前期技术路线的调研

1. 系统架构

本项目基于行为轨迹挖掘的学习伙伴推荐平台，主要包括三个功能模块：学习行为数据集的构建、学习行为轨迹数据挖掘和基于行为轨迹的推荐算法。学习伙伴推荐平台的主要业务流程和核心模块如图 1 所示：

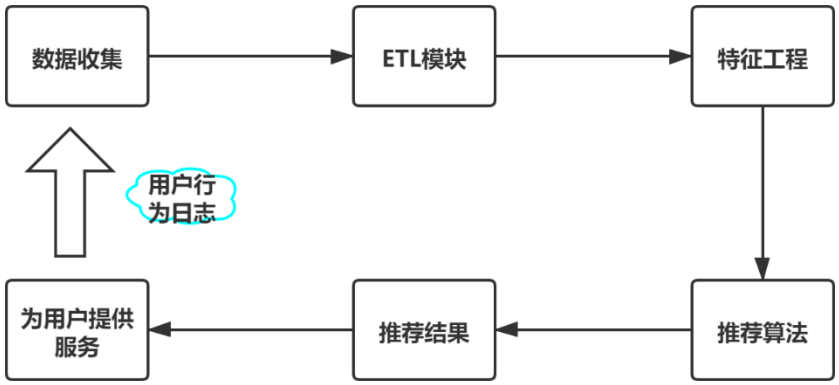

图 1 学习伙伴推荐平台架构图

学习伙伴推荐平台中，通过收集学生的学习行为数据来构建学习行为数据集，利用 ETL 进行学习行为轨迹的特征工程数据挖掘，并根据学习行为轨迹设计相应的学习伙伴推荐算法，给学生推荐具有相同学习兴趣的学习伙伴、或者推荐能够知识能力互补的互助学习伙伴。

为了提高开发效率，考虑采用前沿企业的架构模式，**构建一套通用的算法组件 Doraemon 框架，即使构建推荐业务像搭积木一样，能快速利用各种组件搭建成为一套业务流水线，使项目成果不仅仅局限于某一项功能。**并且便于发现和追踪问题，节省人力成本。具体改进如图 2 所示：

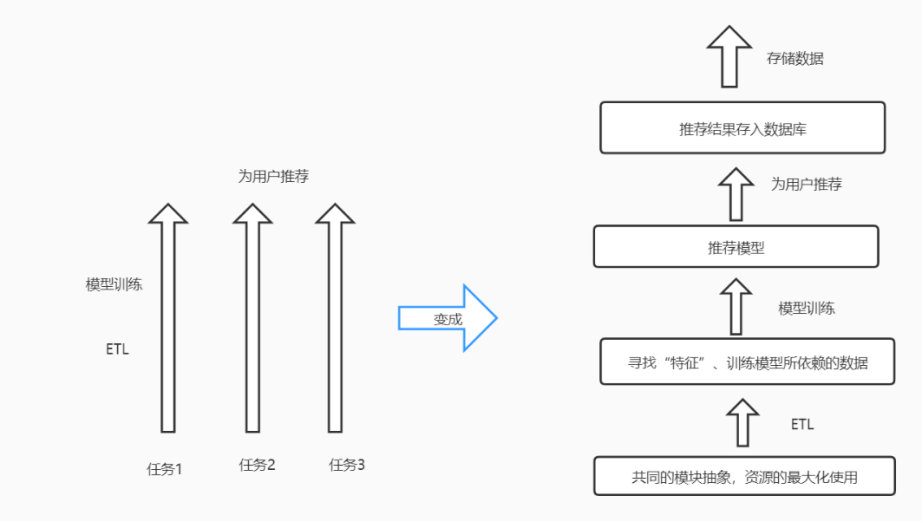

图 2 推荐系统的模块化优化图

在整个系统中，数据的处理尤为重要，对学生行为数据的整个处理周期如图 3 所示，首先通过线上途径获取用户相关学习行为数据，处理后存入数据库，供后续数据挖掘及模型训练等操作。整个数据处理是一个不断迭代优化的过程，形成一个闭环式系统。

