程序设计竞赛组队推荐系统研究与设计

摘 要

大学生程序设计竞赛（如ICPC、CCPC等）在当下的高校中有着非常高的热度，能够有效的锻炼参赛选手的编程能力。竞赛以三人组队形式参加，然而目前学校内的队伍组成还是以相熟的同学互相组队为主，缺乏了组队的多样性和团队知识覆盖的全面性。与此同时，高昂的服务器价格与后续的维护费用阻碍了学校为集训队搭建专门的服务平台。

针对上述问题，本文设计了基于推荐系统和大数据处理技术的程序设计竞赛团队训练系统，为参赛选手提供了组队推荐、题目推荐、个人数据可视化等服务；为参赛团队提供了训练计划列表，团队知识覆盖程度可视化等服务；为学校提供了参赛队伍导出、构建竞赛代码库等服务。

系统主要包括四个部分：数据收集模块、大数据特征处理模块、推荐模块与系统业务模块。对各个模块所用技术以及功能作用说明如下：

1. 数据收集模块：分为离线数据收集与在线数据收集。离线收集采用go-colly框架分布式爬取用户在Codeforces等程序竞赛网站的提交记录；在线收集使用zap与lumberjack进行日志的收集与切割。

2. 大数据特征处理模块：离线收集的数据上传到HDFS中进行存储。Spark使用这些数据生成Embedding向量供接下来的模型训练使用。

3. 推荐模块：使用TensorFlow训练NerualCF模型与双塔模型，并通过TensorFlow Serving提供gRPC和RESTful API 调用。

4. 系统业务模块：前端使用Vue.js + Vite + Element-Plus UI完成业务页面的开发与设计；使用Go语言的Gin框架完成后端开发，它是连接起整个的系统的桥梁，完成了响应前端请求、记录日志、特征拼装、请求模型服务API等任务。

关键词　程序设计竞赛；推荐系统；大数据处理；Embedding

# 绪论

## 研究背景及意义

据最新统计，我国互联网使用人数达到9.89亿，覆盖了全国70.4%的人口，是世界上最大的互联网使用群体（截止至2020.12）[1]。庞大的使用人数，使得如今的互联网有着“人数多、范围广、场景杂”的特点，从而对高校所培养的计算机人才素质提出了更高的要求。

程序千变万化，不变的是它的内在逻辑与思维方式。算法作为程序设计的灵魂所在，在近些年越来越受到国内外高校的重视与青睐。于是各项程序设计竞赛应运而生，高校希望能够通过算法竞赛来以赛促学，激发学生们学习算法的积极性，为学生之后的工作或学术研究打下良好的基础。

国际大学生程序设计竞赛（ICPC）是一项久负盛名的世界性编程赛事，在高校中有着非常高的热度[2]。参加算法竞赛能够有效的锻炼参赛选手分析问题的建模能力与算法实现的编程能力。竞赛以三人团队形式进行，挑战在5小时内编程解决8-13道复杂问题，十分考验团队成员的协同配合能力。然而笔者在经过实际的竞赛训练经历与观察之后，发现在训练与组队的环节还存在着若干问题：

**1. 个人训练问题：“题量过大，信息过载”。**在程序竞赛高度火热的今天，早已不是10年前只有少数几个经典在线测评平台（Online Judge, OJ）的时代了。Codeforces、POJ、HDU等OJ上的题目集都在7000-8000道左右，这个数量还在不断增长当中。参赛选手很难在如此庞大的题目集中合理挑选出自己所需要的训练题目。

**2. 团队组合问题：“熟人互组，依赖经验”。**在每届新生入学之后，因为互相间还不熟悉，人们更倾向于选择自己的同班同学组队。而对于不同班级之间，缺少一个平台能够让其互相了解编程水平和擅长方向。

**3. 团队训练问题：“缺乏计划，定位不明”。**团队在进行训练时，大多都是在被动接受训练安排，缺乏明确且有针对性的训练计划来补齐短板、攻克难关。团队成员在队内的分工方法也时常不明确。

**4. 学校集训队问题：“工作繁琐，资源分散”。**教练的工作繁重，从训练计划到比赛安排，经常使用Excel表格处理大量数据，数据可视化。整个集训队训练资源分散在QQ群中，每一届的队员都会自己制作一些代码模板或经验总结，但是并没有得到很好的应用。

在处理架构上，高昂的服务器价格与后续的维护费用阻碍了学校为集训队搭建专门的服务平台。分布式大数据平台的出现为其提供了一种较为经济实惠的解决方案。开源框架Apache Hadoop可以使用多台廉价的机器组成一个分布式集群，它的三个核心组件HDFS、Yarn、MapReduce构成了大数据处理的三驾马车。HDFS提供了分布式文件存储，Yarn管理整个集群的资源和任务调度，MapReduce实现分布式的并行计算。大数据批处理平台Spark的出现解决MapReduce运行效率低下的痛点，为更快的分布式处理数据提供了支持。这些技术的出现与成熟，为本论文的设计与实现提供了有力支撑。

综上所述，在程序设计竞赛盛行背景下，针对竞赛团队组成过程中所存在的问题，设计基于推荐系统和大数据处理技术的组队推荐系统，对参赛选手们的学习与比赛，有如下几点意义。

（1）对参赛选手：提供题目推荐，使其不再陷入茫茫题海；提供队友推荐，扩大视野范围，选择更适合自己的队友；个人数据可视化，让长处短板一目了然，更清晰的认知自己的实力。

（2）对参赛团队：提供了训练计划列表，让团队的实力提升有条不紊；提供团队知识覆盖程度可视化，及时发现存在的短板，取得更好的成绩。

（3）对集训教练：提供了各个队伍训练情况可视化面板，训练情况轻松掌握；构建集训队知识库，让每一届学生们的知识汇聚于此，薪火相传。

（4）对系统搭建：利用大数据技术，让平台不在依赖昂贵的专用服务器。通过集群也可以支撑起海量日志的存储、处理与分析。

## 国内外研究与发展现状

### 推荐系统研究与发展现状

协同过滤作为推荐算法中影响力最大、应用最广泛的模型，其研究可以追溯到1992年[3]。Xerox的研究中心开发了一种邮件筛选系统，用来过滤一些用户不感兴趣的无用邮件，协同过滤（Collaborative Filtering, CF）算法第一次出现在了人们的视野中。但是协同过滤在互联网领域大放异彩，还是源于电商巨头Amazon对于协同过滤的应用。2003年，Amazon发表了工业界的知名论文Amazon.com Recommenders Item-to-Item Collaborative Filtering[4]，这不仅让Amazon的推荐系统广为人知，更让协同过滤成为很长时间的研究热点和业界主流的推荐模型。协同过滤可以根据相似度计算对象的不同分为UserCF和ItemCF。其中UserCF更符合人们直觉上“兴趣相似的朋友喜欢的物品，我也喜欢”的思想，但是从技术的角度，也存在一些缺点。首先是为了快速找出Top n相似用户，维护的用户相似度矩阵会以n2 的速度快速增长，对存储系统造成了十分巨大的压力。其次针对用户稀疏历史行为时的准确率非常低，在用户反馈困难的领域或是新加入的用户效果不佳。而ItemCF同样也存在着自己的问题：不具备很强的泛化能力，无法将两个物品相似这一信息推广到其他物品的相似度计算上，从而导致热门物品强烈的头部效应和冷门物品无法得到足够的推荐。

2006年，矩阵分解算法（Matrix Factorization, MF）在Netflix举办的奖金挑战赛中表现出色。主要优点在于其利用矩阵分解解决了CF模型在处理稀疏矩阵上的不足。矩阵分解方法分为利用特征值分解、奇异值分解和梯度下降。但MF算法依旧没有解决如何将更多的特征（如上下文信息）加入推荐模型中。

