\lambda_v}{2}\sum_j (v_j - \theta_j)^T(v_j - \theta_j)$
  $+\sum_j \sum_n log(\sum_k \theta_{jk}\beta_{k,w_{jn}}) - \sum_{i,j}\frac{c_{ij}}{2}(r_{ij} - u_i^T v_j)^2$
- 通过使偏导为 0 优化 u_i, v_j
  $u_i \rightarrow (VC_iV^T + \lambda_u I_K)^{-1} VC_iR_i$
  $v_j \rightarrow (UC_jU^T + \lambda_v I_K)^{-1}(UC_jR_j + \lambda_v \theta_j)$
- 定义 $q(z_{jn} = k) = \phi_{jnk}$，然后分离含有 θ_j 项的，并用 Jensen 不等式
  $\mathrm{L}(\theta_j, \phi_j) = -\frac{\lambda_v}{2}\sum_j (v_j - \theta_j)^T(v_j - \theta_j)$
  $+ \sum_n \sum_k \phi_{jnk}(log\theta_{jk}\beta_{k,w_{jn}} - log\phi_{jnk})$
  最优的 ϕ_{jnk} 服从 $\phi_{jnk} \propto \theta_{jk}\beta_{k,w_{jn}}$，文中使用投影梯度优化 θ_j

协同主题回归

参数学习-M 步

- 使用 E 步得到的 U、V 和 ϕ，优化 β 过程，与 LDA 模型中一致，即 $\beta \propto \sum_j \sum_n \phi_{jnk} 1[w_{jn} = w]$

预测

- D 作为观测到的数据，总体预测可以估计为：
  $E[r_{ij}|D] \approx E[u_i|D]^T (E[\theta_j|D] + E[\varepsilon_j|D])$
- in-matrix 问题，$r^*_{ij} \approx (u^*_i)^T(\theta^*_j + \varepsilon^*_j) = (u^*_i)^T v^*_j$
- out-matrix 问题，$r^*_{ij} \approx (u^*_i)^T \theta^*_j$

评估方法

- 评估公式
  $\text{recall@M} = \frac{number\ of\ articles\ the\ user\ like\ in\ topM}{total\ number\ of\ article\ the\ user\ \ likes}$
- in-matrix 预测
  5 倍交叉验证，每篇文章至少出现 5 次平均分配到每组中，少于 5 次的放入训练集
- out-of-matrix 预测
  5 倍交叉验证，平均分配，测试集测试其中从未出现在训练集中的文章即可

