

《智能信息处理》课程考试

形式概念分析在推荐系统中的应用

石现

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 11 月 29 日

形式概念分析在推荐系统中的应用

石现

(大连海事大学 信息科学技术学院, 大连 116026)

摘 要 做为处理信息过载的有效手段, 推荐系统在近些年得到了广泛的研究与发展, 推荐 系统在各领域应用的成功案例也不断涌现, 但是依然面临着很多问题亟待解决。形式概念分析 (Formal Concept Analysis, FCA) 的核心数据结构—概念格 (Concept lattices), 是一种数据分析与规则提取的有效工具。外延与内涵做为概念的组成部分使得形式概念展现出了聚类的特性。概念之间存在的偏序关系也揭示了其泛化与特化的本质。随着其研究的不断深入, 形式概念分析开始逐步应用于数据挖掘、信息检索等领域。

关键词 形式概念分析, 推荐系统, 概念格

Application of formal concept analysis in recommendation system

Shi Xian

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026)

Abstract As an effective means to deal with information overload, recommendation system has been extensively researched and developed in recent years, and successful cases of application of recommendation system in various fields have emerged continuously, but it still faces many problems to be solved urgently. The Concept lattices, the core data structure of Formal Concept Analysis (FCA), is an effective tool for data Analysis and rule extraction. Denotation and connotation as the components of the concept make the formal concept show the characteristics of clustering. The partial order relation between concepts also reveals the nature of generalization and specialization. With the deepening of its research, formal concept analysis has been gradually applied to data mining, information retrieval and other fields.

Key words Formal concept analysis, recommendation system, concept lattice

1 引言

随着近年来互联网的飞速发展, 其用户 的数量也在以惊人速度增长。为了满足大量 用户的各项需求, 种类繁多且量级庞大的数 据支撑着各项互联网服务的正常运转。作为 人类的本能反应, 当想要获取任何事物时, 首先思维中会形成对需求的特征描述, 之后 根据这些特征去寻找目标, 这就是人们自发 产生的搜索行为。而在当今的互联网环境 中, 正是搜索引擎加快了用户获取信息的效率。一旦用户向互联网表述自己的明确需求, 搜索引擎就会根据用户所提交的各项需求信息迅速地将搜索结果反馈给用户。但是, 人们不可能任何时候都对需求具有清晰 明确的认识。面对这种情况, 互联网需要自 发地向用户提供信息资源,

并且尽量确保它 是符合用户潜在需求的, 这就是个性化推 荐。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应用于数据挖掘、信息检索、数据抽取、 软件工程等领域[2,3], 是一种强有力的数据 分析与规则挖掘工具。其主要研究集中在概念格的构造维护算法与实际领域中概念格 的应用两个方面, 具体的研究现状及成果会 在之后内容中介绍。由于概念格存在特殊的 结构及性质, 使得它在多个领域都获得了较 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, 其研究进展和成果也无法与其他领域 相提并论。虽然如此, 但仅就目前研究进展 看, 概念格的自身特质以及相关理论是可以 一定程度上促进个性化推荐问题的解决, 所 以具有一定的研

究意义及价值。

2 形式背景的转化

推荐问题通常面对的数据主要分为显式（如用户对产品的评分）与隐式（如用户的购买、浏览记录）两类。在隐式数据中，通常用户对项目只发生过购买或浏览类的操作，并未对其进行定量地评价，所以用 0 与 1 来表示用户与项目之间是否存在关系。对照形式背景定义中的描述，可以自然地将隐式数据中的用户看作对象，项目作为属性，完成隐式数据到形式背景的转化，转化过程如表 1 所示。

表 1 隐式数据向形式背景的转化

	item1	item2	item3	item4
user1	0	1	1	0
user2	1	0	1	1
user3	1	1	1	1

不难想象，现实很多应用场景中用户与项目之间不仅只存在表示二元关系的隐式数据。但是在经典概念格的研究范畴中，其所依赖的形式背景中只包含二元关系，即对象与属性的关系值非 0 即 1。如果遇到显示数据的情形，由于用户与项目之间的非二元关系，往往不能直接基于多值背景进行概念格的构造，需要将其转化为只包含二元关系的形式背景。

