基于知识图谱的推荐算法研究

杨显鹏

(大连海事大学 信息科学技术学院, 辽宁省大连市 中国 116000)

摘 要 随着互联网的发展，数据开始呈爆炸性的方式增长，随之而来的是海量的脏数据。为了不受这些脏数据的影响得到高质量的数据，推荐系统开始引起人们的广泛关注。传统的推荐算法比如协同过滤( CF) 太过于依赖用户和物品的交互信息，因此遗留的数据稀疏和冷启动问题成为学术界一直需要攻克的难题。最近，知识图谱由于其三元组的易理解性以及丰富的语义信息而被广泛应用在推荐系统中。由此提出一种基于知识图谱嵌入模型 TransE 和图神经网络的推荐算法，使用图神经网络可以提取用户和物品的高阶特征，知识图谱嵌入算法模型可以使提取出的特征信息更加丰富，在相关数据集上的实验结果表明该算法的推荐性能表现优异。

关键词 推荐系统; 知识图谱; 图神经网络; 图嵌入

中图法分类号 \*\*\*\*　　　DOI号 \*投稿时不提供DOI号\* 分类号

Research of Recommendation Algorithm Based on Knowledge Graph

Xianpeng Yang

1)(College of Information Science & Technology , Dalian Maritime University, Dalian , China)

**Abstract**

With the development of the Internet，data began to grow along with huse amounts of dirty data in an explosive way. In order to get higher quality data without being affected by these dirty data，people begin to focus on Recommendation Systems. Traditional recommendation algorithms，such as Collaborative Filtering ( CF), rely too much on the interactive information between users and items，so the problems of data sparsity and cold start have become difficult problems to be solved by the academia all the time. Recently, Knowledge Graph has been widely used in recommendation systems because of its trituple's intelligibility and rich semantic information. For these reasons，we propose a recommendation algorithm model，which is based on knowledge embedding model TransE and Graph neural network. Higher-order representation of users and items can be extracted by using the Graph neural network. Knowledge graph embedding algorithm model can extract the characteristic information in more abundant way. Experimental results on related datasets show that the recommended performance of the algorithm is excellent．

**Key words** recommendation system; knowledge graph; graph neural network; graph embedding

# 引言

推荐系统像很多其他基于海量数据的任务一样受益于深度神经网络的发展，而知识图谱作为典型的图结构数据包含着实体到实体之间的关系，这对用户的兴趣分析和建模具有一定的辅助作用。基于矩阵分解的协同过滤(Collaborative Filtering，CF) 是商业领域最成功的方法之一，然而，基于 CF 的方法依赖于用户和项目之间过去的交互，这将导致冷启动问题( 不推荐没有交互的项目) 。为缓解这一问题，研究人员通常会采取一些措施去整合辅助信息，比如社交网络、图片和评论等。

在众多种类的辅助信息中，知识图谱被广泛使用，其以机器可读的头-关系-尾( head-relation- tail) 三元组形式组成并包含丰富的结构信息。研究人员先后利用知识图谱在节点分类、句子补全和摘要生成等应用中取得了成功。此后出现了基于知识图谱感知的推荐模型，其中许多都受益于图神经网络( Graph neural network，GNN) 捕捉图中的高阶结构并细化嵌入用户和项目的特征。 如 RippleNet传播用户在知识图谱中潜在的偏好并探索其更深层次的兴趣; 图 卷 积 网 络( Knowledge Graph Convolutional Networks， KGCN) 利用卷积操作来产生高阶的连通性的物品特征; 图注意力网络( Knowledge graph attention network，KGAT) 使用注意力机制隐式地为图中不同的邻域节点指定权重系数。

由于推荐系统的高维和异质性，在推荐系统中使用知识图谱仍是一个挑战。一种可行的方法是通过知识图嵌入 (Knowledge Graph Embedding，KGE) 方法先行预处理知识图谱，该方法可将图谱中的实体和关系映射到低维向量表示。常用的 KGE 方法侧重于建模严格的语义相关性( 如 TransE 模型假定头 + 关系 = 尾) ，因此这类方法非常适合于图文应用领域，如知识图谱补全和链接预测。