逻辑回归模型解决了特征融合的问题，它不再依靠相似度计算去对推荐列表进行排序，而是期望将用户的点击率（Click Through Rate，CTR）概率作为推荐排序的基础。2012年，阿里巴巴提出将推荐样本进行分片，然后在每片中进行LR的运算，最终将加权平均概率作为最后的预估值，该模型被称作大规模分段线性模型（LS-PLM）[5]，这已经很接近于三层的神经网络，在深度学习的浪潮席卷之前发挥出惊人的效果。但是基于LR的推荐模型没有实现交叉特征的处理，有可能在某些情况导致辛普森悖论，出现每种特征的CTR概率都很高，但是综合之后反而很低的情况。

为了解决特征交叉的问题，人们从一开始的手动组合到POLY2模型将所有特征两两暴力组合，直到2010年FM模型提出通过引用两个特征隐向量内积作为交叉特征的权重，这才真正解决了特征交叉无法实际应用的问题[6]。传统推荐模型三大模型的演化关系如图1-1所示：

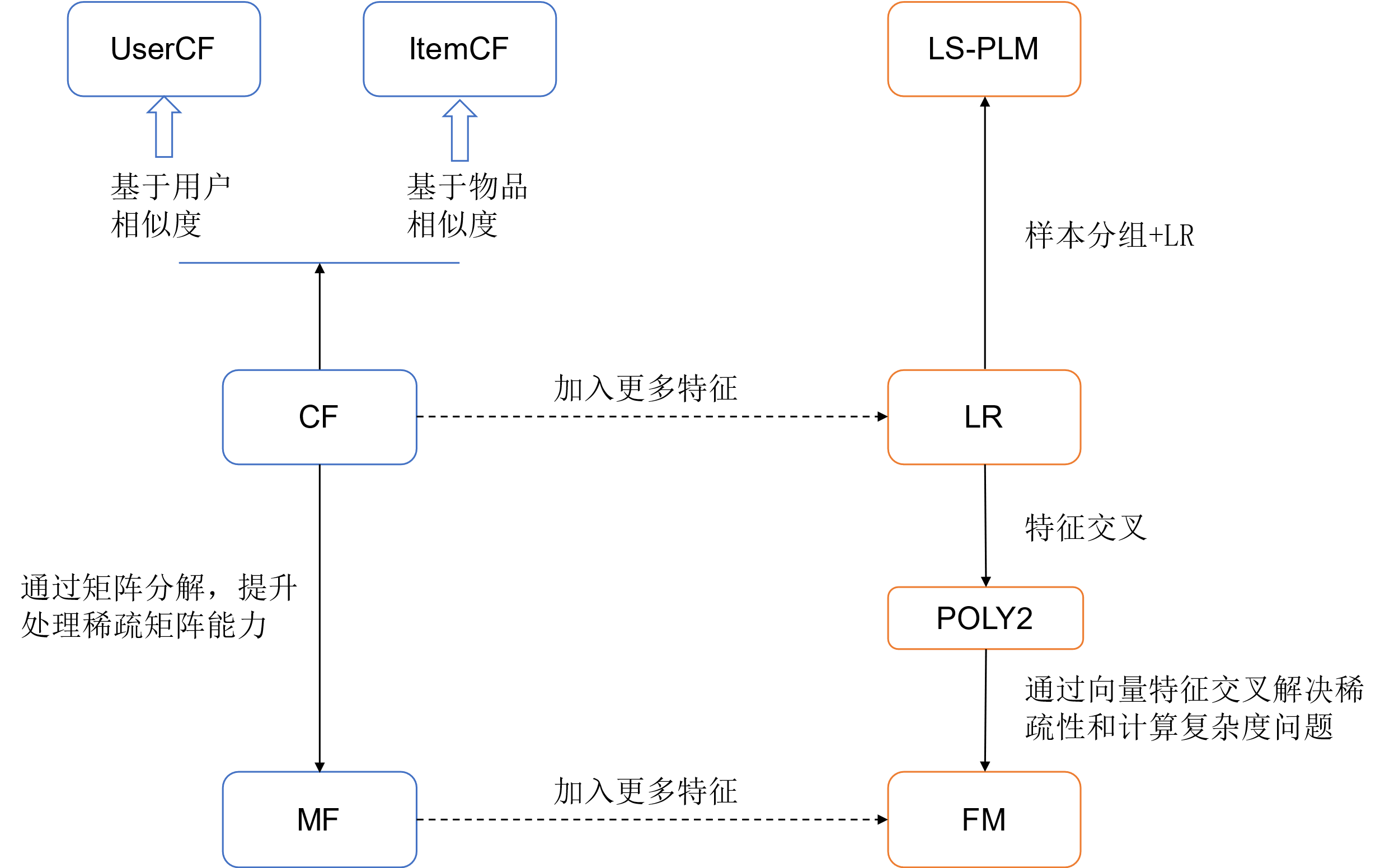


图 1‑1 传统推荐模型演化关系图

随着Alex Krizhevsky在2012年提出了AlexNet，开启了深度学习的时代[7]，推荐算法也深受其影响。而Google于2013年提出的Embedding技术—Word2Vec，解决了传统特征输入中采用One-hot编码导致的矩阵稀疏问题[8]。Barkan等人在2016年提出的Item2Vec，让Embedding技术从词扩展到序列，更加适应了推荐系统的应用场景[9]。但是这些Embedding方法还都是仅仅支持序列信息，无法覆盖图这种复杂的结构。2014年DeepWalk算法采用随机游走的方法将图模型转化为序列模型，成功的解决了Embedding无法运用在图结构的缺点。但随机游走产生的抽样性不强。这个问题再2016年被来自斯坦福大学的Grover等人解决[10]，提出了Node2Vec方法，通过调整游走过程的倾向性来实现相应场景的适应。2018年阿里巴巴公布其Graph Embedding的模型EGES，从工程学的视角解决了多Embedding的融合问题[11]。

在推荐算法的实际应用中，张廉月完成了基于Flink流处理框架的电影推荐系统的设计，其通过分布式大数据计算平台 Apache Spark、Apache Flink 以及多种开源软件，如文档数据库 MongoDB、缓存服务器 Redis、搜索引擎 Elastic Search、消息中间件 Kafka 等，搭建完整的大数据处理系统[12]。景坤使用了推荐系统完成了网约车组队的设计与实现[13]，胡开冉[14]提出了基于FunkSVD 的反向推荐可以大大提高物品的覆盖率，解决推荐算法中的长尾效应。李斌提出一种基于迁移学习的领域自适应推荐方法用于赛事推荐[15]。

### 推荐算法在算法竞赛领域的研究现状

到目前为止，很少有专注于算法竞赛推荐领域的研究：2014年Toledo等人提出了一种基于传统协同过滤的方法，并采用了适合案例的新相似度度量[16]。Year等人在2018年提出基于模糊逻辑的方法[17]。Caro和Jimenez等人考虑了基于用户相似性的方法来进行题目的推建[18]。2020年fantozzi等人使用了欧洲学生在意大利信息学奥利匹克竞赛上的数据训练了基于自动编码器神经网络的ANN模型，并测试两种不同的方法，离线数据和经过改动的增量数据，在运用深度学习推荐题目中做出了尝试[19]。

