《智能信息处理》课程作业

基于形式概念分析的推荐算法研究

于慧敏

作业	分数[20]
得分	

2021 年 11 月 23 日

基于形式概念分析的推荐算法研究

于慧敏

(大连海事大学 信息科学技术学院, 辽宁省大连市 116026)

摘 要 形式概念分析（FCA）的核心数据结构—概念格，是一种数据分析与规则提取的有效工具。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。作为处理信息过载的有效手段，推荐算法在近些年得到了广泛的研究与发展，推荐算法在各领域应用的成功案例也不断涌现，但是现有的推荐算法往往面对的是无法直观反映用户喜好程度的隐式数据，并且随着产品种类的剧增，用户与项目间产生的隐式数据也会变得极为稀疏。而由于稀疏数据环境下用户与项目信息的缺失，最终的推荐效果就会受到影响。针对以上问题，本文提出了一种面向隐式数据的基于概念邻域的推荐算法，以概念格为载体进行推荐问题求解，适用于数据稀疏环境下的推荐。

关键词 推荐算法；形式概念；概念格；形式背景；协同过滤

Research on recommendation algorithm based on formal concept analysis

Yu Huimin

(Computer science and technology, Dalian maritime university, Liaoning Dalian, 116026)

Abstract The core data structure-concept lattice of Formal Concept Analysis (FCA) is an effective tool for data analysis and rule extraction. With its continuous research, formal concept analysis began to be gradually applied to data mining, information retrieval and other fields. As an effective means to deal with information overload, recommendation algorithms have been widely researched and developed in recent years, and successful cases of recommendation algorithms in various fields have emerged, but existing recommendation algorithms often face implicit data that cannot intuitively reflect user preferences, and the implicit data generated between users and items will become extremely sparse as the number of product types increases dramatically. Due to the lack of user and item information in the sparse data environment, the final recommendation effect will be affected. To address the above problems, this paper proposes a recommendation algorithm based on concept neighborhood for implicit data, using concept lattice as a carrier for recommendation problem solving, which is applicable to recommendation in data sparse environment.

Key words recommendation algorithm; formal concept; concept lattice; formal context; collaborative filtering

1 形式概念分析及概念格相关理论

在哲学范畴中，概念被理解为由外延与内涵所构成的思想单元。德国数学家 Wille 在 1982 年首先提出了形式概念分析（Formal Concept Analysis，FCA），用于概念的发现、排序和显示。下面首先对形式概念分析与概念格的基本理论进行介绍。概念格作为形式概念分析的核心数据结构,是基于形式背景中对象与属性之间的二元或多元关系建立起的一种概念层次结构。Hasse 图能够清晰地体现概念格中概念之间的泛化与特化关系，因此被看做是进行数据分析的有力工具。下面给出形式概念分析的相关定义：

定义 1.1（形式背景）：形式背景 $K=(G,M,I)$是由两个集合 G（对象集）和 M（属性集）以及 G 中元素与 M 中元素之间的关系 I 所构成的。集合 G 中的元素称为对象，集合 M 中的元素称为属性。$(g,m)\in I$ 或 $g\,I\,m$ ，则表示对象 g 具有属性 m，或称属性 m 属于对象 g。

表 1.1 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 1.1 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 1.2（伽罗瓦联接）：对于任意的对象集合 $A \subseteq G$，定义函数 $f(A) \in \{m \in M \mid \forall g \in A, g\}$(集合 A 中对象共同具有属性的集合)。相应地，对于任意的属性集 $B \subseteq M$，$g(B) \in \{g \in G \mid \forall m \in B, g\ Im\}$(具有 B 中所有属性的对象的集合)。若 $A = g(B)$，$B \in f(A)$，则称集合 A 与 B 满足迦罗瓦联系，函数 f，g 为迦罗瓦联接。

