基于特征工程与协同过滤的餐饮推荐

**文献综述**

**一. 学术界文献回顾**

作为立于计算机科学、统计数学、社会工程基础等知识之上的交叉程度极高的子领域，推荐系统相关的文献也角度甚广，主要分为学术界模型解释性的文献、工业界效果导向性的文献两大类。在具有强社交关系的数据集上，比如Yelp和Netflix，来自好友的推荐经常是十分重要的，也使得学术界对于CF算法的重视程度超乎以往；而因为该类数据集的超稀疏性，基于记忆的方法受制于很难高效率地找到近邻，因而基于模型训练的方法研究更加活跃，涌现出一系列基于用户互信关系网络（Trust Relationship Network）的矩阵分解和隐特征方法。这类方法一般基于统计学习。

每一种推荐系统的学习模型都有其假设作为前提，比如基于用户的CF方法假设如果两个用户过去对产品有相似的喜好，那么他们现在对产品仍有相似的喜好；基于产品的CF方法假设如果某个用户过去喜欢某种产品,那么他（她）现在仍喜欢与此产品相似的产品。以往在很多推荐模型中，都假设用户满足“i.i.d.”即互相独立、均匀分布（Independent and Identically Distributed）；而这与现实中用户经常互相询问餐馆或电影的推荐的直觉不符。Ma H. et al. [1] 挑战该假设，提出了一种基于概率矩阵分解（PMF）、融合用户-产品的打分矩阵与用户之间的社交网络结构考虑的方法，称之为*SoRec*。*SoRec*不但在预测效果上相比一般的PMF和CPMF（Constrained-PMF）有明显提升，还充分利用了矩阵的稀疏性，在数据量很大的情况下复杂度随着评分数量保持线性增长。其主要思路是使用隐变量模型（Latent Factor Model，简称LFM），将两个不同的数据源之间共享的用户数据抽象为低维的隐变量空间，对两个矩阵同步应用PMF来分别求出用户、评分和社交网络的隐因子空间U、V和Z，便可通过求 和 来补全原矩阵，作出评分的预测。*SoRec*包括一个新的梯度下降解法。另外，作者还强调了用户互信网络的非对称性，以及从它构建互信矩阵时，要考虑采纳用户的影响力来正则化（Regularize）其建立的关系的权重；例如，当某用户关注许多用户时他的关注（社交关系中的边）权重就会下降，而当某用户受到很多关注时，他关注别人的行为就会增加权重。

Guo G. et al. [3] 亦提出一种融合考虑明确的社交互信关系来形成推荐的CF方法，该方法认识到了能够利用社交网络关系的推荐系统（Trust-aware Recommender System，TARs）对于自举（Bootstrap）的意义，并在冷启动上表现不俗。简要步骤如下：（1）根据互信关系的强弱，将某活跃的目标用户的直接好友（Directly Trusted Neighbors）们共同点评过的商铺上的评分求加权均值；（2）将上述评分集作为目标用户口味偏好的一个代表，根据相似度计算方法，为其寻找到更多有类似口味的用户；（3）综合考虑好友和相似用户的性质来产生推荐，实质上就是利用了显性互信关系和隐形互信关系这两层信息。类似的思路Ma H. et al. 在 [2] 和 [12] 中也有推广。

[5] 则在更高的战略层面关注如何将额外的关系信息纳入CF和矩阵分解方法中，因为Yelp等现实数据集中包含了太多实体与实体之间的联系，若每研究一组联系的内在规律都需要重新定制一套方法，显然是费时费力的。作者认为推荐问题，乃至于对于实体之间关系的预测，都可以抽象为关系数据集的矩阵填补问题（[10] 也支持这一观点），可以找到一个隐变量空间，它被所有纳入考虑范围的、有观测数据的每个实体（如用户、商铺、类别等）所共享，并采用 [13] 中的多矩阵同步分解方法对该建模做数学处理。下图显示了这一概括性的模型：



经过在Yelp数据集上的实体选择和简单修改，它可以综合考虑多种来源的信息（实体），对多组关系作预测，以及支持没有直接的互相联系的实体之间的相似度计算。

**二. 工业界技术报告摘要**

与学术界的偏好不同，工业界的推荐系统技术报告和Yelp Dataset Challenge偏好的文献多数并不在精巧的模型或预测的可解释性上下工夫，反而经常更注重数据预处理、特征提取和特征选择，然后对比使用多种较为简单而常用的的机器学习模型来检验效果，从中选择一个较为合适的，甚至为了提升效果对多种模型组合进行Ensemble。常用的模型包括各种简单贝叶斯分类器（Naive Bayes），逻辑回归分类器（Logistic Regression Classifier），支持向量机（Support Vector Machine），随机森林（Random Forest）和梯度提升决策树（Gradient Boosting Decision Tree，简称GBDT）等。

