形式概念分析在推荐系统中的应用

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 做为处理信息过载的有效手段，推荐系统在短时间内得到了迅速的发展。传统的基于邻域的方法忽略了用户与产品间的结构关系，只考虑了同类对象间的相似关系。随着推荐系统的广泛应用，数据稀疏条件下的推荐问题也亟待解决。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，最后结合所提出的全局偏好度与邻域偏好度过滤出最终推荐结。该算法通过在两个公共数据集上的实验，相较于传统的基于邻域的推荐算法，具备较好的推荐效果，并更适用于数据稀疏条件下的推荐。

关键词 推荐算法；协同过滤；形式概念分析；概念格

中图法分类号 **** DOI 号 *投稿时不提供 DOI 号* 分类号

The application of formal concept analysis in the recommender system

Xianpeng Yang

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

Abstract As an effective mean to deal with information overload，the recommendation system has developed rapidly in a short time．The traditional neighborhood based approaches ignore the structural relationship among users or products，and only consider the similarity among the similar objects． With the recommendation system being applied extensively，the recommendation problem in the condition of data sparse also needs to be solved urgently． To solve the key problems the recommendation system faced with，a recommendation algorithm based on conceptual neighborhood is proposed for the implicit feedback data． The algorithm transforms the matrix of scores (relationships) between users and products into binary formal context，and constructs the corresponding concept lattice on the basis of formal context． so that the users and products are gathered in concepts in the form of objects and attributes respectively． As the starting point，the initial concept index can help users obtain the candidate item sets from their neighborhood through the order relationship among the concepts． Finally，the final recommendation can be generated through combining the proposed global preference and neighborhood preference． After the experiment on two public datasets，the proposed method with a better recommendation effect is more suitable for the recommendation under the data sparse condition，compared to the traditional neighborhood based recommendation algorithm.

Key words recommendation algorithm; collaborative filtering; formal concept analysis; concept lattices

1 引言

随着互联网的迅速发展，其用户数量也在迅速增长，但伴随而来的是信息过载问题的出现。海量的信息使用户在获取资源时无从下手。为了解决这样的问题，推荐系统应运而生。通过分析用户以及与用户相关的数据来推测出用户的兴趣所在和行为趋势，以此为基础向用户推荐他们所需要的信息和服务。推荐系统的产生有效的解决了信息过载问题，使数据增长不再成为用户按需获取信息的阻碍。目前推荐系统被广泛的应用在电子商务、视频网站、音乐媒体和社交网络领域。

推荐系统领域的核心部分主要由推荐算法构成，协同过滤(Collaborative Filtering)是目前应用最广泛的方法之一，如图 1 所示。其核心思想是，通过分析用户记录，在用户群体中找到指定用户的相似用户，综合这些相似用户对某一信息的评价，形成系统对指定用户对此信息的喜好程度预测。协同过滤的方法具体可以分为基于用户和基于物品两种，在一些企业的实际应用中更多的选择基于物品的协同过滤方法，因为基于物品的协同过滤算法可以相对于用户而言物品之间的关系更加的稳定，从而训练过程更加可控。

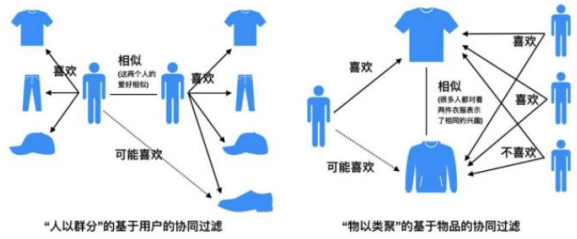

图 1 协同过滤方法

形式概念分析(Formal Concept Analysis，FCA)，是德国数学家 Wille 基于序理论提出的理论体系，其核心数据结构概念格已经广泛地应用于数据挖掘、信息检索、数据抽取、软件工程等领域，是一种强有力的数据分析与规则挖掘工具。由于概念格存在特殊的结构及性质，使得它在多个领域取得了较好的应用效果。但在个性化推荐领域，形式概念分析及概念格理论的应用仍处于探索阶段，所以具有一定的研究意义和价值。