本文在此基础上提出一种基于 GNN 和知识 图谱嵌入模型 TransE 的推荐算法，先把相应的知识图谱信息通过 KGE 算法映射到高维的向量空间，再将图谱输入到相应的 GNN 之中; 实验表明，更高维的语义信息可以提升神经网络的学习能力，使最后的推荐性能有所提升。

相关专业术语介绍

对于一个典型的推荐系统，用户u和物品v的集合通常表示为和。根据用户和物品的历史行为数据，可得到二者间的交互矩阵为。如果一对用户和物品间存在关联（如打分、点击等）那么反之。

对于用来提升推荐性能的知识图谱，用三元组集合来表示，其中每个三元组包含了知识图谱的头节点h、尾节点t和二者之间的关系r。 表示所有的实体(包括头节点和尾节点)集合，为知识图谱的关系集合。

一般来说，知识图谱中的实体有一些代表着物品v，而这些存在于知识图谱中的物品通常多个实体有关联，所以把物品v有关的实体集合表示为 推荐系统可以在这个集合的基础上寻找到用户的潜在兴趣实体。最终的预测函数表示为

式中为用户u对物品v感兴趣的概率；整个函数的所有参数；G为知识图谱的三元组集合。

**2.1 知识图谱嵌入模型TransE**

为使推荐模型和知识图谱更好地融合，使用知识图谱嵌入模型先行处理知识图谱数据，得到语义信息更为丰富的向量; 然后再将得到的相关嵌入向量输入到后续的 GNN 模型中。

传统的知识图谱一般使用本体语言表示，深度学习给予了一个更为明确的思路: 用向量的方式来表示知识图谱。这种形式在需要进行的任务中，如预测、推理等，具有更强的可扩展性与可表达性。嵌入模型目标是把一对对三元组编码为低维的向量形式。知识图谱嵌入模型的目的是向低维向量空间中嵌入多关系数据的实体和关系，同时还能保留数据中的结构信息。

表示学习的目的是将需要表达的对象(知识 图谱等) 表达为机器可以理解的实值向量的形式。对于知识图谱，表示学习目标是图谱中的实体和关系，然后构建模型将实体和关系映射到低维向量空间中进行后续的推理或预测任务，TransE 模型是表示学习的一个经典方法。

TransE 模型属于翻译模型：其将实体和关系 表示为同一空间中的向量，对于给定的三元组 模型将其中的关系看成头节点到尾节点 的平移向量，即; 这种思想来自于词向量空间的平移不变性，TransE 模型如图 1 所示。

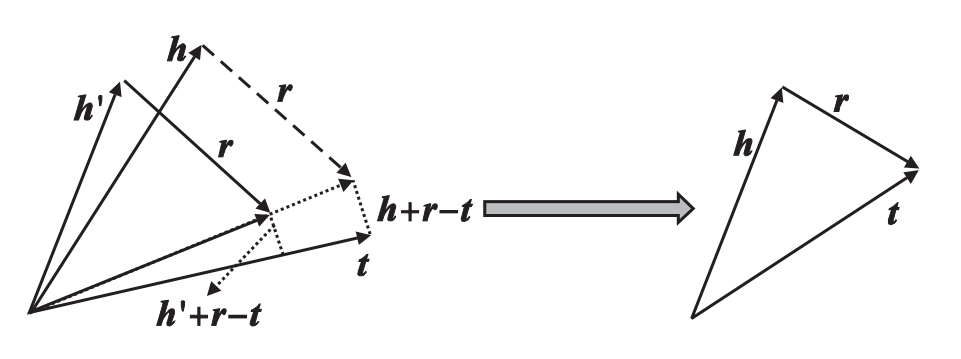
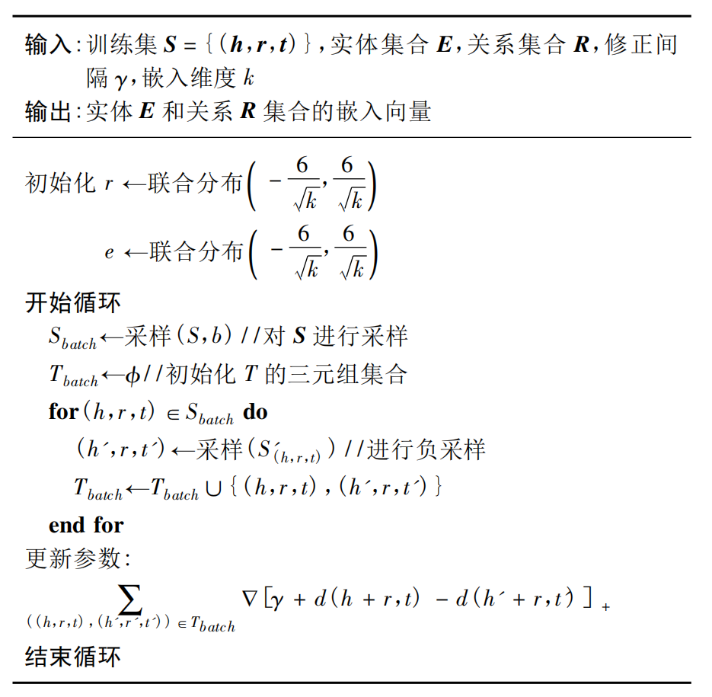