国内对于在OJ中进行题目推荐的尝试有：2015年孙权提出使用协同过滤算法进行题目推荐[20]，2018年朱国进等人采用的基于关联规则挖掘的方法进行推荐[21]，2019年肖春芸等人在华东师范的EOJ平台上使用基于用户的协同过滤算法，来为学生生成推荐题目路径[22]。知名在线测评平台LeetCode也提供了热门题目推荐功能。这些尝试都是基于传统的推荐算法，受到的局限性较大，往往无法真正的对每一个训练队员进行个性化的服务。

## 论文的主要工作

本文首先分析了目前程序设计竞赛各个环节中存在的不足，从个人训练、团队组合到教练员的繁重工作。然后从三个不同的视角给出对应的解决方法。最后使用了大数据技术和基于Embedding与深度学习的推荐算法，在程序设计竞赛组队推荐领域做出了一次尝试。以上三点也是本论文的创新所在。希望该系统设计与实现可以有效的帮助到集训队的同学进行更高效的训练以及减轻教练员重复性的劳动。

系统主要包括四个部分：数据收集模块、大数据特征处理模块、推荐模块与系统业务模块。现对各个模块所用技术以及功能作用说明如下：

**1. 数据收集模块：**分为离线数据收集与在线数据收集。离线收集采用go-colly框架分布式爬取用户在Codeforces等程序竞赛网站的提交记录；在线收集使用zap与lumberjack进行日志的收集与切割。

**2. 大数据特征处理模块：**离线收集的数据上传到Hadoop HDFS进行存储。Spark使用这些数据生成Embedding供接下来的模型训练使用。

**3. 推荐模块：**使用TensorFlow训练NerualCF模型与双塔模型，并通过TensorFlow Serving提供gRPC和RESTful API 调用。

**4. 系统业务模块：**前端使用Vue.js + Vite + Element-Plus UI完成业务页面的开发与设计；使用Go语言的Gin框架完成后端开发，它是连接起整个的系统的桥梁，完成了响应前端请求、记录日志、特征拼装、请求模型服务API等任务。

## 论文的组织结构

本文共分为6章，各章节的安排如下：

第一章为绪论，首先介绍了本论文的研究背景及其意义。然后对推荐系统国内外的研究发展进行了梳理，调查了目前推荐算法在程序设计竞赛领域的研究进展。最后简要说明本文的主要工作和组织结构安排。

第二章为相关技术和理论基础，首先对推荐算法的原理及数学推导做了详细的说明，然后对其评测指标做了简要介绍。主要介绍的算法有基于内容的推荐、协同过滤、矩阵分解、Embedding技术和NeuralCF模型。

第三章为系统分析。在进行系统的详细设计之前，进行需求分析来确定系统最终要实现的功能。最后介绍了系统运行所需的软、硬件资源。

第四章为系统设计。

第五章为系统实现与结果分析。

# 推荐算法的研究与改进

## 传统推荐算法理论研究

### 基于内容的推荐算法

基于内容的推荐算法结构简单、历史悠久，如今依旧在很多场景下表现良好。整个算法的核心思想就是为用户X推荐之前喜欢的物品相似的物品列表Y。下图2-1展示了基于内容推荐算法的Codeforces题目推荐流程：

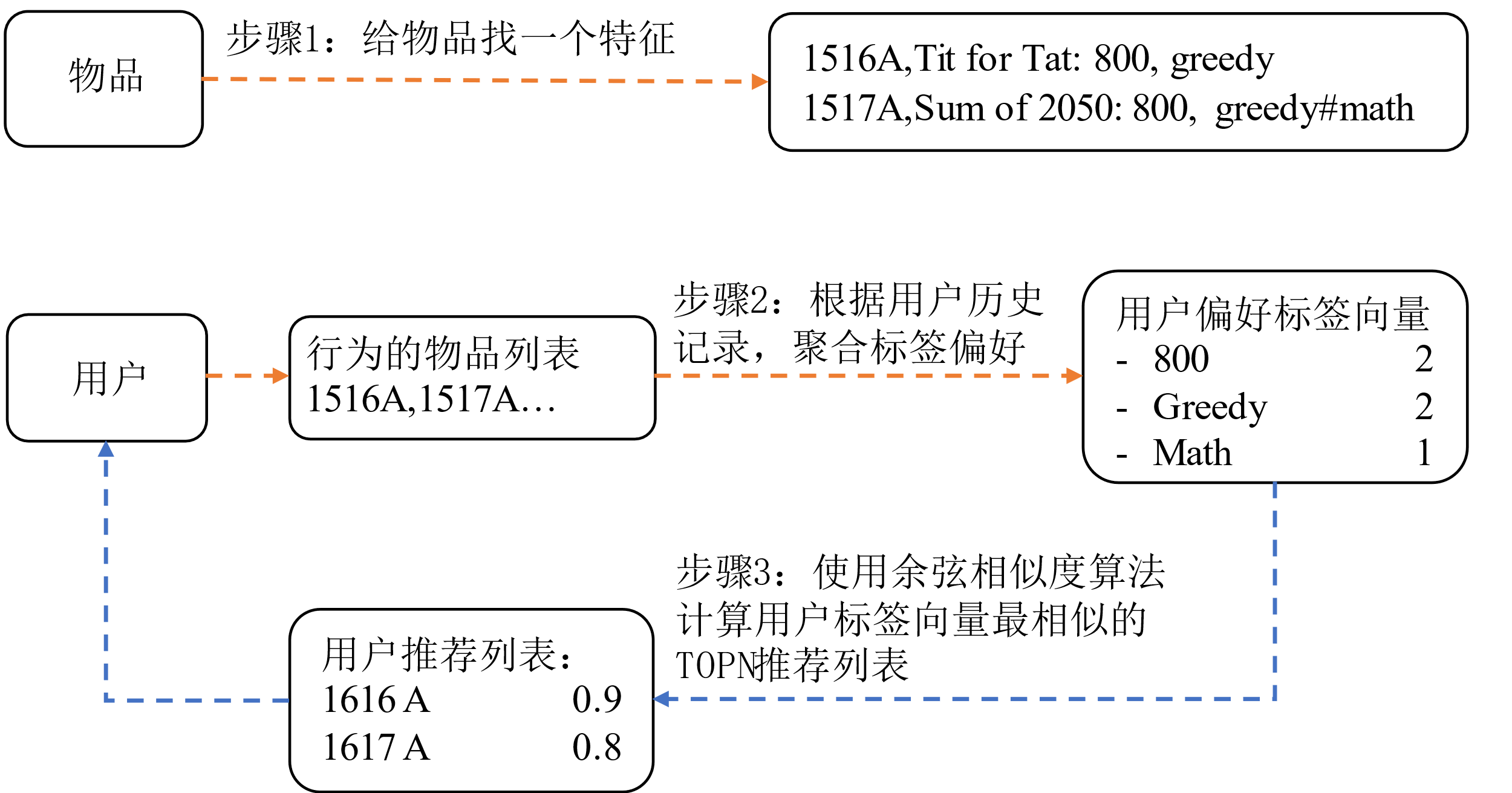


图 2‑1 基于内容推荐算法流程

**1. 步骤一**：为每个物品找一个特征，如难度系数、分类标签等。

**2. 步骤二**：收集用户的历史提交记录，比如用户A做过1516A、1517A这两道题，将题目列表标签做聚合统计，形成用户的偏好标签向量。

**3. 步骤三**：利用余弦相似度算法，计算与用户的偏好标签向量最接近的TopN列表（如上图1616A这道题和用户的偏好相似程度90%，非常可能在选择做），然后推荐给用户。

其优缺点分析如下表所示：

表 2-1基于内容推荐算法优缺点分析

|  |  |
| --- | --- |
| **优点** | **缺点** |
| 不需要其他用户的数据 | 一直处在舒适圈之中，不断推荐水题 |
| 不存在冷启动问题  推荐结果直观易于解释 | 推荐的同质化现象严重  无法与更多的特征交叉融合 |