2 形式概念分析与推荐系统相关知识

2.1 形式概念分析

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 $K=(G,M,I)$ 是由两个集合 G 和和 M 以及 G 与 M 间的关系 I 组成。 G 的元素称为对象， M 的元素称为属性。 $(g,m)\in I$ 或 gIm 表示对象 g 具有属性 m ，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 A 是对象集合 G 的一个子集，定义 $f(A)=\{m\in M\mid \forall g\in A, gIm\}$（ A 中对象共同属性的集合）。相应地设 B 是属性集合 M 的一个子集，$g(B)=\{g\in G\mid \forall m\in B, gIm\}$（具有 B 中所有属性对象的集合）。若 $A=g(B)$，$B=f(A)$，则称集合 满足伽罗瓦联系。

定义 3.（形式概念）背景 (G,M,I) 上的一个形式概念二元组 (A,B)，其中 $A\subseteq G, B\subseteq M$ 且 A 与 B 满足伽罗瓦联系，则称 A 是概念 (A,B) 的外延，B 是概念 (A,B) 的内涵

定义 4. (层次序)设 $C_1=(A_1,B_1)$ 和 $C_2=(A_2,B_2)$

是格中的两个概念，且$A_1 \subseteq A_2 (B_1 \supseteq B_2)$，称$C_1$是$C_2$的子概念，$C_2$是$C_1$的超概念，记为$C_1 \le C_2$，关系$\le$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。

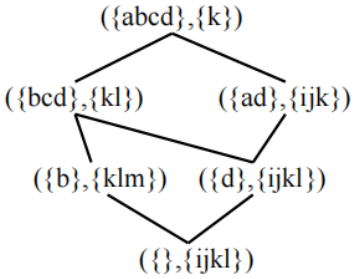

图 2 形式背景下多的概念格 Hasse

定义 5. （邻域概念）当不存在概念(A_3,B_3)满足$(A_1,B_1)<(A_3,B_3)<(A_2,B_2)$时称$(A_1,B_1)$是$(A_2,B_2)$的子节点。当概念$(A_1,B_1)$与概念$(A_2,B_2)$具有相同的上邻（或下邻）时，称$(A_1,B_1)$是$(A_2,B_2)$的兄弟节点。

定义 6.（起始概念）给定一个对象 a，且$a \in A_1$，当不存在概念$(A_2,B_2)<(A_1,B_1)$，并且$a \in A_2$，概念@是对象(A_1,B_1)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$H=\{m_1,m_2,...,m_k\}$，对象集 a 的起始概念$C=(A,B)$，由于$a \in A$，可知$B \subseteq H$。当$B \subset H$时，根据定义 2 可以构造$C^{'}=(g(H),H)$，再结合条件$B \subset H$，得出$C^{'}<C$，显然这与起始概念的定义相矛盾，所以$B=H$，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。

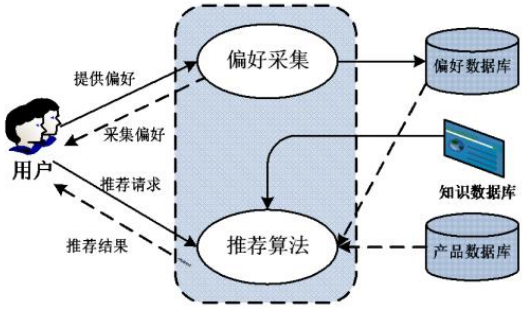

图 3 推按系统通用模型

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。

（2） 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 f, 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。

3 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

	I_1	I_2	I_3	I_4
U_1	0	3	0	2
U_2	0	0	1	5
U_3	0	0	0	2
U_4	5	0	0	0
U_5	1	3	4	0
U_6	0	0	1	0
U_7	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

	I_1	I_2	I_3	I_4
U_1	0	1	0	1
U_2	0	0	1	1
U_3	0	0	0	1
U_4	1	0	0	0
U_5	1	1	1	0
U_6	0	0	1	0
U_7	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。

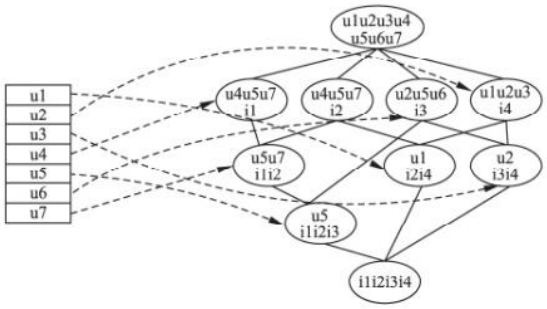

图 4 起始概念索引分概念格

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.

4 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.