图1 TransE模型原理

在训练模型过程中，模型会不断调整其参数，使得知识图谱中的的距离尽可能的小。模型的优化目标为

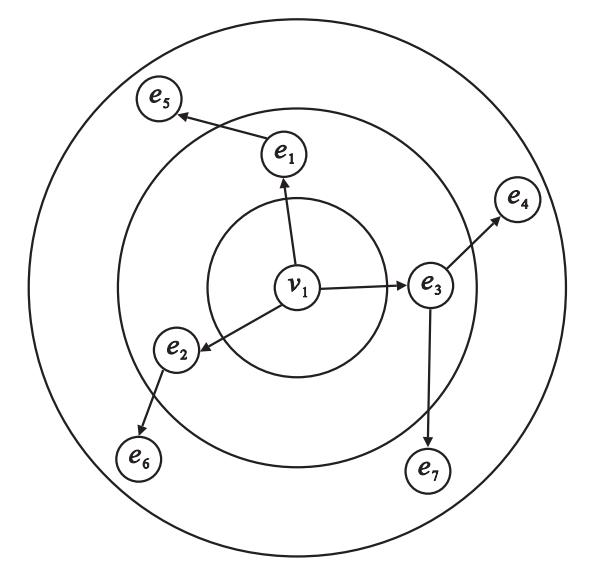
式中表示函数取值大于零时取值不变，小于令时则取零，这种函数一般称之为合页损失函数;为一个正确三元组与错误三元组之前的间隔修正，越大，两个三元组之前被修正的间隔就越大，则对向量的修正就越严格(一般都设置为1)；d为和两个向量之间的距离，一般使用的是L1或L2范数；S为用来训练的三元组的集合。模型的目标是让正确三元组之间的距离变小、错误三元组的距离变大; 所以如果函数取值大于零，则表示需要对模型的参数进行调整，训练流程如表1所示。

表1 TransE算法整体流程



模型在知识图谱上的采样步骤如图 2 所示。 假设用户点击的物品为 v 1，以 v 1 为中心向外扩散 一个步长，将相关的实体放入一个集合 N( v) 中，然后根据集合中每个实体的嵌入特征将集合的所 有特征聚合成一个向量，再以此特征为中心继续 重复之前的步骤向外扩散。

采样后网络会根据图卷积网络(Graph Convolution Networks，GCN)层进行特征提取，GCN模型如图3所示。对每个物品 v，先根据上面的采样结果对邻域节点计算相关系数代表着对用户 u 来说物品 v 和实体 e 的关联程度)，u 代表 不同用户，



概念在哲学中被理解为外延与内涵所组成的思想单元，德国数学家Wille在1982年首次提出了形式概念分析，用于概念发现、排序与显示。概念格作为形式概念分析的核心数据结构，是根据形式北京中对象与属性之间的二元关系建立的一种概念层次结构。概念格能够通过Hasse图清晰地体现概念间的泛化和特化关系，因此被认为是进行数据分析有力的工具。