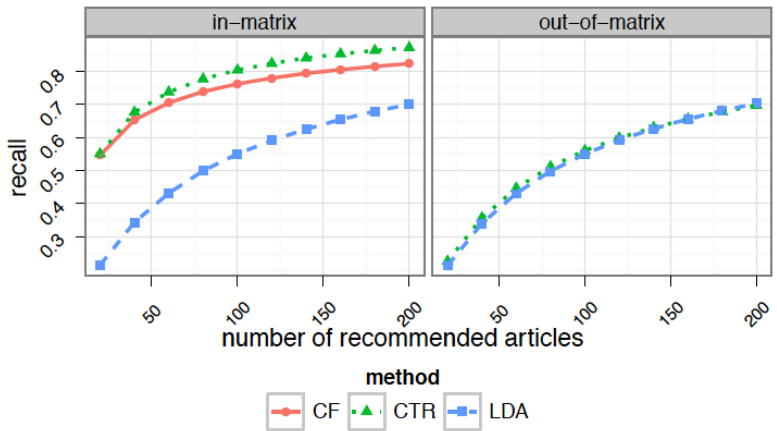


Figure: 对比 LDA、CF、CTR 在两种预测问题的召回率

评估 λ_v 参数影响

- λ_v 的影响
  λ_v 小的时候内容影响小，CTR 贴近 CF，λ_v 大的时候内容影响大，CTR 贴近 LDA

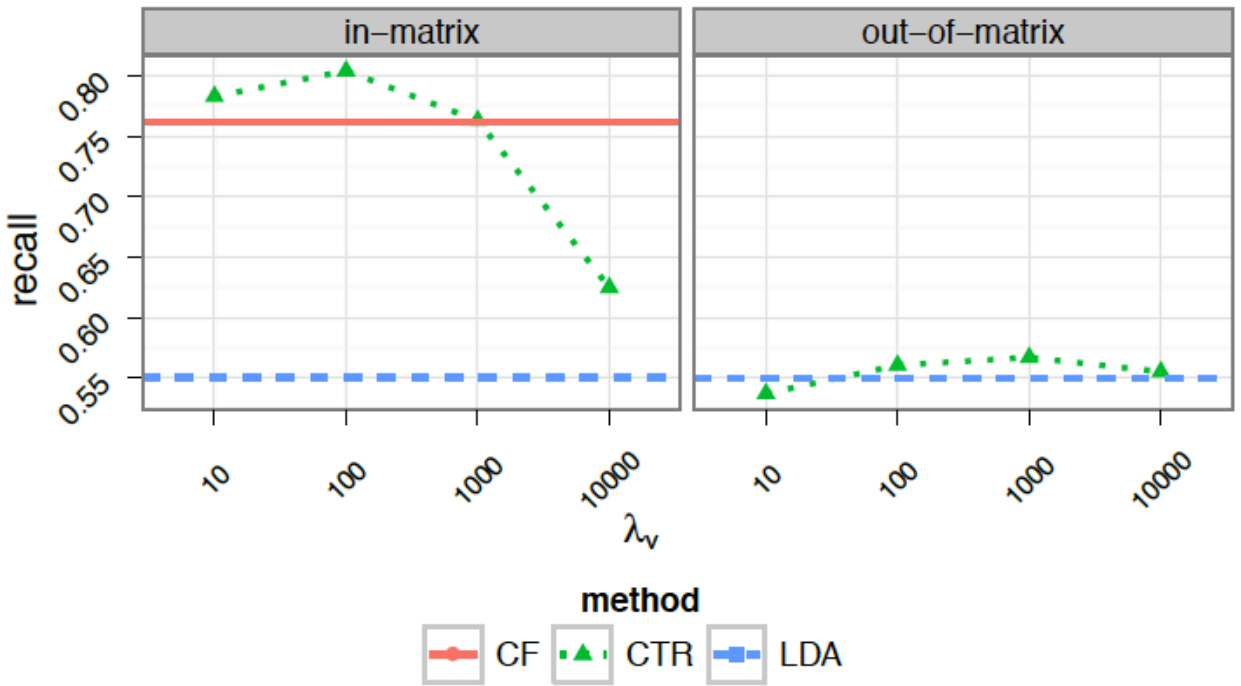


Figure: 在不同 λ_v 下，CTR 与 LDA、CF 召回率对比

评估用户的召回率与其收藏文章数量关系

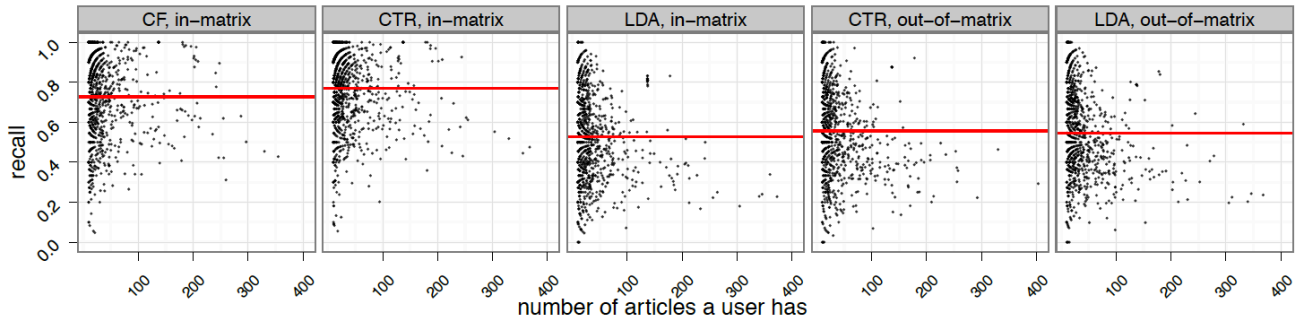


Figure: 一个用户收藏文章数量与召回率的散点图

- 两种问题预测结果都表明，有更多文章的用户预测召回率方差较小，文章少的用户容易产生评分的极值 0 或 1
- 文章很多的用户召回率有降低趋势，因为多阅读量的用户容易有不常见的文章，相对难以预测

