设计研究与应用

自适应推荐系统设计与实现

陈明辉，董 晶

（华北计算技术研究所 公共安全信息化事业部 北京 100083）

摘 要： 推荐系统是现代电子商务中必不可少的系统。由于数据和计算的复杂度，导致普通应用维护和使用推荐系统变得不容易。文中设计和实现了自适应推荐系统平台。通过自适应推荐系统平台，用户可以不需要大量计算资源而进行复杂而缓慢的推荐系统计算任务，只需对每个推荐任务进行算法和相关细节的定制，就可以实现推荐模型的训练。文章最后展示了系统的主要界面，并介绍了使用方法。

关键词： 推荐系统；大数据；自适应；消息队列

中图分类号： TP311 **文献标识码：** A **DOI：** 10.3969/j.issn.1003-6970.2016.12.036

本文著录格式： 陈明辉，董晶. 自适应推荐系统设计与实现[J]. 软件，2016，37（12）：169–175

Design and Implementation of Adaptive Recommendation System

CHEN Ming-hui, DONG Jing

(Public Security Information Division, North China Institute of Computing Technology, Beijing 100083, *China*)

【Abstract】: The Recommendation system is now essential in e-commerce system. Due to the complexities of data and computation, it is not easy to maintain and use recommended systems for general applications. In this paper, an adaptive recommendation system platform is designed and implemented. Through the recommendation system platform, the user can perform complex and slow recommendation system calculation tasks without large amount of computing resources. At the end of this paper, the main interface of the system is presented and introduced.

【Key words】: Recommendation system; Big data; Adaptive system; Message queue

0 引言

互联网的出现和普及给用户带来了大量的信息，满足了用户在信息时代对信息的需求，但随着网络的迅速发展而带来的网上信息量的大幅增长，使得用户在面对大量信息时无法从中获得对自己真正有用的那部分信息，对信息的使用效率反而降低了，这就是所谓的信息超载（Information Overload）问题[1]。在大数据背景下这种情况尤为突出。

目前，针对信息超载，主流的解决方法是搜索引擎。通过搜索引擎，用户可以通过输入查询字段进行相关资料的搜索。然而，早期的搜索引擎并不能够根据不同的用户返回不同的搜索结果，有必有引入一种新的解决方案，这就是推荐系统[2]。

推荐系统的起源可以追溯到认知科学和信息检索等领域的相关研究，它与管理学以及市场营销中的用户行为建模也有密切的关系。上世纪九十年代中后期，互联网迎来了蓬勃发展阶段，信息和商品的数量和种类均呈现快速的增长，用户需要花费大量时间才能找到自己重要的信息或商品。在这样的背景下，协同过滤为代表的推荐技术发表，并受到学术界和工业界的广泛关注。推荐系统，作为一种信息过滤工具，逐渐成为一个独立的研究分支。

现有的推荐系统都是针对某一具体问题进行实现，而实际情况中，尤其在大数据环境下，我们需要一种自适应平台来特定问题进行领域建模，以期实现自适应的分析。

1 推荐系统介绍

1.1 推荐系统数学模型

从 1995 年开始，推荐系统就开始成为了一个独立的研究领域，一般来讲，推荐系统的本质就是为

作者简介： 陈明辉(1992–)，男，硕士研究生，主要研究方向：数据处理与分析；董晶，女，研究员级高级工程师，主要研究方向：信息系统顶层设计、信息集成。

用户 u 推荐感兴趣的物品 v。推荐算法的主要是根据用户的行为和兴趣为用户推荐用户可能感兴趣的物品。数学表征如下：

$$V_u = \arg\max F(u,v)$$

Subject to:

$$\begin{cases} F: U \times V \longrightarrow R \\ \forall u \in U \end{cases}$$

其中，函数 F 用于度量用户 u 对物品 v 的效用函数。效用函数通常用来表示消费者在消费中所获得的效用与所消费的商品组合之间数量关系的函数，以衡量消费者从消费既定的商品组合中所获得满足的程度。目标函数即为满足最大化用户效用的物品推荐。因此推荐系统本质是一个最优化问题模型，问题的关键就在于如何求解这个最优值。

