**Social Recommendation with Missing Not at Random Data心得体会**

此论文的标题，即基于缺失非随机假设的社会推荐。这篇文章，是发表在icdm（international conference on data mining）2018的一篇推荐算法相关的论文，此文的核心价值在于：论文中在社交推荐中引入缺失非随机假设（MNAR假设），并且提出了一种新的推荐算法（SPMF-MNAR）。

传统的、现存的社交推荐方法总是假设数据是随机丢失的（MAR），也就是认为用户购买物品的过程（或用户评分过程）是与用户评分是无关的，但是现实生活中很少存在这种情况。而由于现实的情况是用户通常会选择消费他们喜欢的或者他们朋友购买过的商品、对于一个用户，更多朋友消费该商品时，物品的评分会低于现有模型预测的分数。针对这一问题，此论文在基于两者相关、并不是完全随机缺失的新假设下，提出了对应的新的推荐模型--SPMF-MNAR推荐模型。

SPMF-MNAR推荐模型核心思想是对评分观察过程和评分同时建模。一方面SPMF-MNAR根据社交影响和他们个人喜好预测用户的消费情况（也就是评分观察过程），另一方面采用传统协同过滤模型，根据用户喜好去预测评分。

特别的，和其他专注于探究社交影响对用户评分影响的预测模型不同，此文章专注于探究社交影响对是否会评分的影响过程。并且使用SVI（随机变分intergerence ，即stochastic variational intergerence）接口去加速模型。

根据我对于推荐算法的了解，现在传统的推荐算法主要是基于以下三种基本模式，以经典的协同过滤算法、隐语义算法、基于图的推荐算法为例进行说明：

（1）Collaborative Filtering(协同过滤)算法：协同过滤算法主要分为两个步骤：寻找相似的用户集合，和寻找集合中用户喜欢的且目标用户没有的进行推荐。其中，寻找相似用户集合的具体实现方式为使用Jaccard公式计算Jaccard系数，随后通过计算皮尔逊相关系数得到用户兴趣的相似度，随后计算欧几里德距离和余弦距离，来表示不同用户间兴趣系数的差别完成聚类。随后，通过矩阵分解和ALS算法，计算商品的相似度并完成聚类，以此实现具体项目的推荐。

（2）隐语义（Latent factor model）推荐算法：核心思路是计算矩阵P和矩阵Q中参数值。一般做法就是最优化损失函数来求参数。但是隐语义模型在实际使用中有一个困难，那就是它很难实现实时推荐。经典的隐语义模型每次训练时都需要扫描所有的用户行为记录，这样才能计算出用户对于 每个隐分类的喜爱程度矩阵P和每个物品与每个隐分类的匹配程度矩阵Q。而且隐语义模型的训练需要在用户行为记录上反复迭代才能获得比较好的性能，因此 LFM的每次训练都很耗时，一般在实际应用中只能每天训练一次，并且计算出所有用户的推荐结果。从而隐语义模型不能因为用户行为的变化实时地调整推荐结果 来满足用户最近的行为。

（3）基于图(PersonalRank)的推荐算法：核心思路是将用户行为数据用二分图表示，将个性化推荐放在二分图模型中，用户数据是由一系列的二元组组成，其中每个元组(u,i)表示用户u对物品i产生过行为。那么给用户u推荐物品任务可以转化为度量Uv和与Uv 没有边直接相连的物品节点在图上的相关度，相关度越高的在推荐列表中越靠前。

虽然协同过滤等经典算法是非常流行的推荐算法模型，但是由于这些算法都是通过用户的历史反馈预测用户喜好和生成推荐。但是现实生活中，人们只会消费或者评分很少部分的物品，因此传统协同过滤算法在实用时存在着稀疏问题，而社会计算的提出就是为了解决这个问题。也就是，在推荐算法提出在引入用户的社交信息，这些方法通过对社交关系对用户评分的影响去证明基于社交的协同过滤模型有改善推荐效果的潜力和可能性，增加推荐系统对于用户数据的获取量以解决缺失问题。

但是，现存社交推荐算法统采用信息随机丢失（MAR）假设，也就是说用户购买物品的过程（或用户评分过程）是与用户评分是无关的。事实上，MAR假设不适用于推荐数据，因为用户通常会购买他们喜欢的东西。相关的，通常高评分通常在评分矩阵中，低评分经常缺失。当MAR假设不成立时，之前的社交推荐算法在预测上就存在一定偏差，并且导致推荐效果受损。

而为了解决这个问题，就在这篇论文里提出：建立新的假设--缺失非随机假设（MNAR假设）。这篇文章采用了Epinions和Ciao数据集作为社交推荐分析和实验的数据。

Epinion和Ciao是两个知识分享的网站，由于这两个数据集中，除了包含传统的评分信息、物品购买信息和留言信息之外，这两个数据集还包含了社交信息，每个用户在网站中都有一个朋友列表，这个列表可以直接用来构建一个用户的社交网络。