### 常用向量相似度计算方法

在上文步骤三中提到了计算用户偏好向量和物品向量的相似程度，下面列举出几种常用的相似度计算方式，并对其含义进行说明。

1. **余弦相似度**：如（式2-1）所示，通过度量两个向量夹角的值来表示向量之间的相似程度。余弦相似度计算简单迅速，并且在高维的情况下依旧适用。

2. **皮尔逊相关系数**：如（式2-2）所示，在利用余弦相似度进行计算的时候，若用户没有评价的物品分数会被直接判0，这么做的后果会导致原本高分和低分的物品在计算完相似度后特别靠近。而皮尔逊相关系数将用户还未评分的物品填充样本均值，然后再计算余弦相似度。

3. **Jaccard相关系数**：如（式2-3）所示，杰卡德相似系数从集合的角度衡量相似度，计算方法是看相同元素占总体的比例。而杰卡德距离（式2-4）与之恰恰相反，用不同元素占总体比例来衡量两个集合的区分度。

无论选取哪种相似度计算方式都需要具体问题具体分析，灵活的相似度算法使用可以帮助我们有效提升推荐的效果。

### 基于协同过滤的推荐算法

协同过滤（Collaborative Filtering）的推荐依据是使用行为数据，利用集体的智慧来进行推荐。UserCF与ItemCF的区别与联系如下图2-2所示：UserCF是根据找到和你兴趣接近的人，将他们喜欢的其他物品推荐给你，ItemCF是找到与你所喜欢物品类似的物品，然后推荐给你。左图中User1、User2、User4都喜欢Item1和Item2，那么认为这三个人的兴趣相同，当User2和User4新喜欢一个物品Item4时，便把此物品推荐给User1。右图中User2和4同时喜欢Item2和Item3，系统便认为Item2和3是相似物品。当User1喜欢Item2时，系统也会把Item3也推荐给他。

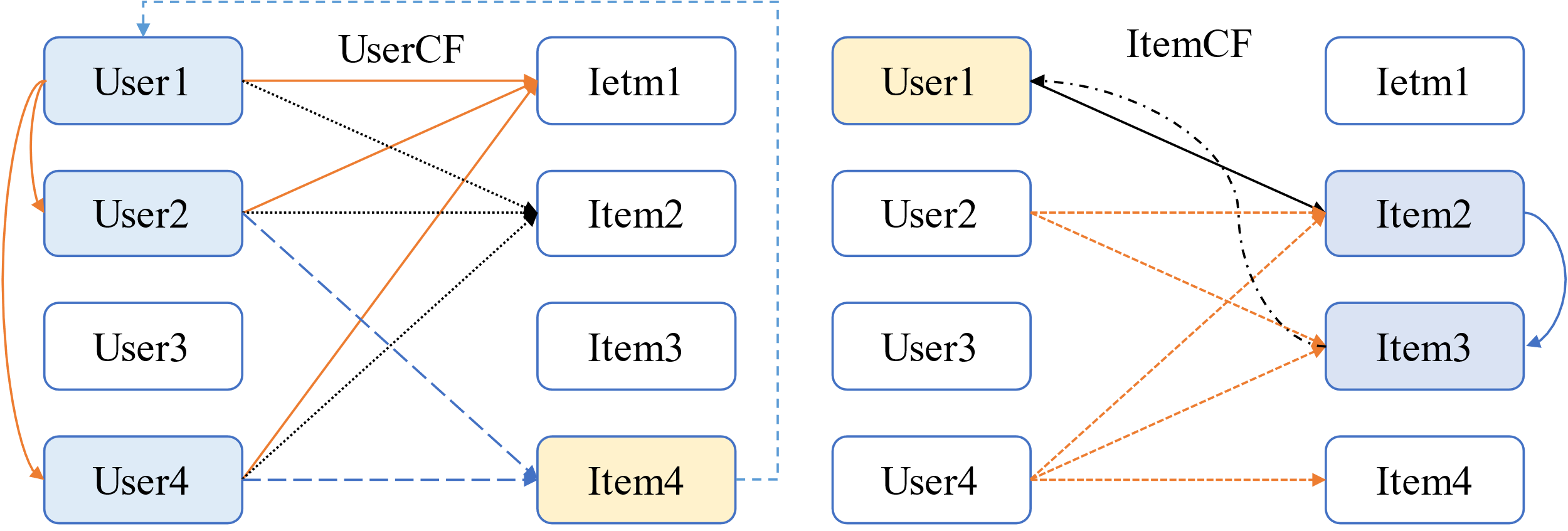


图 2‑2 UserCF与ItemCF原理

在获得了TopN个相似用户之后，我们就可以轻易的获得用户对于其还没有评价的物品的预估值。其实现方式通常为利用其他用户s对该物品p的评分的加权评分，权重为其他用户s与推荐用户u的相似度，最终得到用户u对于物品p的预估评分。整个过程如（式2-5）所示，

协同过滤算法虽然经典，但是缺点也是显而易见的，比如上图2-2里的Uer3，当他还未对某个物品进行评分时就无法找到与他相似的用户，这就是推荐系统中常见的冷启动问题。还有在物品数量和用户数量特别多的时候，维护用户-物品评分矩阵对存储的压力极大，所以许多公司选择了ItemCF作为其推荐算法。

ItemCF的具体流程如下：

1. 获得用户的历史行为信息，组成M×N维的用户-物品矩阵

2. 以列为向量，通过皮尔逊相关系数计算物品之间的相似程度，最终构建出N×N维的物品相似度矩阵。

3. 获得用户的喜爱列表（正反馈可以靠与平均值相比来界定）

4. 针对目标用户u的喜爱列表，分别每个对已评价物品生成TopN个相似物品，组成相似物品候选集。

5. 对相似物品候选集里面的物品，对其按照加权后的总相似度进行排序，最终生成排序列表。加权公式如（式2-6）所示：

其中是目标用户u对未评分物品p的预估情况，H是用户u的喜爱物品集合，权重是未评分物品p与已评分物品h的相似度，是用户u对h的评分。

下表2-2展示了UserCF和ItemCF各自的优缺点与适用场景

表 2-2 UserCF与ItemCF优缺点分析

|  |  |  |
| --- | --- | --- |
|  | **优点** | **缺点** |
| **UserCF** | 基于用户相似度进行推荐，有着更强的社交属性。 | 用户的历史行为可能非常稀疏，找到相似用户的准确率很低。 |
| 适用于发现热点，以及跟踪热点的发展趋势。 | 用户相似度矩阵的维护非常消耗资源，在线存储系统难易承受。 |
| **ItemCF** | 对冷启动优化更好，更适用于兴趣变化较稳定的应用 | 泛化能力较差，冷门的物品会很少被推荐到 |

### 协同过滤的优化—矩阵分解

随着用户和物品的不断增多，由用户-物品所构成的矩阵会以O(N2)的速度不断增大，用户的历史行为数据也随着被稀释。矩阵分解作为协同过滤算法的优化版本，其算法原理如下图2-3所示：

