**Meta-Graph Based Recommendation Fusion over Heterogeneous Information Networks**

# ABSTRACT

异构信息网络（HIN）是现代大型商业推荐系统中数据的自然和一般的表示形式，涉及不同类型的数据。基于HIN的推荐人面临着两个问题：**如何表示推荐的高层次语义**，如何**融合异构信息提出建议**。在本文中，我们通过首先将**元图**的概念引入基于HIN的推荐，然后用**“矩阵分解（MF）+因数分解机（FM）”方法来解决信息融合问题**，从而解决了这两个问题。对于每个元图生成的相似性，我们执行标准MF来为用户和项目生成潜在特征。基于不同的基于元图的特征，我们建议使用FM组与套索（FMG）自动学习观察评级，以有效地选择有用的基于元图的特征。两个真实世界的数据集，亚马逊和Yelp的实验结果，显示了我们的方法相比，先进的调频和其他基于HIN的推荐算法的有效性。

# INTRODUCTION

像亚马逊或Yelp这样的平台上的建议是指向用户推荐诸如产品或业务的项目的问题，以便当用户消费更多的项目时平台可以获得更多的收入。**本质上，如果我们将用户和项目视为二分图，这是对不同类型实体（即用户和项目）的链接预测问题。**如今，大型商业推荐系统融合了更丰富的异构信息。例如，在亚马逊，产品有类别，或者可以属于品牌，用户可以将评论写入产品。在Yelp上，用户可以跟随其他用户形成社交网络，基于位置的业务可以分类，用户也可以向企业写评论。真实世界的推荐系统需要考虑由平台启用和收集的不同类型信息的更丰富的语义。更丰富的异质性因此需要开发一个数学表达式来表达它，并在其上进行计算。

已经提出异构信息网络（HINs）[30]作为许多不同类型的数据，如学者网络数据[32]，社交网络数据[13]，病人网络数据[3]等的一般数据表示。或知识图形数据[5]。在开始时，HIN被用来处理实体搜索和相似性度量问题[32]，其中查询和结果实体被假定为相同类型（例如，使用Person来搜索Person）。后来，它被扩展到处理异构实体推荐问题（即向用户推荐项目）[29,39,40]。为了融入丰富的语义，HIN首先构建了异构网络的网络模式。例如，对于Yelp，网络模式定义在实体类型User，Review，Word，Business等等上。在实体类型约束下的语义相关性可以通过沿着元路径的两个实体之间的相似性来定义[32]。对于传统的协同过滤，如果要向用户推荐企业，我们可以建立一个简单的元路径业务！用户，并从这个元路径中学习进行概括。从HIN的模式中，我们可以定义更复杂的元路径，比如*User* → *Review* →*Word* →*Review* →*Business*。元路径是一种相似性，用于衡量用户是否倾向于喜欢一个企业，如果他/她的评论类似于其他用户针对同一企业的评论。

当将基于元路径的相似性应用于推荐系统时，存在两个主要挑战。首先，**元路径可能不是表征丰富语义的最佳方式**。图1显示了一个具体的例子，其中一个元路径*User* → *Review* →*Word* →*Review* →*Business*是用来捕捉用户的相似性，因为他们都写评论，并提及它所服务的海鲜。但是，如果我们要捕捉U1和U2对同一类业务（如餐厅）进行评级的语义，同时提及同一方面（如海鲜），则元路径失败。我们需要一个方法来捕捉这样复杂的语义。最近，Fang等人[6]和黄等人[10]已经提出使用元图（或元结构）来计算同类型的实体之间的相似性（例如，使用Person来搜索人）而不是HIN，其可以捕获更复杂的元路径不能的语义。但是，他们没有用元图探索异构类型的实体。在本文中，我们将这个想法扩展到异构推荐问题。然而，如何利用HINs推荐的异构类型实体之间的相似性还不清楚，导致第二个挑战。

其次，不同的元路径或元图形导致不同的相似性。**如何以有效的方式组装它们是另一个挑战**。目前有两个原则性的方法。考虑到我们的**目标是实现用户和项目评分的准确预测**，可以将其制定为用户项目评分矩阵的矩阵完成问题。**预测基于HIN的缺失评分的一种方法是使用元路径为用户项目矩阵生成许多特设的替代相似度，然后学习不同元路径的加权机制以组合相似度明确地近似用户项目评分矩阵[29]**。方法不考虑每个元路径的隐含因子，而且每个可选的相似性矩阵对于最终的整体集合都是非常稀疏的。**另一种方法是首先对每个元路径计算出的每个用户项目相似度矩阵进行因式分解，然后使用潜在特征恢复一个新的用于项目集合的用户项目矩阵[40]**。方法解决了每个相似矩阵的稀疏性问题。然而，它并没有充分利用潜在的特征，因为当进行集合时，每个元路径不能看到别人的变量，而只能看到其他变量预测的单个值。