评估文章的召回率与被收藏用户数量关系

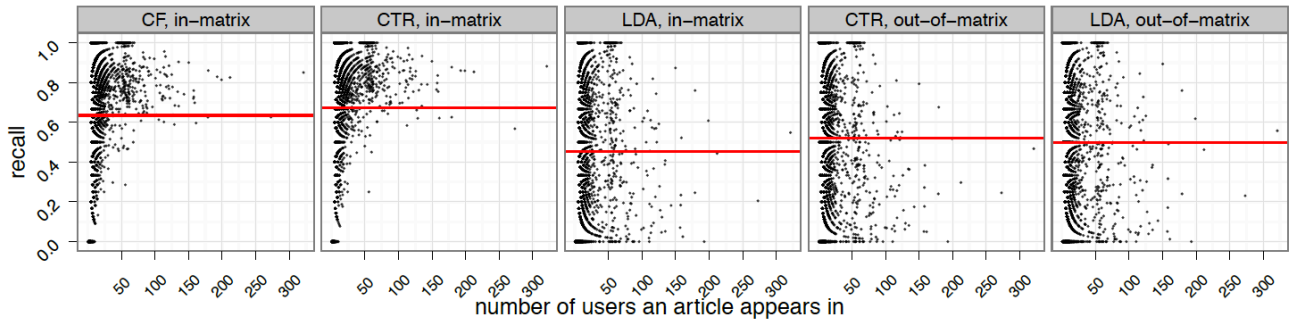


Figure: 一篇文章被收藏的用户数量与召回率的散点图

- 在 in-matrix 问题中，有更多用户的文章容易有更高的召回率，信息更多，更容易预测，在 LDA 方法中这个效应弱一些，因为其不用用户评分信息
- 在 out-matrix 问题中，由于都是新文章，提取不到用户信息，所以就没有上述效应

用户画像

- 通过 CTR 方法得到两个用户的偏好主题与推荐文章 (通过隐向量 u_i)

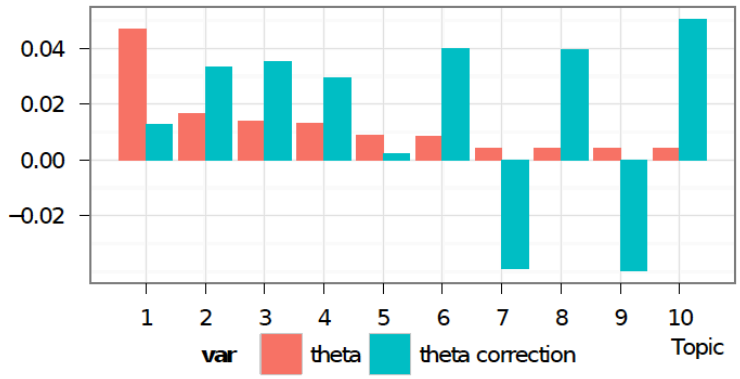
	user I	in user's lib?
top 3 topics	1. image, measure, measures, images, motion, matching, transformation, entropy, overlap, computed, match 2. learning, machine, training, vector, learn, machines, kernel, learned, classifiers, classifier, generalization 3. sets, objects, defined, categories, representations, universal, category, attributes, consisting, categorization	
top 10 articles	1. Information theory inference learning algorithms 2. Machine learning in automated text categorization 3. Artificial intelligence a modern approach 4. Data xmining: practical machine learning tools and techniques 5. Statistical learning theory 6. Modern information retrieval 7. Pattern recognition and machine learning, information science and statistics 8. Recognition by components: a theory of human image understanding 9. Data clustering a review 10. Indexing by latent semantic analysis	✓ ✓ × × × ✓ ✓ × ✓ ✓
	user II	in user's lib?
top 3 topics	1. users, user, interface, interfaces, needs, explicit, implicit, usability, preferences, interests, personalized 2. based, world, real, characteristics, actual, exploring, exploration, quite, navigation, possibilities, dealing 3. evaluation, collaborative, products, filtering, product, reviews, items, recommendations, recommender	
top 10 articles	1. Combining collaborative filtering with personal agents for better recommendations 2. An adaptive system for the personalized access to news 3. Implicit interest indicators 4. Footprints history-rich tools for information foraging 5. Using social tagging to improve social navigation 6. User models for adaptive hypermedia and adaptive educational systems 7. Collaborative filtering recommender systems 8. Knowledge tree: a distributed architecture for adaptive e-learning 9. Evaluating collaborative filtering recommender systems 10. Personalizing search via automated analysis of interests and activities	× ✓ × ✓ ✓ ✓ ✓ ✓ ✓ ✓