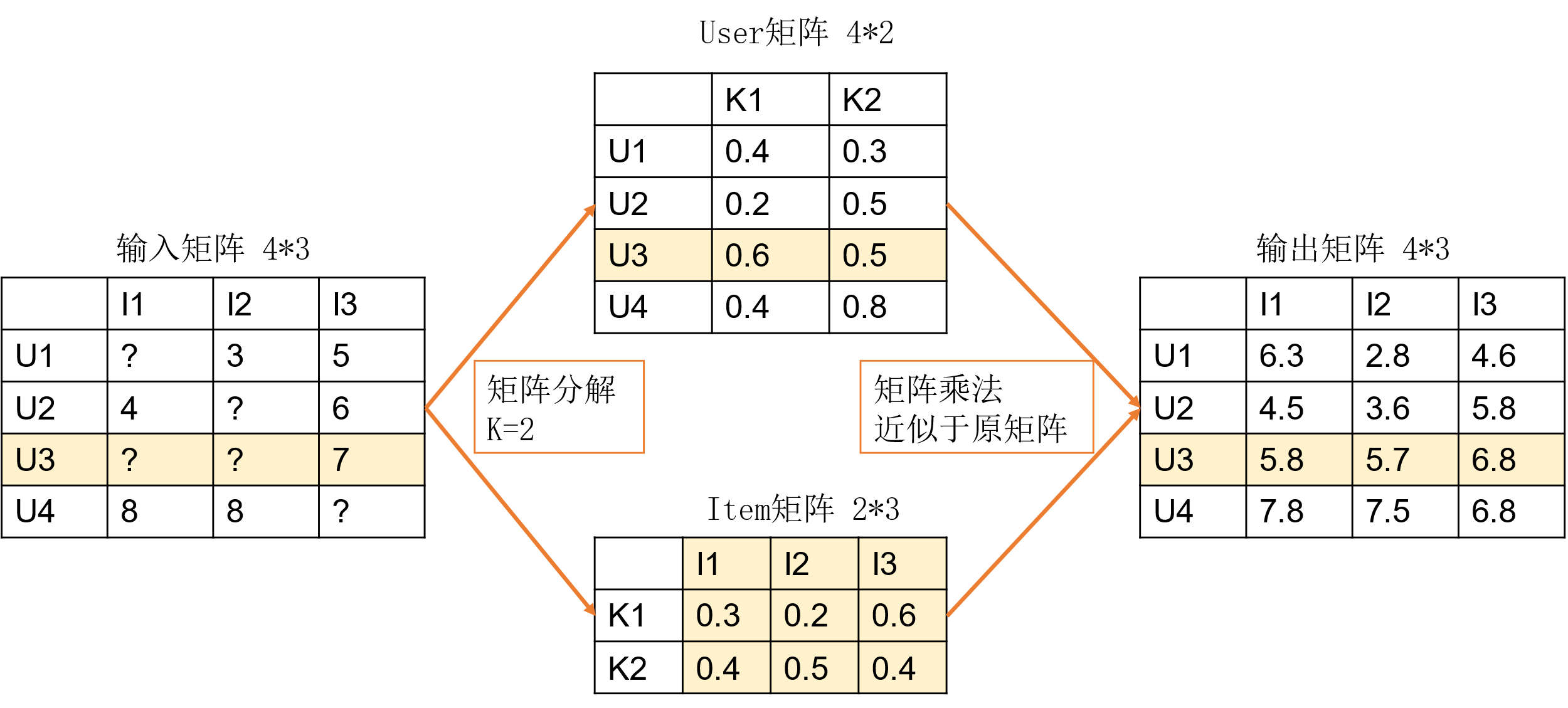


图 2‑3 矩阵分解原理

矩阵分解首先将输入矩阵分解成为为两个子矩阵（User矩阵与Item矩阵）。其中若用户数为M，物品数为N，则User矩阵为维，Item矩阵为维，K表示特征的数量。分解所得到的向量被称之为隐向量，它们用来将用户的兴趣和物品特征关联起来。而输出矩阵通过两个分解矩阵相乘后每个单元格都有值，而这些缺失值的填充值就是对用户还未评价物品的预测值。在相乘之后，输出矩阵不一定和原矩阵一致，而我们可以通过多轮训练，利用梯度下降的方法不断调整K1、K2的参数，最终让这个差值降到最小。

矩阵分解模型最终的输出是在分解后最终产生的用户特征向量和物品特征向量，它们都是K个维度，代表着K个不同的隐含兴趣点。得到用户和物品的隐向量之后，就可以通过相似度计算来进行排序，最终得到我们所需要的推荐列表。

表示用户u对物品i的预估评分，若将其表示两矩阵相乘，则如（式2-7）所示：其中表示物品i在Item矩阵中所对应的列向量，表示用户u在User矩阵中对应的行向量。

我们求解矩阵分解的目标函数如下（式2-8）所示：其表示了对于K维的用户矩阵p\*和和物品矩阵q\*，让用户已评分和通过隐向量乘积生成的“预估评分”差值最小，将预估的值尽可能的精准。

为了避免部分噪声点的影响，减少过拟合现状，对（式2-8）加入L2范数作为正则化项后如（式2-9）所示： 为正则化系数，代表约束模型稳定性的能力。

在确定损失函数（式2-9）之后，可以利用梯度下降求解最优化问题。首先，对 求偏导的结果如（式2-10）所示：

对求偏导的结果如（式2-11）所示：

利用上式的偏导结果，沿着梯度的反方向更新参数，新的参数等于旧的参数减去学习率\*变化率：称为学习率，也叫做学习步长，而刚刚求得的偏导便是参数在这个方向的变化率。更新公式如（式2-12、2-13）所示：

在不断进行梯度下降的过程中，满足两种情况即可终止训练。一是训练的轮数超过了我们预定义的次数；二是损失值（Loss）已经低于一个限度。即可结束训练，否则不断重复上述的梯度反向传播的过程。

下表2-3展示了矩阵分解算法的优缺点

表 2-3 矩阵分解的优缺点分析

|  |  |
| --- | --- |
| **优点** | **缺点** |
| 将高维的矩阵映射成多个低维矩阵的乘积，解决了数据稀疏性的问题 | 推荐结果无法直观的进行解释，其隐空间的维度无法与现实中的概念结合起来 |
| 最终的预测精度比较高 | 模型训练比较耗时，通常以天为粒度更新 |

## Embedding技术对推荐算法的改进

### Embedding向量的含义与运用方法

Embedding技术是近几年推荐系统研究的前沿，大量的论文都在攻克这一技术方向。最难能可贵的是，其在真实的应用中收获了非常好的效果。

Embedding从直观上来讲就是一个向量，里面的元素都是小数，如[0.8, 0.9]就是一个二维的Embedding向量，若有100个小数，就是一个一百维的Embedding向量。

从实际意义上来讲，每个维度上的小数表示了一个强度或是概率。如数组第一维表示数据结构、第二位表示图论。那么用户Embedding[0.8, 0.9]表示这个人擅长0.8强度的数据结构，0.9强度的图论。物品Embedding[0.4，0.6]表示这道题目有0.4的强度是数据结构、0.6的强度是图论。通过余弦相似度([0.8, 0.9],[0.4, 0.6])就可以计算出该用户做出这道题目的概率。同理，我们也可以轻松的通过运算得出用户之间的相似程度和物品之间的相似程度。但一般来说Embedding的维度都是成百上千维，很难用实际的含义去对应每一维数字代表的含义。虽然无法解释，但是这样的隐含兴趣向量就和2.1.5节中进行的矩阵分解时生成的隐向量一致，可以在推荐领域发挥出巨大的价值。下表2-4展示了Embedding的应用范围：

表 2-4 Embedding向量的用途

|  |  |
| --- | --- |
| **序号** | **Embedding向量的用途** |
| **用途一：** | 物-物推荐，相关推荐、看了又看、做了又做 |
| **用途二：** | 人-物推荐，猜你喜欢、猜你想看、猜你想练 |
| **用途三：** | 人-人推荐，猜你认识、推荐队友、找到同好 |