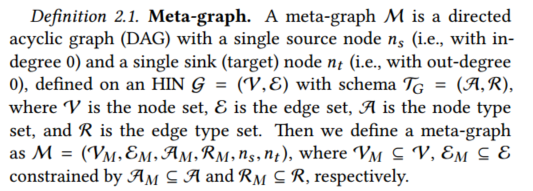
为了解决上述挑战，我们提出了一个新的原则性的方法来充分结合不同的潜在特征。首先，我们不是将元路径用于异构推荐[29,40]，而是将**推荐问题的元图的概念引入到推荐问题中**，这使得我们可以将更复杂的语义纳入到我们的预测问题中。其次，不是直接计算恢复的矩阵，而是**使用所有元图的所有潜在特征**。受著名PCA + LDA用于人脸识别[2]的启发，**首先采用PCA（主成分分析）进行无监督降维，然后应用LDA（线性判别分析）发现进一步缩小维度，我们将矩阵分解（MF）+分解机（FM）[25]应用于我们的推荐问题。**对于每个元图，**我们首先在元图的引导下计算用户项相似度矩阵**，**然后使用无监督的（**without seeing the ratings**）MF将其分解成一组用户和项目的潜在向量**。**从许多不同的元图形计算出来的用户和项目向量的不同集合**，我们使用FM来组装它们以从评价矩阵中学习。为了有效地选择有用的元图，我们建议使用**FM with Group lasso（FMG），即“l2, 1-范数正则化”来学习参数**。通过这种方式，我们可以自动确定一些新的问题：**1.哪一种元图应该被使用；2.为每个元图生成的用户和项目向量；3.确定应该如何加权。**在Amazon和Yelp这两个大型实际数据集上的实验结果表明，我们的方法可以成功地胜过其他基于MF，基于FM和现有HIN的推荐技术。我们的代码可以在//github.com/HKUST-KnowComp/FMG上找到。

# 2 FRAMEWORK

## 2.1 Meta-graph based Similarity

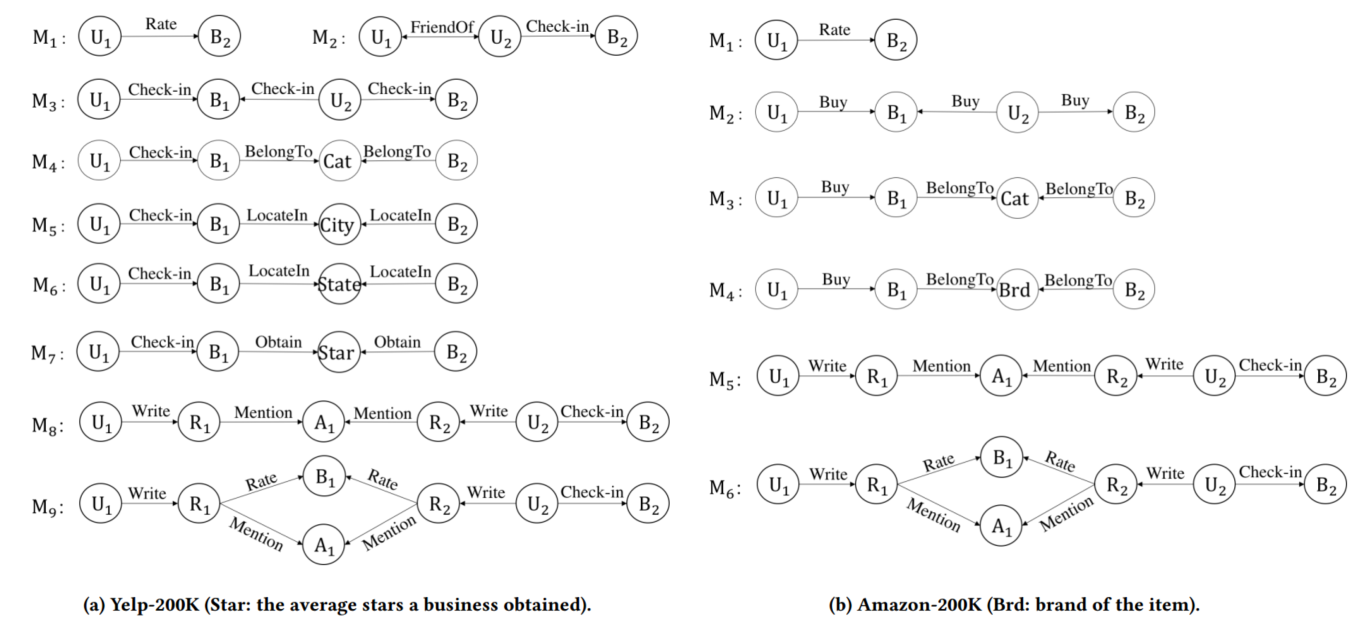
HIN和HIN Schema（一个实体类型的模式图及其关系）的引入已经在[32]中被引入。这里我们跳过正式的定义，只说明图1中的原始HIN和图2中相应的模式。这里我们重点介绍与本文相关的概念。首先，我们正式定义HIN中的元图来进行推荐。