定义1. (形式背景)一个形式背景是由两个集合和和以及与间的关系组成。的元素称为对象，的元素称为属性。或表示对象 具有属性，如表1所示。

表1 形式背景示例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 考古遗址 | 沙滩 | 欧元 | 溪流 | 滑雪区 |
| 雅典 | 1 | 1 | 1 | 0 | 0 |
| 因斯布鲁克 | 0 | 0 | 1 | 1 | 1 |
| 巴黎 | 0 | 0 | 1 | 1 | 0 |
| 罗马 | 1 | 1 | 1 | 1 | 0 |

表1为四座欧洲城市与城市特有属性构成的形式背景，可以看到表中各对象与属性的关系值非 0 即 1。例如，如果巴黎流通欧元，那么它所对应的属性值则为 1，否则为 0。这种类型的形式背影由于其仅用 1 和 0 表示关系的存在与否，所以形式背景仅包含二元关系，基于形式背景下的概念格称为经典概念格。当然现实中对象与属性之间的关系不仅仅只有 0 和 1 的形式，还可能有多个离散值或连续实值的的情况存在。

**定义2.** （伽罗瓦联系）设是对象集合的一个子集，定义（中对象共同属性的集合）。相应地设是属性集合的一个子集，（具有B中所有属性对象的集合）。若，，则称集合 满足伽罗瓦联系。

**定义3.**（形式概念）背景上的一个形式概念二元组，其中,且与满足伽罗瓦联系，则称是概念的外延，是概念的内涵

**定义4.** (层次序)设和是格中的两个概念，且()，称是的子概念，是的超概念，记为，关系称为概念格次序。

概念格就是由形式背景下的所有概念以及概念之间的层次序构成的，通常以Hasse图的形式对其格结构进行可视化，图2为表1形式背景下概念格的Hasse图。为便于表示我们将表2.2所示形式背景中的对象按从上至下的顺序一次标注为a、b、c、d，属性按从左至右的顺序标记为i、j、k、l、m。

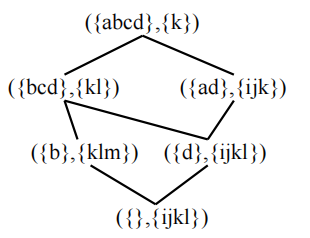


图2 形式背景下多的概念格Hasse

**定义5.** （邻域概念）当不存在概念满足时称是的子节点。当概念与概念具有相同的上邻（或下邻）时，称是的兄弟节点。

**定义6.** （起始概念）给定一个对象a，且，当不存在概念，并且，概念@是对象的起始概念。

命题1.一个对象具有的属性集即为此对象起始概念的内涵集。

证明：设对象a的属性集，对象集a的起始概念，由于，可知。当时，根据定义2可以构造，再结合条件，得出，显然这与起始概念的定义相矛盾，所以，命题1得证。

**2.2 推荐系统**

数据规模的快速增长与用户日益多样化的需求促使了推荐系统的产生。相比于更早起步的搜索引擎，推荐系统做为另一种能够有效处理信息过载问题的手段，虽然其处理方式有所不同，但推荐系统的本质就是发掘用户与产品间的内在联系，最终能够达到对用户需求或兴趣的一种预测。人们最为了解的也是推荐系统较早应用的领域就是电子商务领域。种类繁多的商品信息量早已超出了人们自身的检索能力范围。在搜索引擎与推荐系统的成功应用下，用户不仅可以使用搜索引擎主动过滤出符合需求的有效信息，也可以在推荐系统的作用下，被动地接收到自己可能感兴趣的商品信息。不仅是在电子商务方面，推荐系统的研究发展进一步完善了互联网的使用环境，为用户提供了便利。通常的推荐系统都有几个相对独立的模块组成，如图3所示。

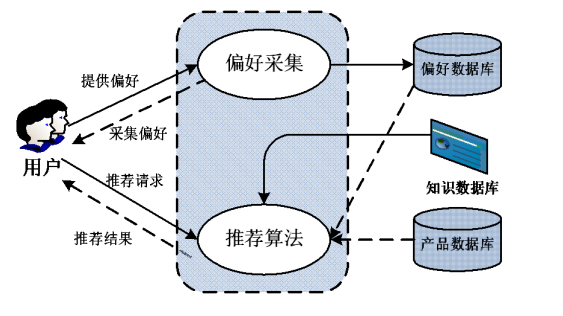


图3 推按系统通用模型

作为推荐系统的核心，推荐算法直接影响着该系统的推荐水准。它的主要作用就是基于数据模型，采取不同方式对用户的偏好以及需求进行预测，进而将用户可能感兴趣的产品形成推荐列表呈献给用户。推荐算法发展至今，已经得到了大量学者的深入研究，也涌现出了侧重点不同，处理方式不同的推荐方法，下面主要对一些常见的推荐方法进行介绍，并对比分析了方法自身的优缺点。