### Embedding向量的生成方法

生成Embedding向量的方式主要有三种：第一种是基于内容的Word2Vec方法以及后续发展出来的Item2Vec、Node2Vec等词嵌入方式；第二种方法为本文在2.1.4小节中介绍的矩阵分解，其中间生成的隐向量也是一种Embedding向量；最后一种是直接基于DNN深度学习的生成方法。

在介绍Word2Vec算法之前，需要先说明一下在其出现之前人们实现词向量化的方法——One-Hot编码。

One-Hot编码将词典集合中的词映射为一个维向量，将其本身设为1，其他位置设为0。这样就可以通过编码唯一的表示一个词。“I Love ICPC, I Like Code.”的One-Hot编码如表2-5所示：

表 2-5 One-Hot编码举例

|  |  |
| --- | --- |
| 词 | One-Hot编码 |
| I | 1 0 0 0 0 |
| Love | 0 1 0 0 0 |
| ICPC | 0 0 1 0 0 |
| Like | 0 0 0 1 0 |
| Code | 0 0 0 0 1 |

这种编码的优点就是简单，只需要将每个词Hash一下就可以为其分配编号，但是随着词数量的增多，One-Hot编码的缺点也显而易见。主要的缺点有三点：

**1. 维度膨胀极快：**随着语料库的增加，编码长度可能突破上亿维。

**2. 向量之间没有联系：**无法通过One-Hot编码之间的相似度运算得到任何有用的信息。

**3. 数据稀疏。**经过One-Hot编码后的向量几乎全都为0。

以Word2Vec为代表的Embedding技术的出现解决了上面所描述的问题。基本思想为将低维空间中的维稀疏One-Hot向量映射到高维度空间中长度为维的一个点。这里是我们指定的一个参数，表示Embedding向量的长度。然后通过简单的点积运算就可以得到两个Embedding向量的相似度。Word2Vec有两种经典的模型：CBOW和Skip-gram。它们的结构如图2-4所示：