3 形式背景下的概念格的生成

为了够造以形式概念为单位的邻域环境，在确定了形式背景之后，需要在其基础上构造出完整的概念格。为了便于讨论概念格的构造过程，给出表 2 所示的形式背景，其中对象表示推荐背景下的用户，属性表示项目。

表 2 形式背景下的概念格的生成

	属性1	属性2	属性3	属性4
对象1	0	1	0	1
对象2	0	0	1	1
对象3	0	0	0	1
对象4	1	0	0	0
对象5	1	1	1	0
对象6	0	0	1	0
对象7	1	1	0	0

虽然表 2 能够清楚地反映用户（对象）与项目（属性）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，概念格中的层次序也能清楚地反映对象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。

4 概念格中起始概念索引的构造

4.1 Godin 算法

Godin 算法是一种经典的渐进式构造方法。该方法首先将概念格初始化为空，之后每当有新的对象加入，就在原有概念格的结构上进行动态调整。随着后续的数据对象加入，如果使用批处理方式，则需要在扩充的形式背景基础上重新进行所有概念与关系的生成，影响了够格效率。

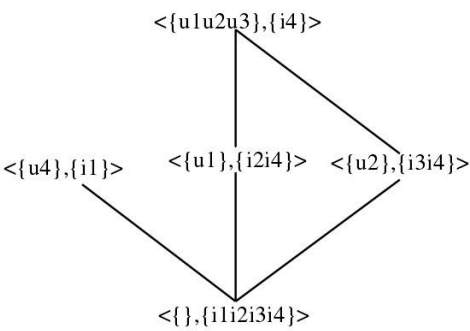


图 4.1 概念格的渐进式构造中间结果

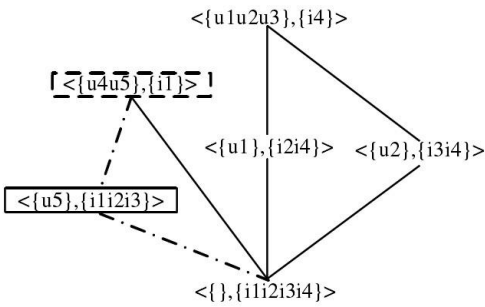


图 4.2 概念格的渐进式构造中间状态

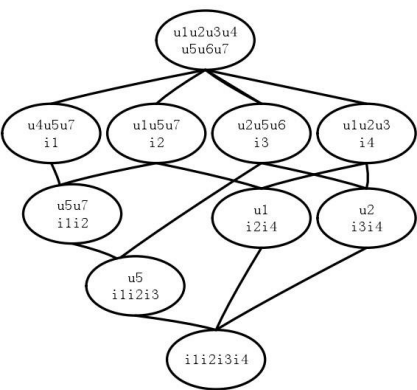


图 4.3 概念格构造完成的结构图

图 4.1 所示为表 2 中包含对象 1—对象 4 的渐进式够格中间结构。记对象为 u，属性为 i。当 u5 加入时，对其属性集 {i1,i2,i3} 与概念格中的现存概

念的内涵分别求交集,并根据结果进行分类处理。

(1) 当与概念的内涵求交集,得出集合 $\{i1,i2,i3\}$ 。而图 4.1 中的结构中并没有内涵为该集合的概念存在。这时候需要新生成概念,并加入到原有的格结构中。如图 4.2 中所示,为新生成概念,并且与概念之间建立了父子关系。(2) 当与图 4.1 中的概念的内涵进行交集运算后,得到集合 $\{i1\}$,即与概念的内涵相同,这也说明了对对象 $u5$ 的属性集包含于该概念的内涵中。此时需要对的外延进行更新,结果如图 4.2 中虚线框中的概念所示。值得注意的是,更新后的概念与新生成的概念和原有概念之间均存在直接父子关系,出现关系冗余的情形。观察图 4.2 可知,需要将概念与原有概念之间存在的直接父子关系删除,才能得到正确的构造结果。