1. 基于内容的推荐、

基于内容的推荐算法源于一个基本的假设：“用户可能会喜欢与他曾经喜欢过的物品相似的物品”，其通过建模计算用户曾经有过的显示反馈和隐式反馈的物品集合与所有物品的相似度，按照相似度的大小排序到推荐列表，并考虑时间因素、购买地点等因素来优化推荐结果。此外还可以使用基于统计和机器学习方法从 用户历史反馈数据中建模学习用户喜好，替代传统相似度的方法。基于内容的推荐算法的主要优缺点在于可以解决新物品冷启动的问题，不受R的稀疏性的影响；推荐结果有不错的可解释性。其缺点在于需要复杂的特征工程构造物品属性特征，否则会严重影响推荐结果；缺乏多样性，推荐结果会与用户曾经显示和隐式反馈的物品高度相似；新用户的冷启动问题。

1. 基于协同过滤算法

基于协同过滤的推荐算法是目前应用最成功的推荐方法其利用用户和物品历史的反馈数据,挖掘用户和物品本身的相关联性并基于此进行推荐。具体地, 这类方法可以被分为 3 类: 基于用户的推荐、基于物品的推荐和基于模型的推荐。 基于用户的协同过滤方法是基于假设 “用户可能喜欢与他相似用户喜欢的物品”, 通过用户历史反馈记录计算用户间的相似度, 利用其相似的用户对物品的反馈, 来预测对应用户的反馈情况, 并进行推荐. 这类方法的主要优点在于避开了对物品自身属性的特征挖掘, 缺点在于在用户数量变化很大的情况下, 算法效率较低, 并且面临新用户的冷启动问题. 基于物品的协同过滤方法与基于内容的推荐算法假设类似, 不同在于其使用物品历史被反馈的数据来判断物品之间相似性。其优点在于计算简单, 因为物品反馈结果变化比用户要低很多, 相较于基于用户的协同过滤算法, 更可以通过离线计算, 定时更新来完成, 其缺点则是无法在不离线更新物品相似性时推荐新的物品给用户。基于模型的协同过滤方法是为了解决基于用户、物品的协同过滤方法所面临的数据稀疏、难以在大数据量级上返回即时结果的问题.。其通过历史数据利用机器学习方法训练得到一个预训练模型 , 从而可以实时预测任意用户对某一物品的喜好。

**2.3 形式概念分析在推荐系统领域的应用**

形式概念分析及概念格相关理论在推荐系统方面的应用研究仍处在探索阶段。2006 年 Boucher Ryan 等人首次提出了将形式概念分析与协同过滤算法结合的思想，该文献将概念格作为用户与产品间关系信息的存储载体，通过利用概念之间的偏序关系，探索性地搜索近邻概念，并从中获取推荐候选项，虽然并未具体提出明确的搜索策略，但为之后的进一步研究提供了方向。

Tomohiro Murata 等人结合形式概念分析提出了一种基于知识的推荐模型，该模型的核心结构主要分为三个部分：（1）知识源本体，用于知识表示，该部分描述了产品来源与相关特征的综合信息；（2）用户配置文件本体，用来有组织的存储用户的历史和行为信息，通过分析使用用户的请求与喜好，从而是搜索更加快速；（3）形式概念本体，它是对所有实体和其属性以及实体间关系的形式化描述，提供了以个捕捉关键区别的通用映射域，加快了推荐候选项集的生成。通过以上三部分的协同工作，最终为用户提供个性化的推荐项。另外，也有研究者为了应对多值背景将模糊形式概念分析应用在了推荐问题中。国内也有部分学者将形式概念分析相关理论与推荐系统进行了结合。文献中提出了一种基于概念格的图书协同推荐模型，利用概念之间的偏序关系，寻找与目标用户相近的用户群体，从这些相似用户的阅读记录中挑选书籍推荐给目标用户。文献从大量的社交数据中抽取用户知识，以概念格为载体，构造了用户属性概念格和用户社交概念格结合带重启的随机游走算法，进行朋友推荐。