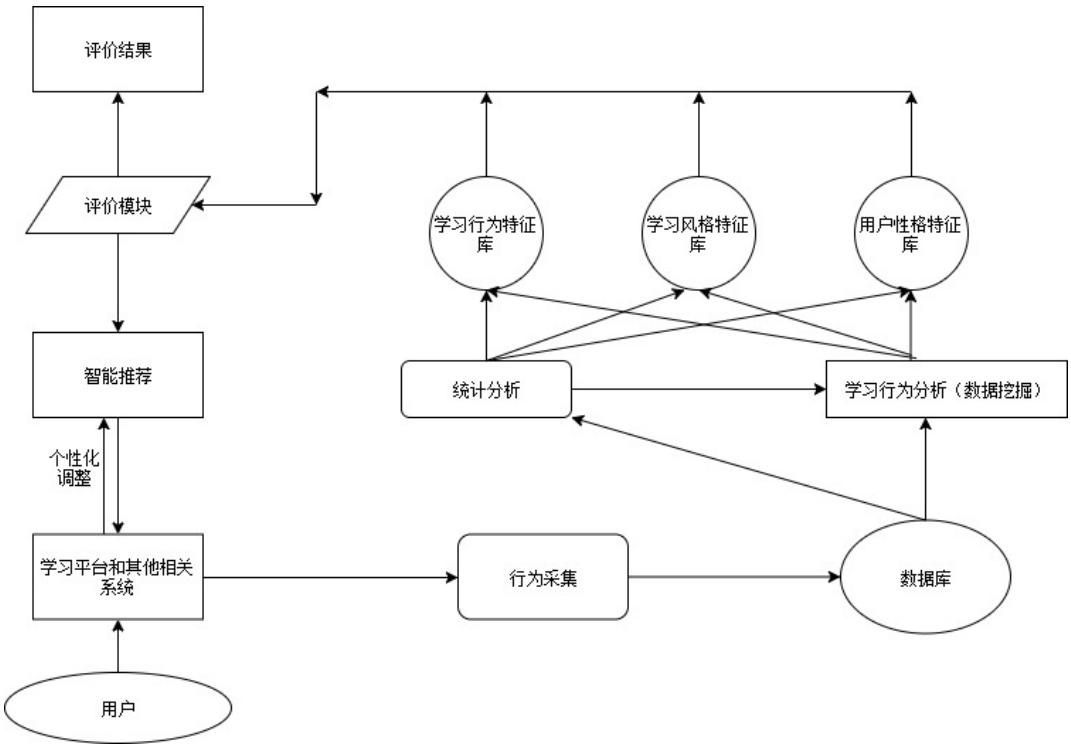

图 3 学生行为数据处理过程

2. ETL 模块

ETL 设计主要分为数据抽取（Extract）、数据转换清洗（Transform）以及数据的加载（Load）三部分，如图四所示：

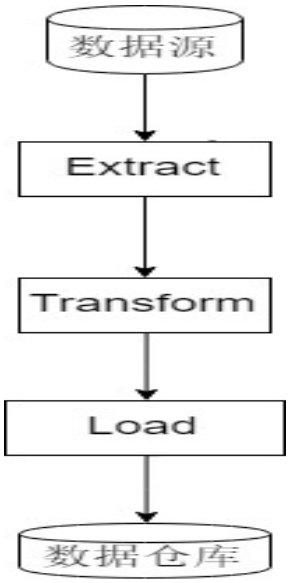

图 4 ETL 处理流程

1.数据抽取：

1.1 前期调研：弄清数据源所在业务系统的数据服务器运行什么 DBMS（Data Base Management System）,并考虑是否存在手工化数据、是否存在结构化数据。

1.2 数据源抽取处理：本项目采用的数据仓库为当前国内主流数据仓库——基于 Hadoop（名为 Hive 的大数据分析平台）的开源数据仓库。对于存放 DW 数据库系统相同的数据源，在 DW（Data Base）数据库服务器和原业务系统建立直接链接关系，然后用 Select 语句直接访问。对于数据库系统不同的数据源，则通过 ODBC（开放式数据库连接 Open Data Base Connectivity）的方式建立数据库链接，或者通过工具将源数据导出成.txt/.xls 文件,也可以通过程序接口来实现。

2.数据清洗转换：这部分顾名思义，任务就是过滤掉不符合要求的数据。这些数据主要分为不完整的数据、错误的数据和重复的数据。对于不完整的数据，要筛选出来并且补全后才能写入数据库；对于错误的数据，通过 SQL（structured Query Language 结构化查询语言）找出来，要么修正后抽取，要么剔除；对于重

复的数据，如何同一信息，从不同数据源反复收集,对于这些情况，也要经过处理。

3.数据加载：将清洗完的数据写入 DW 中去。

6.1.3 特征工程模块

关于特征工程有这样一句话：**“数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。”**所以特征工程需要最大限度地从原始数据中提取特征以供算法和模型使用。