目前公布的前五轮大奖成果中的11篇文献里，有7篇都是自然语言处理（Natural Language Processing）和文本挖掘（Text Mining）相关领域的研究与应用，其中5篇有关话题模型（Topic Model）或情感分析（Sentiment Analysis）等；有鉴于此，可以认为在餐饮数据集上的NLP已经被探究得相对比较透彻。而无论是强调特征工程的传统CBF，还是基于模型与社交网络关系的CF方法，在该类数据集上的实际部署和调优还比较少见。

Finn L. [7] 建议不要太过关注精巧模型中的数学技巧而忽略了那些直觉上对模型的预测准确度影响最大的因子：用户的打分主要还是由他自身评分习惯与商铺平均分值来综合决定，而来自朋友的影响只能位居次席。通过对数据的仔细分析，文中指出了Yelp中社交关系的若干特点：（1）只有0.004%的信任或关注关系不是双向的，即若用户i在用户j的朋友列表中，基本上反之亦然，也即可以认为Yelp数据集的互信网络是无向性质的，而少数不符合的情况应该归咎于发布公开数据集前对它进行扰动造成的小瑕疵；（2）超过一半的用户只点评或浏览点评而不参与任何社交关联，而其余用户中的96.7%竟然可以通过互信关系连结在一起，最后剩下的4K用户互相之间结成2-3人的小团体互相关注，不与外界来往；（3）如同许多文献中提到的那种，再一次验证了社交网络中用户的朋友数量基本符合幂律分布（power law）。本文第一部分提出并检验了一些新的特征在逻辑回归中的表现，其结果证明不同类别的商铺在同一算法下，与基准测度的准确率偏差比较大，因此根据类别信息来微调分类器的参数或特征是很有必要的。

Keven Hung. et al. [6] 对Yelp数据集进行了深入的特征探索，并使用了若干精巧的可视化手段，如混淆矩阵（Confusion Matrices）等，发掘了其中蕴含的一些重要规律：（1）评分的分布明显向四至五颗星的好评倾斜，它们的比例占据超过80%；（2）随着网站建立以来，评分的均值稳步缓慢地趋于乐观（Optimistically Biased），方差（Variance）越来越大，而早期用户的评分相对要更高一些。据此，该技术报告提出计算残差和（Residual Sum，即RSS）对算法的评估指标Precision与Recall作校准，其中RSS即为方差的非归一化表达，规范化定义如下：

该校准方法有充足的必要性和可解释性：例如，在一个极端下，若预设几乎所有的评论都是5星的先验， SVC分类器很有可能试图将所有的评论都预测为5星以最大化准确率（Accuracy）；在这种情况下，5星的RSS准确率和召回率都会非常高，而相对地其他评级的准确率和召回率都会很低，使得总的RSS较大，显示了这种高偏差的分类方法对于不同评级分类准确率的高差异。经实验发现，2星的评论尤其难以预测对，1、2、3星的评价都经常被误分类为4星好评；而仅能准确预测好评（4-5颗星）的分类算法并不是我们需要的。因此实验团队对中差评采用了过采样（Oversampling）方法来减小训练数据的偏差（Bias），事实证明对结果有相当程度的改善。该实验也有力地说明了基于特征工程与机器学习的推荐算法对训练数据集的敏感，与此相反的是基于矩阵分解的方法对训练数据的bias并不敏感，对负样本没有硬性需求。另外，实验团队认为虽然用户历史平均评分值与商铺历史获得平均评分值这两项特征对于预测结果最相关，设计算法时却应该基于鲁棒性（Robustness）的考虑尽量避免使用它们，以免在遇到冷启动问题时束手无策，比如当用户-商铺对中用户没有评分过、商铺没有收到过评分时或两者皆无时，这两项特征是无法获得的，也不宜用默认值去填补。

Sawant S. [9] 提出将Yelp数据集表达为一个加权二部图（Weighted Bipartite Graph）来压缩用户-产品评分矩阵的信息，进而将推荐问题抽象为图映射（Graph Projection）问题的做法，以缓解其稀疏性对CF中典型的基于记忆的方法（例如K-近邻算法）造成的计算困难和精度问题；其中二部图中的边只能从用户集连接到商铺集，与边关联的权重则是评分。文中通过定义每位用户各异的 “推荐力” （Recommendation-power或称Resource）来计算用户之间的相似度（Similarity）：