文章的隐空间 (偏离向量 ε 的影响)

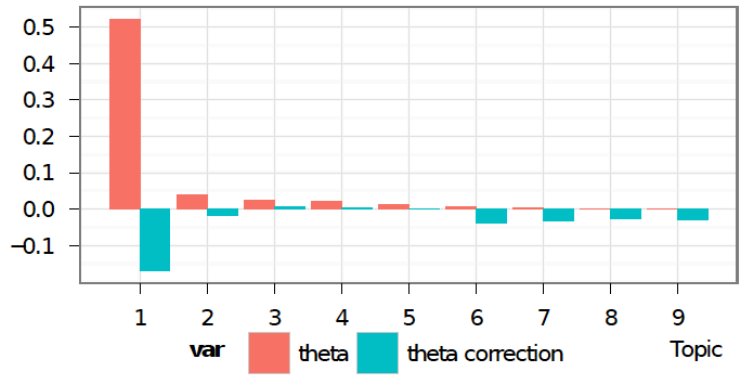
title	# dataset	# Google	avg–like	avg–dislike
1. The structure and function of complex networks	212	5,192	0.909	0.052
2. Emergence of scaling in random networks	193	8,521	0.899	0.058
3. R: a language and environment for statistical computing	113	837	0.827	0.047
4. A mathematical theory of communication	129	39,401	0.817	0.062
5. Maximum likelihood from incomplete data via the EM algorithm	157	22,874	0.864	0.055
6. A tutorial on hidden Markov models and selected applications in speech recognition	135	11,929	0.822	0.048
7. The structure of collaborative tagging systems	321	648	0.903	0.055
8. Why most published research findings are false	161	713	0.846	0.049
9. Phase-of-firing coding of natural visual stimuli in primary visual cortex	8	64	1.057	-0.004
10. Defrosting the digital library bibliographic tools for the next generation web	179	37	0.840	0.042

- 偏离向量 $\varepsilon_j^T \varepsilon_j = (v_j - \theta_j)^T (v_j - \theta_j)$，上表表示最大偏离量的 10 篇文章的概况和预测情况
- 表格第二三列为文章出现在数据中次数及 Google Scholar 中引用次数
- 这些偏移量大的往往引用量很高，是很大众化文章，这些文章常被不同领域人阅读
- 后两列表示喜欢和不喜欢的平均预测评分

文章的隐空间 (偏离向量 ε 的影响)



topic 1: estimate, estimates, likelihood, maximum, estimated, missing, distances
topic 10: parameters, Bayesian, inference, optimal, procedure, prior, assumptions



topic 1: neurons, responses, neuronal, spike, cortical, stimuli, stimulus

Figure: 主题参数分布图

- 左图表示偏移量大且很受普遍关注的文章，偏移量来自众多读者的偏好，隐向量会多出很多文本本身以外的主题。
- 右图表示不那么受普遍关注的文章也可以有较大偏移，但由于读者少，主要主题不会变，偏移量调整隐向量，一般不会推荐给文章主题爱好以外的用户

总结与展望

- 这篇文章整合矩阵分解方法和 LDA 主题模型预测推荐，后者预测物品隐向量
- 引入偏移量更好衡量内容的隐向量与该文章的隐向量，从而反映用户影响
- 作为一个传统方法，测试其他方法做 baseline
- LDA 表示文章主题仍有不足，所以后续引入深度学习来搞这个，参考 CDL 这篇文章 (Wang H, Wang N, Yeung D Y. Collaborative deep learning for recommender systems[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1235-1244.)