系统采集的数据种类有很多，部分是已经整理好的学习数据特征，还有是需要进一步整理的，**对于不同数据来源的大致处理流程如图 5 所示。**

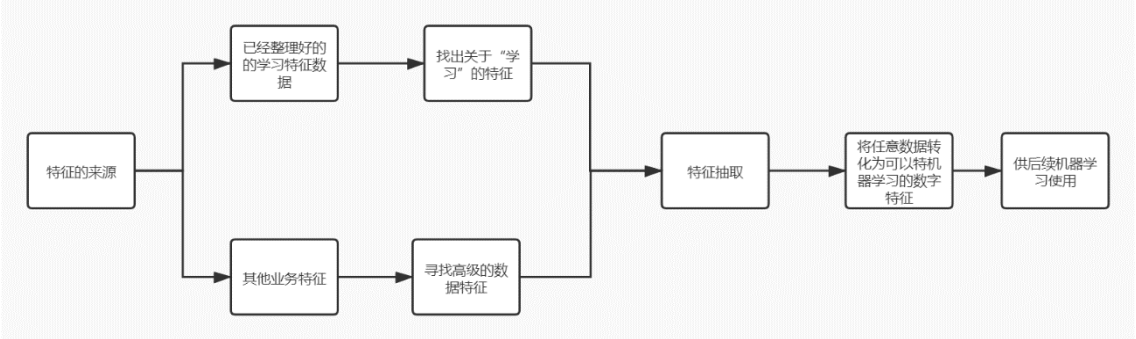

图 5 对数据特征处理的流程

推荐系统采用各种机器学习算法来学习用户的行为特征，而这些可被推荐算法用于训练的数据是可以“被数学所描述”的，**特征工程即是根据推荐算法的需要，通过特征处理及降维等操作，将 ETL 后的数据转换为推荐系统可以学习的特征**。特征工程的处理流程如图 6 所示。