定义 1.3（形式概念）：形式背景 $K=(G,M,I)$上的形式概念是以二元组 $C=(A,B)$的形式存在的，其中 $A \subseteq G, B \subseteq M$，且集合 A 与 B 满足迦罗瓦联系，则将对象集 A 称作概念 $C=(A,B)$的外延，属性集 B 为概念的内涵。

对于形式背景 $K=(G,M,I)$，$A_1, A_2, A \subseteq G$，$B_1, B_2, B \subseteq M$ 存在以下性质：

（1）$A_1 \subseteq A_2 \Rightarrow f(A_2) \subseteq f(A_1)$，$B_1 \subseteq B_2 \Rightarrow g(B_2) \subseteq g(B_1)$

（2）$A \subseteq g(f(A)), B \subseteq f(g(B))$，$A \subseteq g(B) \Leftrightarrow B \subseteq f(A)$

（3）$f(A_1 \cup A_2) = f(A_1) \cap f(A_2)$，$g(B_1 \cup B_2) = g(B_1) \cap g(B_2)$

（4）$f(A_1 \cap A_2) = f(A_1) \cup f(A_2)$，$g(B_1 \cap B_2) = g(B_1) \cup g(B_2)$

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 1.1 所示为表 1.1 中形式背景下概念格的 Hasse 图。为便于表示我们将表 1.1 所示形式背景中的对象按从上至下的顺序依次标注为 a、b、c、d,属性按从左至右的顺序标记为 i、j、k、l、m。

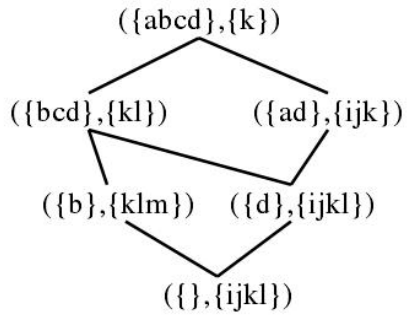

图 1.1 形式背景下的概念格 Hasse 图

2 形式概念分析在推荐领域的应用

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[1]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。文献[2]将概念格应用在了广告业词汇的推荐上，通过构造基于以广告公司为对象，所购买的广告词为属性的形式背景，并对其建立概念格，使用挖掘关联规则的方式来为广告公司提供个性化的广告词推荐服务。Tomohiro Murata 等人[3]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[4]。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[5]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。还有学者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 基于概念邻域的推荐算法

3.1 问题的形式化描述

为了将推荐的应用背景映射到形式背景中，将用户看做形式背景中对象，产品作为属性，那么对特定用户进行产品推荐就可以看做对形式背景中对象进行属性推荐。为了选取候选属性，结合概念间的层次序做如下分析：

（1）在概念 Ci 的父节点中，其内涵集应包含于概念 Ci 内涵集，所以对概念 Ci 的父节点进行探索无法从中获取对象 u 的属性以外的其它属性。

（2）概念 Ci 的兄弟节点的内涵集包含其与概念 Ci 共同所属父节点内涵集中的所有属性，同时也包含在概念 Ci 的内涵集中未出现过的属性，所以在概念 Ci 的兄弟节点中可以获取额外的属性作为候选属性。

（3）概念 Ci 的内涵集包含于其子节点内涵集，显然其子节点的内涵集中可以提供额外的属性作为候选属性。

根据以上分析可知，对于起始概念 Ci，只有在它的兄弟节点与子节点的内涵集中存在未在概念 Ci 的内涵集中出现过的属性，即候选属性。达到最终的推荐目的需要经历两个阶段。第一阶段在于推荐候选项集的构造，其作用在于将推荐产品的选取范围由整个产品集缩小为包含少量产品的候选项集，可以视作一次粗略地筛选。第二阶段的关键在于效用函数 Pre()的确定，它的作用就是进一步过滤掉候选项集中那些不适合作为最终推荐项的产品，同时将剩余的产品项作为推荐结果反馈给用户。