图 2‑4 CBOW模型与Skip-Gram模型

这两种模型的区别就在于，CBOW模型是通过中心词周围的词来预测中心词；Skip-gram模型是通过中心词来预测周围词。

CBOW简化版的模型（仅输入一个词）如下图2-5所示：

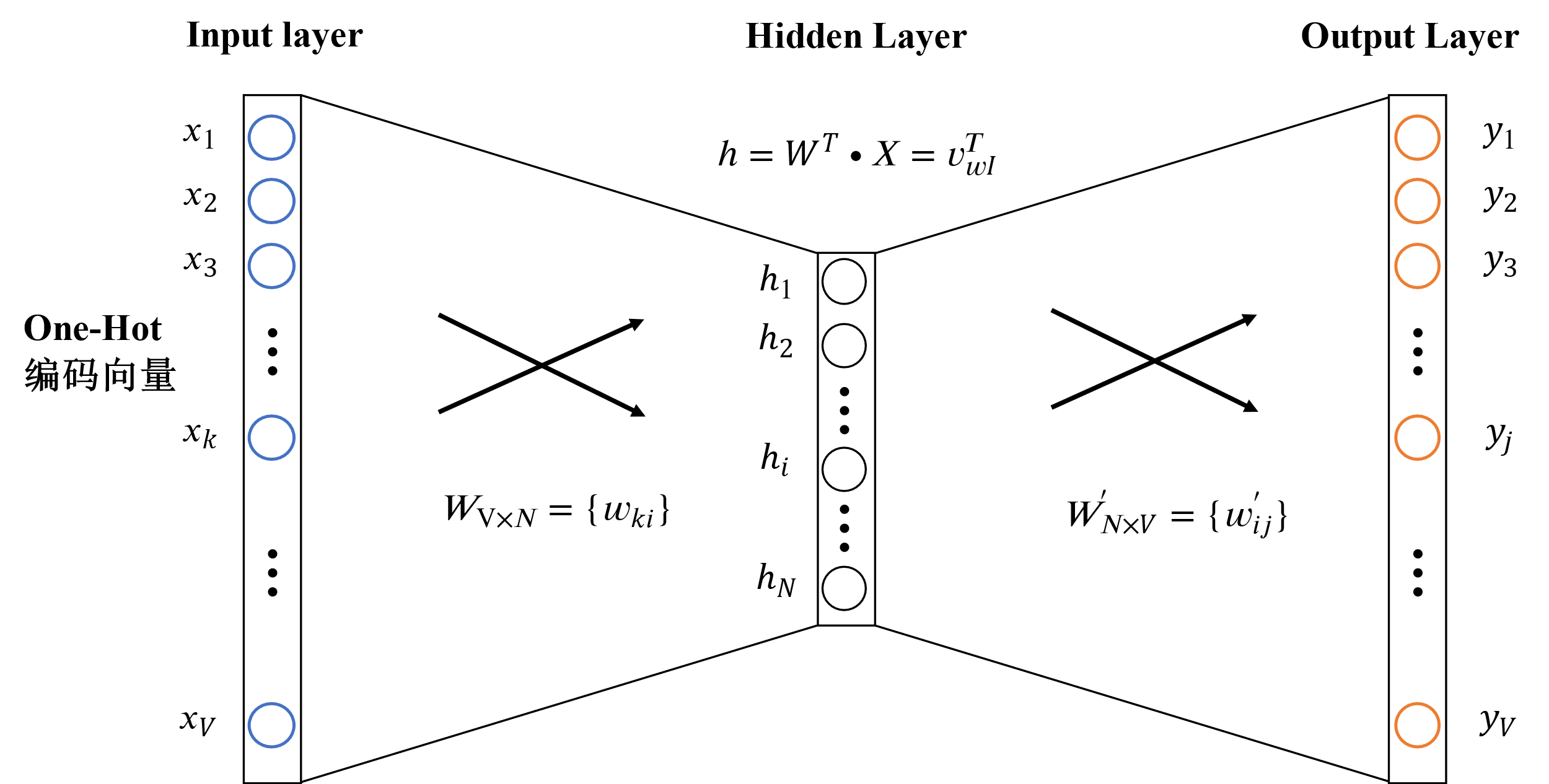


图 2‑5 CBOW简化版模型

**1. 输入层：**约定输入的词库大小为，中间的隐层大小为。若将输入单词标记为，则输入层是其经过One-hot编码后的向量表示，一共有个神经元，除了第个神经元为1，其余各项均为0，例如。

**2. 权重矩阵：**输入层向量到隐层之间的权值可以用一个的矩阵表示，其每一行的值对应着一个我们最终需要的Embedding向量。将One-Hot编码过后的向量映射为一个N维的短向量，在这里我们实现了One-Hot编码的降维。

**3. 隐层：**隐层得到的结果是由输入向量乘权重矩阵得到的。若One-Hot编码的第k位为1，相乘后隐层的值等于权重矩阵的第k行。其数学形式如（式2-14）所示：

**4. 权重矩阵**：在隐层到输出层之间有另一个权重矩阵。通过这个权重矩阵我们可以预估出集合V中每个词可能的得分。其公式如（式2-15）所示：

其中表示权重矩阵的第j列。

**5. 输出层：**在输出时将每一个词的得分通过softmax函数归一化得到一个在之间得概率，其表示为在输入词为时输出词是的概率。可以用（式2-16）对其进行表示：

将（式2-14、2-15）带入（式2-16）整理可得（式2-17）：其中和分别是输入词的两种不同的N维向量表示形式。

其优化的目标函数L如（式2-18）所示：

确定了目标函数后，就可以利用梯度下降和链式求导规则，对两个权重矩阵进行更新，整个下降的过程如图2-16所示：

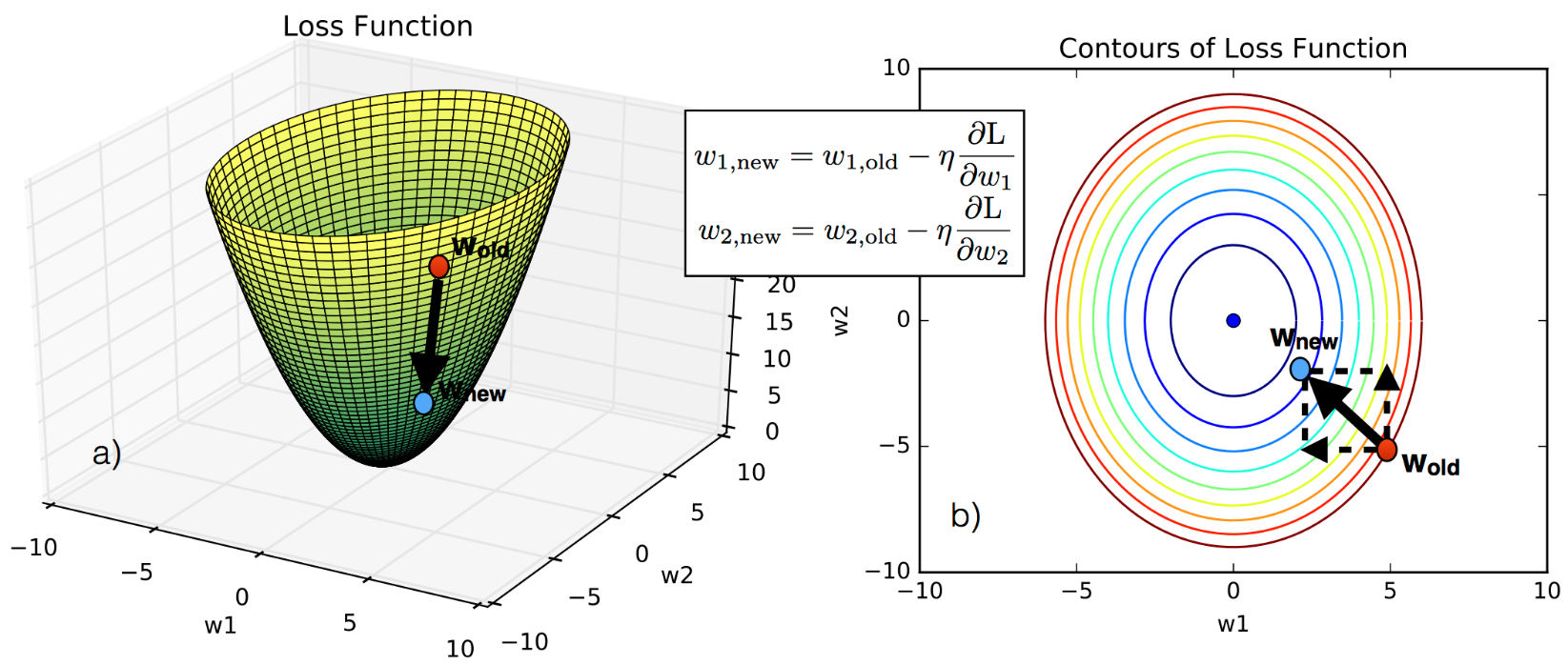


图 2‑6 Word2Vec梯度下降求最优权值

当我们有了Word2Vec之后，若将其（文本ID，文本内容）的输入变为（用户ID，用户提交ID）就可以得到一个行为上的Embedding向量，这就是Item2Vec算法。

### Embedding向量的快速检索

在生成完Embedding向量之后，如何快速得到每个用户最相似的TopN列表有三种方法。第一种方法就是离线使用两层循环暴力为每个用户的Embedding向量挨个计算余弦相似度，然后缓存到Redis里面。这种方法在数据量小的时候还可以使用，一旦数据量过大，离线计算的时间就是我们不能接受的。第二种方法是通过离线局部敏感哈希（LSH）来进行相似搜索。其原理是将Embedding向量分为很多个桶，然后在桶和桶的附近进行搜索，可以极大的增加搜索速度。第三种方法为使用Facebook开源的近邻搜索库Faiss，可以实现ms级别的在线搜索（百万级别）。

下图2-7展示了局部敏感哈希高维空间点向低维空间映射后的结果：

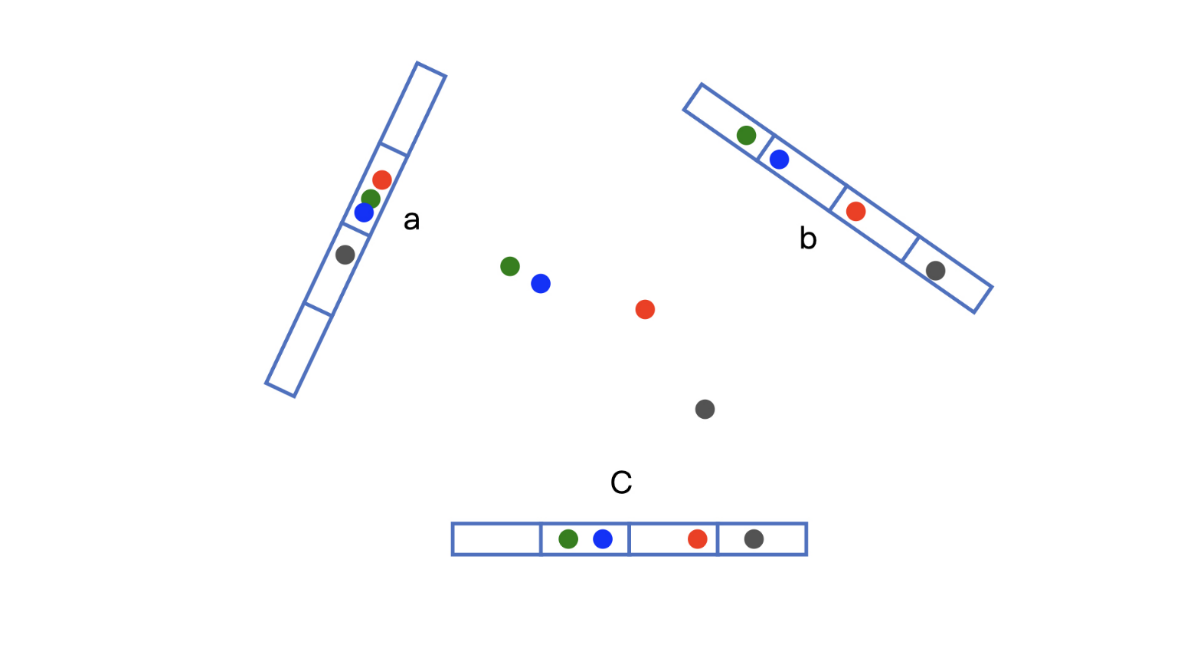


图 2‑7 高维空间点向低维空间映射

上文提到了Embedding向量是数据经过One-Hot编码后映射到高维空间的一个点，而局部敏感哈希就是将其压缩回低维的欧式空间，其原本在高维中相似的点映射到低维依旧靠近。但是原本相隔“距离”较远的点可能会变成接近的点。解决这个问题的方法是通过多个哈希函数将其进行分桶，如果两个点同时掉入多个哈希函数生成的桶中时，那么它们相似的概率就会大大提高，减少误判的情况。