# 概念格及起始概念索引的构造

**3.1 基于用户记录的概念格的生成**

推荐系统一般通过日志系统获取用户行为数据，并按照一定的格式进行再处理。一条用户记录通常表示某用户对某物品在某一时刻进行了某种操作，包含用户标识、物品标识以及其他一些辅助信息。例如一个 7 × 4 的用户-产品评分矩阵，其中包含了 7 位用户对 4 件产品的所有打分，如果分值为 0则说明该用户未访问过该产品，如表 2 所示。

表2 用户-产品评分矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | I1 | I2 | I3 | I4 |
| U1 | 0 | 3 | 0 | 2 |
| U2 | 0 | 0 | 1 | 5 |
| U3 | 0 | 0 | 0 | 2 |
| U4 | 5 | 0 | 0 | 0 |
| U5 | 1 | 3 | 4 | 0 |
| U6 | 0 | 0 | 1 | 0 |
| U7 | 4 | 2 | 0 | 0 |

虽然用户-产品矩阵能够清楚地反映用户与产品之间的关系，但并不能从中直接提取用户或产品的邻域信息。而概念格具有明显的聚类特性，概念间的层序关系也能清楚反映对 象(用户)集间的泛化与例化关系，通过概念之间的关系可以直接获取邻域信息。

表3 形式背景

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | I1 | I2 | I3 | I4 |
| U1 | 0 | 1 | 0 | 1 |
| U2 | 0 | 0 | 1 | 1 |
| U3 | 0 | 0 | 0 | 1 |
| U4 | 1 | 0 | 0 | 0 |
| U5 | 1 | 1 | 1 | 0 |
| U6 | 0 | 0 | 1 | 0 |
| U7 | 1 | 1 | 0 | 0 |

根据概念格的定义，对照表2，可以直接将用户-产品矩阵 转化为形式背景． 用户集合作为形式背景中的对象集合，产品集合作为形式背景中的属性集合，用户与产品间的关系( 评分) 直接映射为对象与属性间的关系，如果某对象具备某种属性，则表示该用户对该产品进行了打分。由于是在经典概念格范畴内研究问题，对象与属性之间只存在0 和 1 的关系( 二元关系) ，即对象是否具有这个属性。所以只关注用户是否对产品进行了评分，而不考虑评分值。由表1 中用户-产品矩阵转化而来的形式背景如表4所示。由于需要利用概念格的结构关系，所以采用渐进式算法中的 Godin 算法来构造概念格。

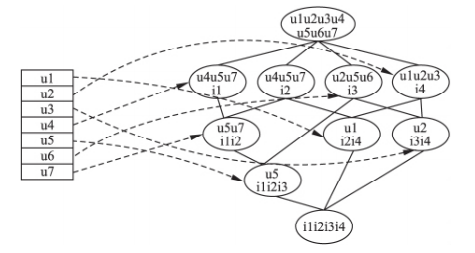


图4 起始概念索引分概念格

此算法是在有新的对象插入时，在原有概念格基础上进行调整。基于4中形式背景生成的概念格如图4右侧部分所示，节点上两排字符分别表示概念中的外延与内涵，即概念的属性以及包含这些属性的对象，节点与节点间的连线表 示概念之间的关系。

**3.2 起始概念索引的构造**

假设对用户 u 进行推荐，虽然每个用户具有唯一标识，但是在概念格中，外延中包含用户( 对象) u 的概念通常不止一个，所以需要从这些概念中确定唯一一个概念作为探索邻域的起点。由定义 5 可知，对象 u 的起始概念包含了该对象具有的所有属性，也是所有包含该对象的概念中唯一包含其所有属性的概念，所以起始概念恰好描述了该对象的所有特性，更适合作为探索的起点。然而在推荐的过程中，如果每次都要从概念格中定位目标对象的起始概念，则会产生大量重复性探索，显然这会影响整个算法的效率。在数据库技术中，索引的引入显著提高了数据检索的效率． 如果可以一次性的建立所有用户( 对象) 与其起始概念之间的对应关系，那么当对用户进行推荐时，就可以通过索引直接确定与之对应的起始概念位置，有效避免了大量的重复检索。