3.2 构造推荐候选项集

根据概念间的偏序关系，越是处于概念格下层的概念所包含的对象越特殊，因为这些对象具有更多的多属性。在确定推荐对象即目标用户的起始概念之后，本文利用概念格中概念之间的偏序关系，通过探索邻域概念中的内涵集合直接获取推荐候选项。为了构造推荐候选项集，并尽量使候选项构造充分，本文在这里采用控制递归深度的方法探索起始概念的子节点及兄弟节点。

3.3 用户对产品的偏好度计算

在候选项集生成之后，需要确定效用函数 Pre( )以计算产品对于用户的推荐度，并对每个用户的候选项集进行再次过滤，从而得到最终推荐项。在 Top-N 推荐中需要从中筛选出用户最有可能感兴趣的 N 项产品，传统的基于邻域的协同过滤算法通常利用用户相似度或物品相似度的方法来计算产品对于用户的推荐度。本文以概念格作为数据载体，结合概念相似度提出的全局偏好度与邻域偏好度两种偏好度定义，提出概念相似度的相关定义及度量方法。

3.3.1 概念相似度

在推荐问题的研究过程中经常涉及到用户相似度或产品相似度等概念。通过计算相似度，能够量化同类事物间的相关程度,便于比较与推荐结果的生成。但是在概念格中，每个概念都是由对象和属性共同构成的，所以本质上表现的是用户与产品间的关系，而不能单独用户或产品的角度去思考问题。如果只考虑二者其一，无异于传统的相似度度量方法，更无法通过挖掘概念间的内在联系去改进推荐效果。为了将问题的求解过程置于概念格的结构背景下，本文引入概念相似度的相关理论及度量方法。

与形式概念分析相比，领域本体中的概念相似度问题已经得到了更为广泛的关注与研究，例如基于字符、基于图以及基于知识的相似度度量方法。但在领域本体问题中，概念是作为一种数据标签的表现形式，而形式概念分析中的概念是不包含数据标签的，它可以看做是由对象与属性所构成的双聚类结构（对象聚类与属性聚类）。下面给出相似性度量的形式化定义：

定义 3.1 一种相似度测量方法 S 就是定义在集合 X 的笛卡尔积上的非负实值函数，形式如下：

$$S: X \times X \rightarrow R$$

S 满足以下条件：

1. $\exists s0 \in R: -\infty < S(x,y) \leqslant s0 < +\infty, \forall x, y \in X$
2. $S(x,x) = s0, \forall x \in X$
3. $S(x,y) = S(y,x), \forall x,y \in X$

如果 S 又同时满足：

1. $S(x,y) = s0 \leftrightarrow x = y$
2. $S(x,y)S(y,z) \leqslant [S(x,y) + S(y,z)]S(x,z), \forall x, y, z \in X$

那么相似度测量函数 S 又称为度量相似度函数。在模式识别与数据挖掘中，相似性函数通常是基于实数值与离散值的向量集合定义的。对于离散值向量的相似度度量方法来说，通常都是基于集合

间的关系运算与基数构造而来的。

为了将基于集合运算的相似性计算方法推广到形式概念中去，需要首先对形式概念的构成进行分析。由之前定义可知，形式概念本质上是由两个集合所构成，分别为对象集与属性集。所以，从形式概念的结构特点可以容易地想到通过分别计算对象集之间与属性集之间的相似度，并分别加权求和的方式来构造一种基于概念的相似度计算方法。按照这种方式就得到了一种称为加权概念相似度（weighted concept similarity）的概念相似度度量方法。通过引入这种概念相似度度量方式，在加权概念相似度的基础上进一步提高了概念相似性度量的精度，更深层地挖掘了形式概念之间的共有信息与内在联系。

4 实验及结果

为了评测本文提出算法在数据稀疏情况下的推荐性能，本文设计实验在数据稀疏度较高的数据集中进行，通过以一定概率删除数据集中用户产品交互记录的方式来模拟数据稀疏性逐渐增大的推荐环境。设定消减概率 p=[0.2，0.4，0.6，0.8]，消减概率表示删除一条记录的可能性，随着概率 p 的增加，越来越多的记录从数据集中剔除，数据稀疏度逐渐增大，在固定推荐列表长度为 10 的前提下，对四种算法推荐结果的准确率进行评估。