## 深度学习在推荐系统中的应用

如果将矩阵分解看成一个简单的神经网络，那么其结构如图2-8所示：

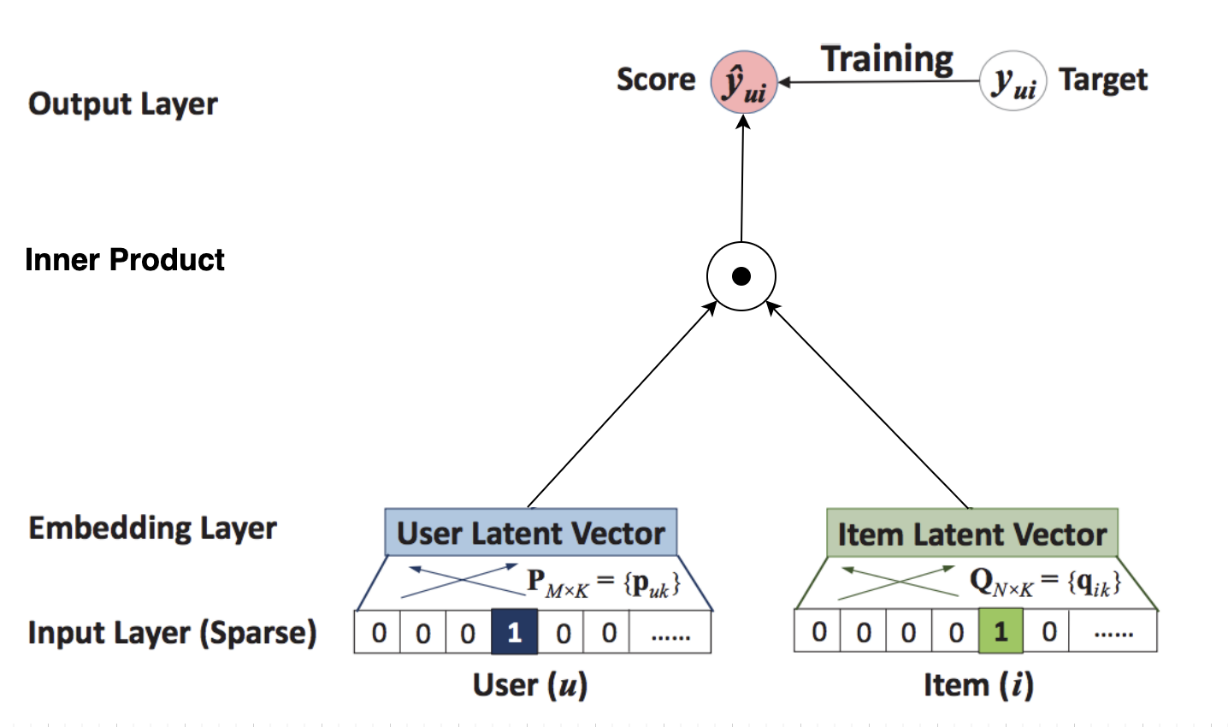


图 2‑8 矩阵分解的神经网络结构

输入稀疏的One-Hot向量，经过分解得到Embedding层，然后通过点积运算得到预测的得分矩阵，最后进行梯度下降来更新Embedding层。但是，矩阵分解在Embedding层之上直接利用内积得出最终结果，会导致特征之间还没有得到充分交叉，会有欠拟合的风险。

NeuralCF将简单的点积运算变成了多层的神经网络，提高了模型的拟合能力。其结构如图2-9所示：

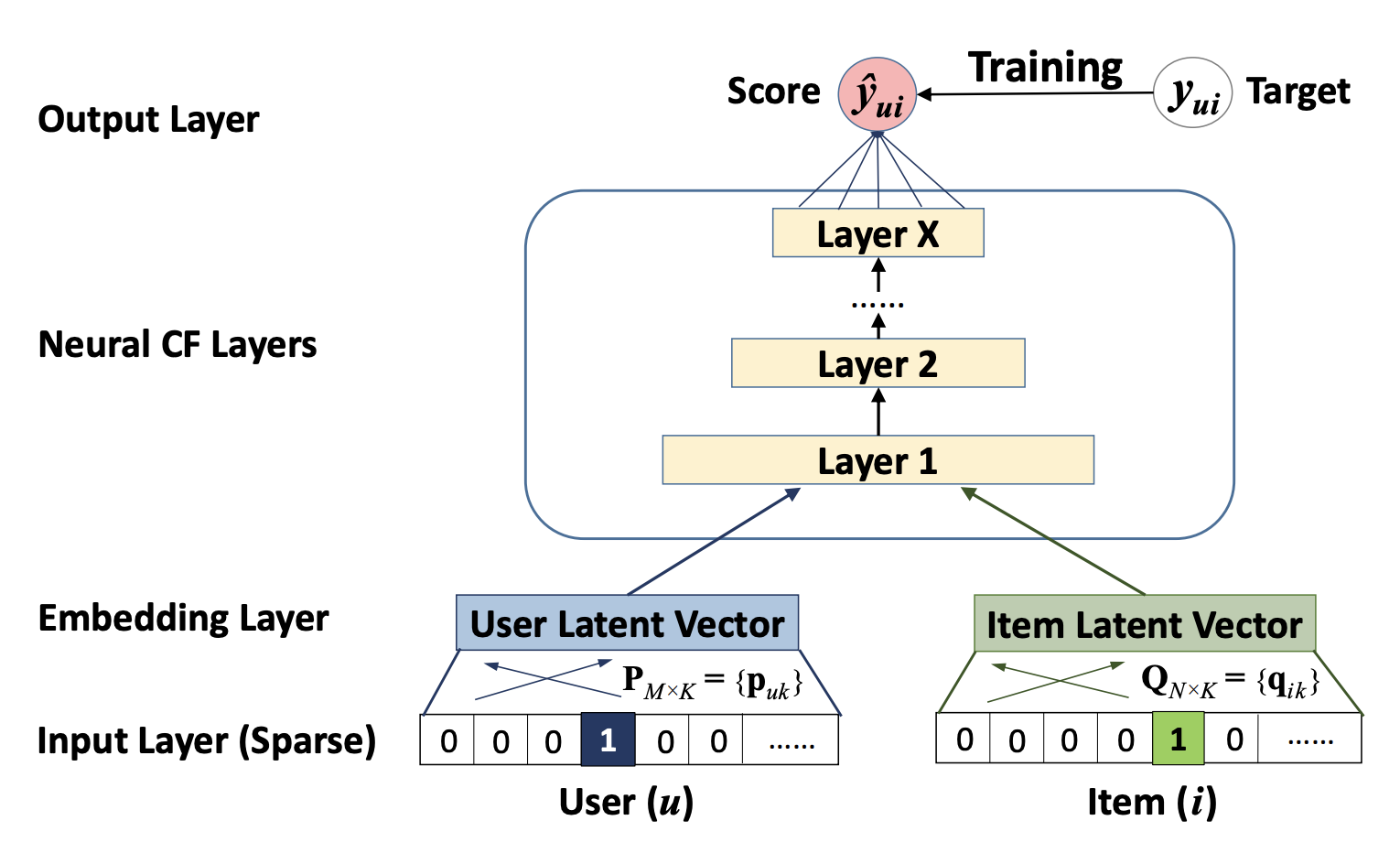


图 2‑9 NeaulCF神经网络结构

## 本章小结

首先对推荐算法的原理及数学推导做了详细的说明，主要介绍的算法有基于内容的推荐、协同过滤、矩阵分解、Embedding技术和NeuralCF模型。