1.2 推荐系统应用分类

依据应用场景的不同，推荐系统主要分为以下几类：

1）视频推荐。视频推荐最早最成熟的莫过于 Netflix，在推荐系统领域，同 Amazon 是最有代表性的两家公司。Netflix 在相关资料中声称，有 60% 的用户是通过推荐系同获发现新兴趣。其他如 Hulu，youtube，优酷，爱奇艺，都有相应的推荐系统进行视频推荐。

2）音乐推荐。音乐推荐最典型的比如网易云音乐和豆瓣音乐。其中豆瓣音乐不允许用户点歌，而是通过用户的反馈喜欢，不喜欢或者跳过来训练用户的兴趣模型。网易云音乐在提供在线音乐搜索和电台服务。

3）社交网络推荐。社交网络国内以微信，微博最热。国外则是 FaceBook 和 twitter。在社交网络中，用户可以发挥社群优势，推荐和分享好多有用的知识和物品。社交网络中的推荐系统主要包括好友推荐，物品推荐和群组推荐。

4）阅读推荐。已经下线的 Google Reader 作为早期一款流行的社会化阅读工具。它允许用户阅读关注用户分享的文章。现在最流行的杂志推荐系统 Flipboard，其前身是 Zite，同为流行的阅读推荐工具。它可以将 Facebook 或者微博上流行的内容整合起来以杂志的形式提供给用户阅读。

5）基于位置的推荐。基于位置的推荐通常见于基于地图的应用上，如大众点评，美团，饿了么外卖和百度地图的附近搜索。随着移动设备的飞速发展，位置信息作为一种隐形公开的数据很容易被 App 后台获取，位置作为一种很有效的上下文信息，可以基于位置为用户提供感兴趣的餐饮，娱乐等服务，引导用户消费。

6）个性化邮件。通常，个人每天都收到很多邮件，其中部分邮件对我们来说非常重要，有些不太重要和一些垃圾邮件。垃圾邮件可以通过垃圾邮件过滤器删除，这是一个专门的研究领域，这里不讨论。但在普通邮件中，如果你能发现消息的用户重要，允许用户浏览，毫无疑问会大大提高用户的生产力。

1.3 推荐算法分类

大部分的推荐算法其工作原理还是基于物品或用户的相似性进行推荐，大致上可以分为如下几种：基于人口统计学的推荐（Demographic-based Recommendation）[3]，基于内容的推荐(Content-Based Recommendation)[4]，以及基于协同过滤的推荐（Collaborative Filtering-Based Recommendation），以及混合型推荐（Hybrid Recommendation）[5]基于协同过滤的推荐被研究人员研究的最多也最为深入，它又可以被分为多个子类别，分别是基于用户的推荐（User-Based Recommendation），基于物品的推荐（Item-Based Recommendation），基于社交网络关系的推荐（Social-Based Recommendation），基于模型的推荐（Model-based Recommendation）[6]等。

2 自适应推荐系统设计

2.1 需求分析

自适应推荐系统相对于传统推荐系统的额外需求如下：

1）自适应。针对不同的应用场景和业务需求，用户只需要定义好输入参数，格式，方法以及输出参数，格式，方法，系统就能够自动进行推荐系统模型的训练，而不需要用户介入。

2）可配置。对于同一个推荐算法模型，用户可以配置不同的输入参数来进行模型训练。

3）时效性。在线推荐中需要处理各种实时用户日志，比如用户的浏览记录，点击记录等，这些对数据的实时处理要求比较高。自适应推荐系统需要能够绑定特殊的数据源，并且能够进行实时计算。

4）智能。系统自身能够预留，动态分配，实时释放系统资源。大数据环境中的资源虽然数目非常多，但是针对不同的计算需求，系统会需要不同的计算资源。比如计算密集型任务对 CPU 数目和速度要求比较多。而数据密集型则对内存要求比较高。