根据数据分析，得出来了四个数据分布特点：

1. 用户倾向于消费（评分）他们喜欢的物品、
2. 用户倾向于消费（评分）他们的朋友消费过的商品、
3. 朋友对用户消费的影响大于对评分的影响、
4. 消费了这个商品的朋友越多，用户对该商品的评分越低

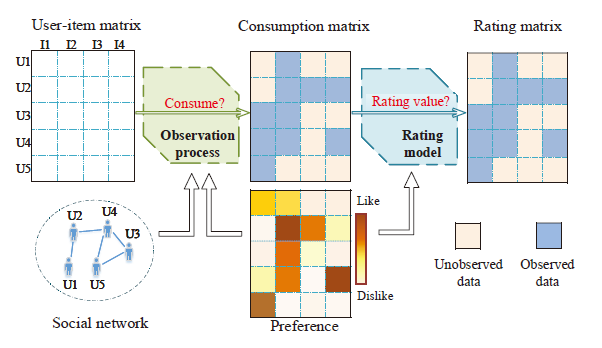
而论文根据这四个特点得到了两个结论：

1.在社交推荐中应用MNAR模型是有益的。一方面，基于对数据分析结果的观察，可以得出结论评分数据不是随机丢失的这个特点。特别的，用户倾向于消费他们喜欢的物品，并且低评分通常是缺失项：在处理MNAR数据但是并不基于MNAR的假设时，推断会变差，预测的准确性也会受损。另一方面，用户消费更容易收到社交关系影响。所以说研究社交关系对消费的影响可以更准确的研究用户真实的喜好。

2.用户的消费情况可以认为收到用户个人喜好和社交影响双方面的影响：一方面，喜好应该发生在选择之前。当用户决定要不要购买某个商品时，他们会衡量他们时候会喜欢这个商品。另一方面，在社交网络中，我们都属于内容共享的社区。用户会被显式链接，并且会有意无意向他们的朋友推荐商品或者服务。用户朋友购买过的商品更容易引起用户的注意力，并且会存储进情景记忆中，这些都会促进用户的消费。和朋友中流行的商品相比，用户对于那些不流行商品的消费更多是基于他们自己的喜好。因此，当物品被更多朋友分享消费时，评分会下降。

而为了证明这个假设是正确的，且为了将这个假设投入使用，论文中提出了一种新的假设来证明新的假设是合理的、可用的，就是此文提出的SPMF-MNAR推荐模型。

和现有社交推荐方法根本性的不同是，SPMF-MNAR推荐模型中要在社交推荐中引入MNAR假设，并且对评分的观察过程进行建模。一方面，根据社交影响和用户喜好生成了他们的消费矩阵。（逻辑如下图）



另一方面，采用传统的CF模型对用户消费过的商品做评分预测。

而这个推荐模型做的假设有以下四点：

1. 用户对物品的喜好可以通过概率矩阵分解生成。对于每一个用户物品对，生成一个隐藏的连续变量zij。这个变量刻画了用户i对物品j的喜好。并且这个关系假设满足高斯分布。

2. 对于每一个元素xij​∈X,有βi0​的概率购买行为是因为用户自己的喜好的概率 xij​∼p(xij​∣zij​,di​,c) ,有βik​的概率是被他第k个朋友影响 xij​∼p(xij​∣xgki​j,η​) .这里βik​ 刻画了针对于每个人的社交影响的重要性。其中βI​=[βi0​,βi1​,…,βi∣gi∣​]，且满足∑k=0∣gi∣​βik​=1，服从狄利克雷分布（即多元β分布），参数为α。

3. 当xij​是通过用户个人喜好生成的时候，即τij0​=1时，可以使用伯努利对数模型可以用来描述用户趋向于购买他们喜欢的商品。引入偏移量di​定义了每个用户的接受阈值：当用户对一个物品的喜好大于接受阈值时，即zij​>di​，用户就会倾向于购买某个物品，当用户对一个物品的喜好小于接受阈值时，即zij​<di​时，用户就会倾向于拒绝购买某个物品。同时，引进比例参数c，c > 0，去控制用户喜好性zij​对消费行为xij​的影响。当比例参数c值变大时，用户会对zij​更敏感。

4. 当用户的购买行为xij​是他第k个朋友影响生成的时候，τijk​=1，可以对xij​进行伯努利分布和他第k个朋友的购买情况xgik j​的伯努利参数进行建模。