其中 为用户u对用户v的推荐力，也即u的推荐对v造成选择影响的程度； 是该用户给出过的评分之和，而 是商户收到过的评分之和；B是一个包含了所有用户u和用户v均评分过的商铺b的集合。通过该相似度度量方法，可以将原本十分稀疏的评分矩阵用K-Means聚类（Cluster）方法进行压缩，使得每一个聚合用户组或聚合商铺组之间产生更多的边，缓解任意两个用户都没有光顾过许多相同的餐馆的问题；通过绘制在设置产生不同数量的用户聚合组和商铺聚合组时对应的RMSE和MAE曲线，可以最优化参数的选择，结果显示这种相似度的度量方法能够取得比PCC和Cosine更好的效果。

**三. 存在问题**

总体而言，学术界侧重CF的研究更关注精巧的、可解释的矩阵分解模型等技术，而在工业界、KDD Cup等实际应用场景上，被证明更有效果的反而是大量的特征抽取、特征构造和特征选择技术，再加上合理的常用机器学习模型的选择、参数的调整，获得比较理想的推荐结果和较高的运行效率。无论是矩阵分解中着重挖掘的用户-产品评分矩阵与社交网络结构中蕴含的关联信息，还是实际业务场景中来之不易的丰富数据特征，实际上都左右着对推荐效果，很难说其中哪一项的影响要远超过另一项；因此，寻找合适的方法将它们结合考虑始终是非常必要的。

**四. 参考思路**

与本研究目的最相关的要属Chen T. et al. [11] 所阐述的 “基于特征的矩阵分解” 思路，能够将特征工程的结果与用户-产品评分矩阵中的隐维度结合考虑。文中提出了一种对用户评分的建模：

前两项是考虑显性本身对于评分的作用（用户对于所有物品的打分的平均分，物品收到的平均打分）；而一对用户-产品（User-item Pair）对应的特征在隐空间（Latent Space）分解后相乘（上述公式的第三项）就代表了隐因子空间各处的评分的预期。至于社交网络关系，也可以直接认为是一种特征，稍做处理就可以灵活地纳入训练模型。

文献主要作者，来自上海交通大学Apex实验室的陈天奇等人，据此基于C++开发了一套支持单机运算大规模矩阵分解（SVD++）的工具集SVDfeature，用于带附加特征的CF方法；它使得工程师可以集中精力负责特征工程，而不再需要太关注模型的效率、可扩展性和如何利用关系数据的问题。利用SVDfeature Toolkit，作者的团队在2011年、2012年的KDD Cup分别取得了总成绩第三和第一。然而自此以后因为开发XGBoost、cxxnet和mshadow等库使得SVDfeature疏于维护，而因为Mac OS X不支持Pthread的匿名信号量（Anonymous Semaphore）使得Toolkit的流水线（Pipeline）出错，从10.7一直到10.11也没有解决无法编译的问题。SVDfeature虽然很难纳入各Python科学计算工具集进行使用，但它的思路依然对本研究的手段与目的很有启发意义。

**五. 引用文献列表**

1. Ma H, Yang H, Lyu M R, et al. Sorec: social recommendation using probabilistic matrix factorization[C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 931-940.

2. Ma H, King I, Lyu M R. Learning to recommend with social trust ensemble[C]//Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009: 203-210.

3. Guo G, Zhang J, Thalmann D. A simple but effective method to incorporate trusted neighbors in recommender systems[M]//User modeling, adaptation, and personalization. Springer Berlin Heidelberg, 2012: 114-125.

4. Yang B, Lei Y, Liu D, et al. Social collaborative filtering by trust[C]//Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press, 2013: 2747-2753.

5. Gupta N, Singh S. Collective Factorization for Relational Data: An Evaluation on the Yelp Datasets[R]. Technical report, Yelp Dataset Challenge, Round 4, 2015.

6. Kevin Hung, Henry Qiu. Oversampling with Bigram Multinomial Naive Bayes to Predict Yelp Review Star Classes[R]. Technical report, Yelp Dataset Challenge, Round 4, 2015.

7. Finn L. Link Prediction in the Yelp Social and Review Networks[J].

8. Carrillo N, Elmaleh I, Gallego R, et al. Recommender Systems Designed for Yelp. com[J]. 2013.

9. Sawant S. Collaborative filtering using weighted bipartite graph projection: a recommendation system for yelp[C]//Proceedings of the CS224W: Social and Information Network Analysis Conference. 2013.

10. WU J L. Collaborative filtering on the netflix prize dataset[J]. 2010-08-01]. http://dsec, pku. edu. cn/-jinlong, 2010.

11. Chen T, Zheng Z, Lu Q, et al. Feature-based matrix factorization[J]. arXiv preprint arXiv:1109.2271, 2011.

12. Ma H, Zhou D, Liu C, et al. Recommender systems with social regularization[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 287-296.

13. Singh A P, Gordon G J. Relational learning via collective matrix factorization[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008: 650-658.