根据命题1可知，一个对象的起始概念的内涵集是唯一的，即该对象的属性集，所以该对象的起始概念在概念格中也具备唯一性。根据以上理论基础，构造包含所有对象的起始概念索引的方法如下:

函数1. ConstrutIndex（）

输入：概念格L

输出：起始概念索引集ICindex

说明：构造所有对象的起始概念索引

1.for C in L

2. if len(Map) == 所有对象个数

3. break//构造完成

4. for j in C.extent:

5. if j in ICindex.keys():

6. continue

//设置标记变量 f，表示当前概念 C 是否是

对象 j 的起始概念

7. f=True

8. for k in C.chd

//判断概念 C 是否为对象 j 的起始概念

9. if j in L[k].extent:

10. f = False

11. break

12. if f is True:

// 将映射 j→C，即对象 j 与起始概念 C 的

对应关系添加至索引集中

13. ICindex.append(j->C)

函数 1 中对概念格的所有概念逐个遍历，因为在概念格的存储结构中每个概念都包含了其父节点及子节点的所有信息，所以可以直接获取概念的子节点，并根据起始概念的定义找出符合条件的对象图4左侧部分代表索引键( 对象) ，虚线则代表对象与该对象起始概念的对应关系.通过构造起始概念索引，使得在推荐时可以以目标对象作为索引键直接访问起始概念.

# 结论

协同过滤算法作为推荐系统产生之初就出现推荐算法之一，时至今日依然在不同的推荐背景下发挥着重要作用。针对推荐系统所面临的关键问题提出了一种面向隐式反馈数据的基于概念邻域的推荐算法。将用户与产品的评分( 关系) 矩阵转化为二元形式背景，以此为基础构造出相应的概念格，将用户与产品分别以对象与属性的形式聚集在概念中，并通过概念间的偏序关系，以对象( 用户) 的起始概念为起点探索其近邻概念并获取候选项集。但还存在着可以改进的部分，如可将用户评分划分出不同区间来增加更多的语义信息等。

参 考 文 献

[1]陈昊文. 基于形式概念分析的推荐算法研究及应用[硕士学位论文].郑州大学,2017.

[2] Zhang Xizheng, Cai Yueyue, Luo Wen.Research on Personalized Knowledge Recommendation for Leading Users Based on Fuzzy Concept Lattice in Innovation Community. 2017,38(11):2553-2559.

(陈昊文,王黎明,张卓.基于概念邻域的Top-N推荐算法.小型微型计算机系统,2017,38(11):2553-2559.)

[3] Chicaiza, J. and P. V. Díaz (2021). "A Comprehensive Survey of Knowledge Graph-Based Recommender Systems: Technologies, Development, and Contributions." Inf. 12(6): 232.

[4]Diaz-Agudo B., Caro-Martinez M., Recio-Garcia J.A., Jorro-Aragoneses J., Jimenez-Diaz G. (2019) Explanation of Recommenders Using Formal Concept Analysis. In: Bach K., Marling C. (eds) Case-Based Reasoning Research and Development. ICCBR 2019. Lecture Notes in Computer Science, vol 11680. Springer, Cham.

[5] Chemmalar Selvi G.Lakshmi Priya G.G. Rating Prediction Method for Item-based Collaborative Filtering Recommender Systems Using Formal Concept Analysis.2020,EW,EAI

[6] 张伟. 社交网络中基于形式概念分析的用户推荐[硕士论文].西华大学,2015.

[7]Li X, Murata T,et al. A Know ledge-based Recommendation Model Utilizing Formal Concept Analysis . International Conference on Computer & Automation Engineering,2010,4:221-226.

[8] Maio CD, Fenza G, Gaeta M,et al.Rss-based e-learning recommendations exploiting fuzzy FCA for knowledge modeling [J].Applied Soft Computing,2012,12(1):113-124.

[9] Fang P, Zheng S. A Research on Fuzzy Formal Concep-t Analysis Based Collaborative Filtering Recommendation System[J]. 2nd International Symposium on Knowledge Acquisition and Modeling,2009, 3:352-355.