Definition 2.1. **Meta-graph:有向无环图**



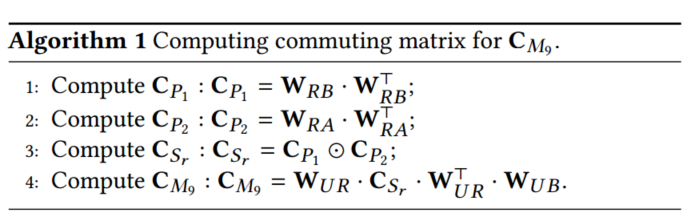
我们在图3中显示了亚马逊和Yelp在本文中使用的所有元图。我们可以看到它们是以U（用户）作为源节点和B（用于Yelp和Amazon的产品的业务）作为目标节点的DAG。这里我们使用Yelp数据上使用的M3和M9来说明计算问题。

给定上述的元图定义，我们要计算源节点和目标节点之间的相似度。本来，**交换矩阵[32]已被用来计算元路径的基于计数的相似性矩阵**。这表明基于计数的元路径的相似性可以通过矩阵序列的乘法来计算。在实践中，如果邻接矩阵W是稀疏的，我们可以非常有效地实现它。



对于元图，问题变得更加复杂。例如，对于图3（a）中的M9，有两种方式可以通过元图，即（U; R; A; R; U; B）和（U; R; B; R; U ; B）。

请注意，R表示HIN中的实体类型Review。这里在路径（U; R; A; R; U; B）中，（R; A; R）意味着如果两个评论都提到相同的A（方面），那么它们有一些相似之处。类似地，在（U; R; B; R; U; B）中，R; B; R意味着如果两个评论都评价相同的B（商业），那么他们也有一些相似之处。我们应该定义我们的相似性逻辑，当一个flow通过元节点从源节点到目标节点有多种方式时。当有两条路径时，我们可以允许一个流经任一条路径，或者我们约束一个flow满足他们两个。通过分析以前的策略，我们发现类似于将这种元图分解成多个元路径，然后采用我们后来的计算。我们选择latter，这需要比简单乘法多一个矩阵运算，即Hadamard乘积或元素乘积。算法1描述了计算图3（a）中M9基于计数的相似度的算法，其中是Hadamard乘积。在获得CSr的情况下，通过乘以一系列矩阵来获得整个可交换矩阵CM9是比较容易的。在实践中，不仅限于图3（a）中的M9，本文定义的元图可以通过对应矩阵的两个运算（Hadamard积和乘法运算）来计算。



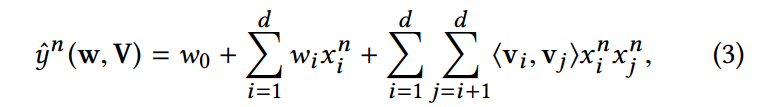
## 2.2 Meta-graph based Latent Features

然后我们得到L个不同的用户 - 项目相似度矩阵，我们使用**矩阵分解**来获得用户和项目的潜在特征，以减少噪声和原始相似矩阵的稀疏性问题。最先进的MF技术可以用于任务[14,22,36]。基于用户偏好受少数因素控制的假设，相似度矩阵R可以分解为两个低秩矩阵U和B，它们分别表示用户偏好和项目的潜在特征。通过解决（1）中的优化问题，可以获得用户和项目的低维表示。