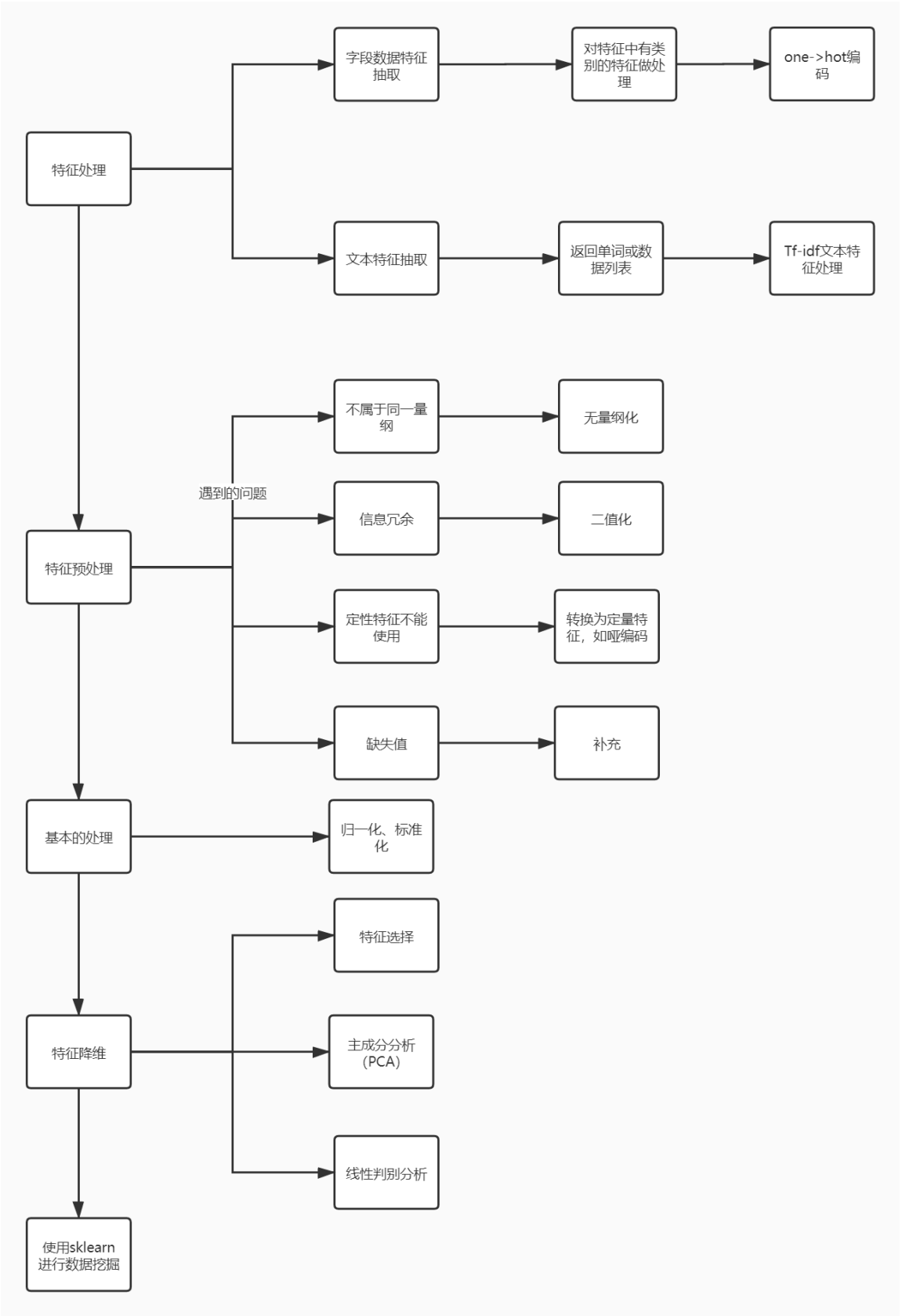

图 6 特征工程的处理流程

3. 基于学生行为数据的模型构建及智能推荐算法

• 关于模型的构建

基于用户的行为记录（常去的地方、常看的书），用户产生的相关信息（年龄 、性别、成绩信息），构建学生学习行为算法模型。大数据背景下分析在线学习平台中记录的大量用户学习行为数据,通过追踪用户在学习过程中产生的各种学习行为数据，构建层次性模型，进而得出分析结果。

• 推荐算法包含模型训练、预测两个核心操作，模型训练是将搜集到的数据处理后供推荐算法进行学习，模型预测就是根据推进算法，基于构建好的模型，对符合用户需求的结果进行预测。模型训练及模型预测概览图见图 7。

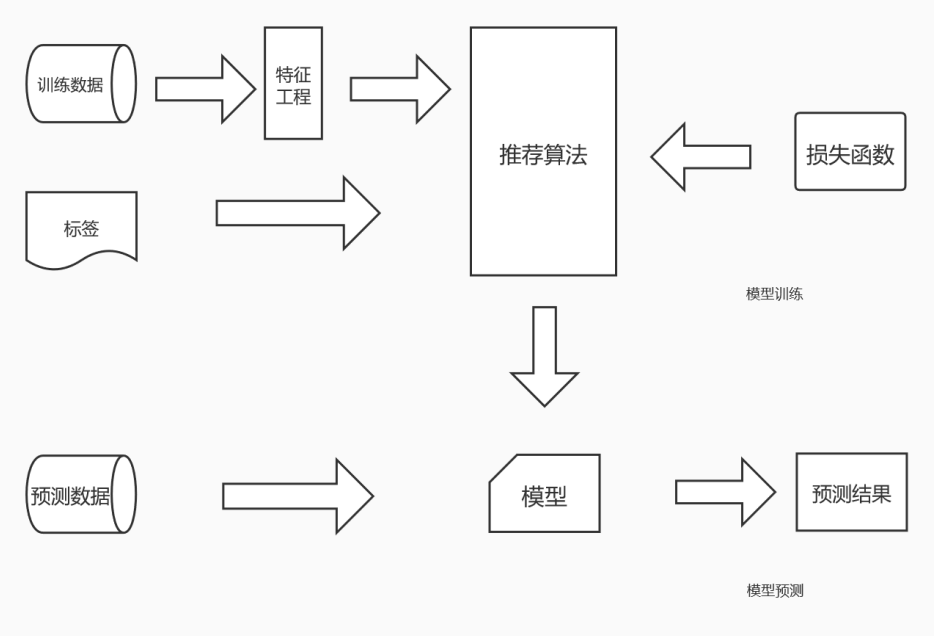

图 7 模型训练及模型选择

考虑到本项目的用户需要“推荐物”，同时又是“被推荐物”的特殊关系，推荐算法采用“人以类聚”的协同过滤和基于模型的推荐算法。首先建立好学生行为数据模型，根据用户同一或相似的一些行为特征，将用户进行分类，然后根据其他数据进行精细化推荐。

4. 该项目的其他模块

•推荐结果储存模块：因原始数据已经足够支撑推荐结果，不考虑实时更新问题，采用 CouchBase 等可以横向扩容的数据库

• 服务模块：为实现用户个性化多种推荐结果，出基本相似结果的推荐外，考虑与搜索引擎结合，提供更优服务

项目拟解决的问题:

• 当前高校在校生一个人学习效率低下，难以找到志同道合学习伙伴的问题

• 众多大学生因缺乏学习伙伴，有问题不能及时与他人交流，互相提高的问题

• 通过组件化、模块化解决推荐系统业务开发时效率低下、资源浪费的情况

项目预期成果

• 系统能通过对用户各行为数据的分析，智能推荐志同道合的学习伙伴

•用户可根据自己的需求对不同类型的学习伙伴（如具有共同兴趣的、可以互帮互助的、可以组队学习的）进行个性化选择

• 推荐系统的构架做到通用性、模块化、组件化，能有的学习行为模型