在以上假设成立的情况下，具体的推荐算法执行步骤是：首先计算朋友对消费的影响：η0​∼Beta(ζ), η1​∼Beta(ζ)。随后，对每个用户i而言，计算偏差ai​∼N(0,σa2​);和潜在喜好ui​∼N(0,σu2​I)与计算接受的阈值（接受阈值）di​∼N(0,σd2​)再计算不同因素的重要性βi​=Dir(α)。而对每个物品j而言，先计算偏差bj​∼N(0,σb2​);再计算潜在价值vj​∼N(0,σv2​I)。得到用户和物品的信息后，对于每个物品和用户的组合对(i,j):先计算用户对物品的喜好zij​∼N(uiT​vj​+ai​+bj​,σz2​)再计算元素对消费的贡献τij​∼Multinomial(βi​)。当τij0​=1时，说明消费受到了用户喜好的影响xij​∼Bernoulli(xij​∣σ(c(zij​−di​)))。当τijk​=1(1≤k≤∣gi∣)时，则说明消费是受到第k个朋友对用户的影响，即xij​∼Bernoulli(xij​∣η1​xgki​j​+η0​(1−xgki​j​))。而对于每一个观察的评分rij​则是根据用户的喜好推测用户评分rij​∼N(zij​,σr2​)。

随后，对论文中提出的推荐模型在两个真实数据集Epinions和Ciao进行实验：实验的指标是MAE 平均绝对误差、RMSE 均方误差和NDCG(评价排序质量)。对比的基线模型是Useravg和PMF、对比的其他模型包括不采用MNAR假设的社交推荐算法:SoRec，RSTE，SocialMF，TrustMF，TrustSVD，这些都是社交推荐中不使用MNAR假设的最优的模型；还对比了采用MNAR假设的非社交推荐算法最优模型。包括GPMF和论文作者自己设计的SPMF-MNAR-nos模型作为比较去验证对社交关系建模对评分观察过程的影响。（SPMF-MNAR-nos是SPMF-MNAR的简化版，排除了社交影响，只通过用户喜好生成用户购买情况）

最终的实验结果表明，无论从MAE和RMSE的衡量标准上讲还是从NDCG的标准上讲。SPMF-MNAR模型在所有数据集上都优于其他模型。SPMF-MNAR模型比其它方法的改善主要体现在两个方面，

（1）在社交推荐中的评分不是随机丢失的，如果方法采用传统的MAR的假设，忽略用户是否消费这个物品和用户评分的联系，预测结果会受损。

（2）推荐系统中融入社交信息。和以往探究社交影响对用户评分的研究不同，SPMF-MNAR模型致力于探究社交影响对用户评分过程也就是是否会产生购买行为进行研究，并且观察它如何改善推荐效果。

事实上，在数据集上的分析表明在内容共享的社交网络中，相互连接的用户通常会不自觉的相互推荐，因此他们的消费会被社交关系强烈影响。所以，SPMF-MNAR的推荐效果比不应用社交关系的SPMF-MNAR-nos模型更好。

**成绩∶97**

**指导教师签名：吴共庆**

2019年5月18日

改进思路：

此文章在现在有的社会计算的推荐思路的基础上提出了新的要求，即考虑了用户的社交关系对于用户的行为产生影响的同时，不再是认为评分与购买行为无关，即引入MNAR假设，并且对评分的观察过程进行建模。一方面，根据社交影响和用户喜好生成消费矩阵。并采用协同过滤算法进行评分预测。

此文章的意义是毋庸置疑的，但是并非没有改进的必要和空间：

比如，对社交影响建模可以进一步提升推荐效果。这篇文章中只探究社交影响对用户评分过程（也就是购买行为）的影响，并探究它如何提升推荐效果。事实上为了探究基于社交模型的影响效果，完全可以用稍微复杂的模型，例如论文：‘’A matrix factorization technique with trust propagation for recommendation in social networks‘’中提出的SocialMF社交分析模型或者论文‘’Social collaborative filtering by trust‘’中提出的TrustMF社交分析模型。以提高对于社交分析对数据特征提取的效果。（这一点也正如论文中“结论”部分提到的“探究更多会影响用户消费的因素，比如物品流行性和用户位置”）

本文最重要的不足就是对用户消费过的商品做评分预测的环节采用了传统的CF模型。CF模型的种种特性虽然很好，但是绝非到现在还先进。比如，本文的协同过滤的本质是基于用户的协同过滤，User-based的CF算法存在两个重大问题：

 1. 数据稀疏性。一个大型的电子商务推荐系统一般有非常多的物品，用户可能买的其中不到1%的物品，不同用户之间买的物品重叠性较低，导致算法无法找到一个用户的邻居，即偏好相似的用户。

这个问题已经使用社会计算的方式解决了，这算是解决了一个问题。

 2. 算法扩展性。最近邻算法的计算量随着用户和物品数量的增加而增加，不适合数据量大的情况使用。这样做在处理大规模数据的时候效率明显会非常低，因为我在大学生创新项目中实现过CF算法，因此深有体会。

可以考虑通过改进基础的CF算法可以解决这个问题：因为物品直接的相似性相对比较固定，所以可以预先在线下计算好不同物品之间的相似度，把结果存在表中，当推荐时进行查表，计算用户可能的打分值，可以同时解决上面两个问题。

当然，也可以直接用更先进的算法来替换掉User-base的CF算法，比如2017年提出的MLR算法，当然，做出这种改进的话，需要对代码进行较大规模的重构，但我认为理论上是可行的。