## 2.3 Recommendation Model

在后面，我们获得了L组用户和项目潜在特征，对于观察到的评级中的样本xn，即一对用户和项目，由ui和bj表示，我们连接所有相应的用户和项目特征 所有的L元图：

请注意，对于不同的矩阵，F可能不同，但为了简单，我们保持恒定。xn表示第n个样本的连接的特征向量。每个用户和项目可以分别由L×F个潜在特征表示。



将FM模型应用于基于元图的潜在特征时存在两个问题。第一个问题是在处理多个元图时会带来噪声，从而影响模型的预测能力。而且，在实际中，由于某些元路径提供的信息可能被其他元路径覆盖，因此一些元图可能是无用的。第二个问题是计算成本。所有这些特征都是由标准矩阵分解生成的，这意味着设计矩阵，即馈给FM的特征是密集的。它增加了学习模型的参数和在线推荐的计算成本。为了缓解上述两个问题，我们提出了一种新的正则化FM，即集合套索正则化[42]，这是一组变量的特征选择方法。参数p的组套索正则化如下：

# 5 RELATED WORK

## 5.1 Heterogeneous Information Networks

已经提出HIN作为许多真实世界图或网络的一般表示。元路径被开发为由HIN网络模式定义的一系列实体类型。基于元路径，已经提出了几种相似性度量，例如PathCount [32]，PathSim [32]和Path Constrained Random Walk [16]。已经证明这些措施对于许多实际网络中的实体搜索和相似度测量是有用的。随着元路径的发展，许多数据挖掘任务已经被启用或增强，包括推荐[29,39,40]，相似搜索[27,28,32]，聚类[33,35]，分类等。 [1，12，18]和链接预测[31,43]。最近，已经提出了元图（或元结构）来解决HIN中更复杂的语义[6,10]。在实体被限制为相同类型的情况下，仍然将元图应用于实体相似性问题。在本文中，我们将这个想法扩展到推荐问题。推荐问题要求我们近似大规模的用户项目评分矩阵。我们不考虑在线计算每个相似度，而是考虑计算矩阵，并设计使用由不同元图生成的用户项矩阵的最佳方法。最终预测。

## 5.2 Recommendation in HIN

现代电子商务网站使我们能够在提出建议时融入不同的信息。传统的推荐系统必须加强，以利用异构信息提供的丰富的语义。例如，Ma等人[20]将社会关系作为正则化的因素纳入到推荐系统的矩阵分解中。在[4,34]中，项目的元数据被建模来改进推荐任务。在[19,21]中，评估文本与评级预测任务中的评级一起进行分析。Ye等人[38]提出了将用户偏好，社交网络和地理信息相结合的概率模型，以加强兴趣点推荐。以前的方法已经证明了异构信息在提高推荐准确性方面的重要性和有效性。然而，这些方法中的大多数不同地处理了不同的异构信息，因此**失去了存在于信息之间的信息**。

已经提出基于HIN的建议来避免对不同类型的信息进行不同的处理。基于元路径，有几种方法可以解决基于HIN的推荐任务。在[39]中，基于元路径的相似性被用作矩阵分解框架中的正则化项。在[40]中，多元路径用于学习用户和项目潜在特征，然后用于恢复由加权机制组合的相似性矩阵。在[29]中，用户对项目的评分被用来建立一个加权HIN，在此基础上使用基于元路径的方法来衡量用户推荐的相似度。不同元路径的组合是明确的，使用相似性而不是潜在特征。正如引言中所讨论的，所有上述方法都没有充分利用基于元路径的特征，而**基于因子分解的方法**却可以做到。

## 5.3 Factorization Machine

因子分解机[25]是一种最先进的推荐框架，它可以对特征之间的交互进行建模，例如评分信息，项目类别，文本，时间等。因此，它是一个强大的框架，整合内容功能的协作 - 基于过滤的推荐。许多方法和系统已经基于FM开发[9,11,26]。不同于以往只考虑显式特征的方法，我们通过基于不同元图的矩阵分解生成潜在特征。对于使用原始显式特征的FM，MF可以被认为是与PCA类似的步骤，以执行降维以降低原始特征的噪声。

# 6 CONCLUSION

在本文中，我们提出了一种基于异构信息网络（HIN）的推荐方法。我们介绍一种融合网络异构信息的原理。通过使用从HIN模式派生的元图，我们可以制定用户和项目之间复杂的语义。我们使用矩阵分解来以无监督的方式从每个元路径获得用户和项目的潜在特征。我们使用一个组套索正则化因子分解机融合从不同元图提取的不同组语义信息来预测链接。实验结果证明了我们方法的有效性。今后，我们计划探索更丰富的信息，丰富网络中的特征和语义，并用并行计算和深度学习来进一步完善我们的系统。