4.2 起始概念索引的构造

当需要对目标用户 u 进行推荐时,虽然每个用户都拥有唯一标识,但是在概念格的所有概念中,外延包含用户(对象) u 的概念通常不只存在一个。所以,在概念格范畴中,如果选择从概念的角度出发,则需要确定唯一一个与用户 u 相关的概念作为出发点。由之前定义可知,对象 u 的起始概念的内涵集包含了该对象所具备的所有属性,也是所有外延集包含该对象的概念中,唯一的内涵集包含其所有属性的概念。通过分析以上特性可以看出,任一对象的起始概念是所有概念中唯一能够完整描述该对象特性的概念,所以更适合作为寻找其邻域概念的起点。然而在整个推荐过程中,如果每一次推荐都需要从概念格中定位目标用户(对象)的起始概念,那么就会产生等同于用户数量的定位次数,这其中必定会出现大量的重复性搜索,也会一定程度上影响整个算法的执行效率。在数据库相关技术中,索引的引入很大程度上提升了数据检索的效率。依据索引的构造思想,假设可以一次性地建立所有用户(对象)与起始概念之间的对应关系,那么当对某一用户进行推荐时,就能够借助已有的对应关系快速获取与之相应起始概念的位置,有效地减少了大量的重复工作。

5 总结

本章分别介绍了推荐系统与形式概念分析相关理论知识,对主流的个性化推荐方法进行了列举

介绍,并对推荐系统中仍普遍存在的问题及各领域中推荐系统的应用情况作了叙述,给出了形式概念分析及概念格理论的相关定义及性质,并详细分析了形式概念分析及概念格理论在推荐系统中的应用现状。从形式背景的转化、概念格的生成与起始概念索引的构造三个方面,对基于概念邻域的推荐的数据处理工作进行了叙述。介绍了不同类型多值背景向单值背景的转化方法,通过分析起始概念定位中存在的问题,提出了起始概念索引的定义与构造方法。

参考文献

- [1] R Bell,Y Koren,C Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems[C].Acm Sigkdd International Conference on KnowledgeDiscovery & Data Mining, 2007:95-104
- [2] Wang Li-ming , Zhang Zhuo.An algorithm for mining closed frequent itemsets based on apposition assembly of iceberg concept lattices[J].Journal of Computer Research and Development, 2007,44(7): 1184- 1190.
- [3] Zhang Zhuo,Du Juan,Wang Li-ming. Formal concept analysis approach for data extraction from a limited deep-web database[J].Journal of Intelligent Information Systems, 2013, 41(2):211-234.
- [4] John S.Breese,David Heckerman,Carl Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[J].Fourteenth Conference on Uncertainty in Artificial Intelligence,2013,7(7):43-52.
- [5] Polat H,Du W.SVD-based collaborative filtering with privacy[C].Acm Symposium on Applied Computing,2005,1:791-795.
- [6] Koren Y, Bell R,Volinsky C.Matrix factorization techniques for recommender systems[J].Computer,2009,42(8):30-37.
- [7] Vucetic S,Obradovic Z.Collaborative filtering using a regression-based approach[J].Knowledge Inf Syst .2005,7(1):1-22.
- [8] Rendle S, Freudenthaler C, Gantner Z,Schmidt-Thieme L.Bpr: Bayesian personalized ranking from implicit feedback[C]. In: Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence,2009:452-461.
- [9] Deshpande M, Karypis G. Item-based top-n recommendation algorithms[J].ACM Transactions Inf Syst , 2004, 22(1):143-177.
- [10] Linden G,Smith B,York J.Amazon.com Recommendations Item-to-Item Collaborative Filtering [J].IEEE Internet Computing,2003,7(1):76-80.
- [11] Koren Y.Factorization meets the neighborhood:a multifaceted collaborative filtering model[C]. In:Proceedings of the 14th ACM

SIGKDD international conference on knowledge discovery and data mining, 2008, pp 426–434.

- [12] Xue G-R, Lin C, Yang Q, et al. Scalable collaborative filtering using cluster-based matrix smoothing[C]. *Sigir: International ACM Sigir Conference on Research & Development in Information Retrieval*, 2005, 114–121.
- [13] 谢志鹏, 刘宗田. 概念格的快速渐进式构造算法[J]. *计算机学报*, 2002, 25(5): 490–496. [14] 沈夏炯, 韩道军, 刘宗田, et al. 概念格构造算法的改进[J]. *计算机工程与应用*, 2004, 40(24): 100–103. [2] 张云中. 基于形式概念分析的领域本体构建方法研究[D]. 吉林大学, 吉林, 2009.