图 4.1 展示了数据集上的四种算法的在不同消减概率下的准确率。可以明显看出 itemkNN 算法对于稀疏数据的适应性是最差的，userkNN 在 p=0.2 之后的准确率与召回率的下降速率虽然减缓，但其召回率与准确率均低于在 CNCF-1 和 CNCF-2 算法。随着消减概率的增加，传统的 itemkNN 和 userkNN 算法都会因为数据愈发稀疏而获取不到足够的邻域信息，从而直接影响了最终的推荐效果。由之前章节中的概念格示意图可以看出，如果把格中的概念看做顶点，则其 Hasse 图为连通图，从其中任意一点出发，都能到达其他任意各点，再结合递归探索过程，能够很容易访问到邻域概念，并构造候选项集，保证了在相对稀疏的形式背景下，依然能够获取足够的候选项进行推荐。综上所述，本文提出的算法在数据稀疏的情况下能够取得更好地推荐效果。

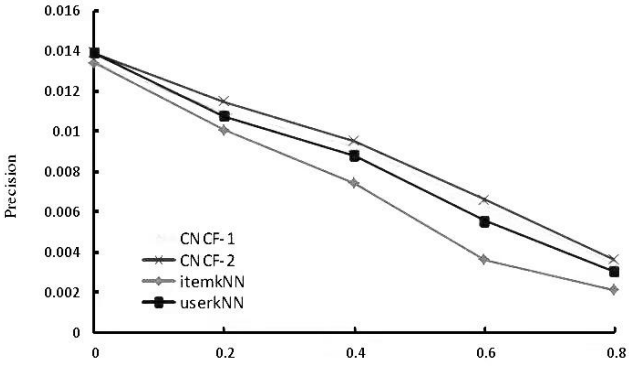

图 4.1 算法准确率实验

综上所述，当在模拟的稀疏环境下进行推荐时，本文提出算法的推荐结果的指标值均高于两种比对算法，所以整体来看，相较于传统的基于邻域的协同过滤算法，该算法能够保持较高的推荐水准，并且对稀疏数据具有更好地适应性。

5 总结与展望

协同过滤算法作为较早出现推荐算法之一，如今依然在个性化推荐领域发挥着重要作用，本文首先对形式概念分析及概念格的相关理论介绍，然后介绍了目前形式概念分析在个性化推荐领域的研究状况。并依次从构造候选项集与用户偏好度计算对 CNCF 算法的核心部分做了详细描述。构造候选项集中之所以采用递归方式是为了充分扩充候选属性（产品）集合。为了在候选项集中进一步筛选，提出用户对产品的全局偏好度与邻域偏好度，通过单独使用全局偏好度与将全局与邻域偏好度结合衍生提出了效用函数。

最后在实验部分，主要对稀疏环境下的推荐效果进行了评估。通过在数据稀疏环境下的比较，本文提出算法具备更好的推荐效果。

下一步的工作为，由于形式背景中只包含二元关系，所以算法对于结构类似的隐式反馈数据具有较好的操作性。而现实中却也存在着一些直接体现用户对产品喜爱程度的显式数据，例如用户对项目的评分。虽然存在着多值背景向单值背景转化方式，但是转化过程不可避免地存在着信息的损失。后续会尝试将算法推广到多值的模糊背景下，结合模糊概念格理论，直接在多值背景进行推荐问题求解。

参 考 文 献

[1] Boucher-Ryan PD,Bridge D,et al. Collaborative Recommending using Formal ConceptAnalysis[J].Knowledg-e-Based Systems, 2006, 19(5): 309-315.

[2] Dmitry I. Ignatov,Sergei O. Kuznetsov,et al. Concept based Recommendations for Internet Advertisement[C]. 6-th International Conference on Concept Lattices and Their Applications, 2008.

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

[4] 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.

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