因此，自适应推荐系统必须首先智能化的提供资源分配的功能，这样才能体现出自适应推荐系统的巨大潜力。

5）自治。自适应推荐系统能够对自己资源作有效的评估，选择可以接受的作业运行。一个计算机由于自身条件限制，并不能执行用户提供的所有任务，因此，每个计算机都需要有自治功能，一方面，能对自身情况作有效评估，另一方面，当系统有故障时，能够通知中央管理器，对作业进行迁移。

6）重设计而不重开发。针对开发者，能够分离算法层和程序代码层，无疑对对程序运行了解很少的科研工作者很有帮助。作为信息网格的主要客户群体—科研工作者，需要一个算法描述平台来大量，高性能的处理自己的计算需求，成为一个必然的趋势。

7）可扩展。不仅仅针对推荐系统一个层面的应用，自适应推荐系统也可以存储其他的算法，如传统的统计算法，机器学习算法等。这样使用的对象和作用域都可以扩展。

8）高吞吐。自适应推荐系统中，由于同时要进行多个推荐模型任务的数据记录以及模型计算，因此，高吞吐量是自适应推荐系统设计的核心。整个系统的设计必须围绕高吞吐，高并发为核心展开。

9）模块化。从功能上讲，自适应推荐系统已经涉及到现在最新的数据存储核处理技术，因此整个系统的结构是复杂的。模块化作为一种通过将系统切分为更小的，独立的子系统的软件设计方法。这种方法要求模块之间相互独立，通过清晰的接口定义来进行模块之间的交流。

2.2 分层设计

针对上述自适应推荐系统的需求，设计出一个可拓展的自适应系统。该系统主要用户接入层，转换层，平台层和并行计算底层构成。如下图 1 所示：

用户接入层中，终端用户通过访问 web 界面进行自己的相关信息管理，包括以下几个部分：

1）用户管理。该模块用于管理用户信息。对于管理员，拥有对所有用户的管理功能。

2）权限管理。管理使用权限。这些权限包括 CPU 数目，算法类型，作业时长，指令集要求，底层约束等。

3）数据管理。管理用户自己的打包数据。包括提交离线数据集，绑定数据接收接口，数据获取接口等方式。同时，这些数据还包括用户提交的算法描述文件并行度，

4）作业输出格式等。对于用户，系统还提供了统一的推荐系统接口来进行数据获取。

5）作业管理。管理用户的历史，当前运行，故障等作业。

对于转换层，有以下部分组成：

1）数据切分模块。该模块负责对用户提出的数据切分算法，对数据进行有效，安全的切分。很容易想到，大部分的并行执行算法都是从分割任务和数据开始。

2）算法模块。根据用户接入层里面的算法描述语言，生成可以本地执行的文件，默认的，为了保证安全和跨平台，我们代码都统一编译为.class 文件，最后，用于底层平台的 JVM 进行执行。

3）资源描述模块。该模块是对现有资源的描述分配模块，客户执行一个作业时，必须有一个作业环境描述，这样方便底层的资源进行有效的分配。

4）沙盒模块。该模块是可执行文件(.class)与平台曾的交互环境，这个环境中封装了作业处理细节，最终回显计算信息到输出文本。

5）监控模块。对于整个系统，提供系统监控模块进行统一的监控和异常恢复。

6）日志模块。日志模块是自适应推荐系统各个操作的记录器,记录详细的操作日志。

3 关键技术

3.1 模块间通信

模块间通信是实现各个模块之间有序组织，高效配合的办法。通常，各个模块之间需要通信来满足模块之间的交互需求。传统的模块间通信包括客户机服务器，分布式计算，对等系统，面向服务（SOA），微服务，消息队列等方法。在自适应推荐系统中模块间通信使用面向服务的软件架构用于模块间的信息通信。在自适应推荐系统中，使用 SOA 作为模块间通信的方法。

3.2 训练集数据绑定

自适应推荐系统中，需要绑定用户提供数据的接口。系统使用消息队列作为训练数据接收的中间件。系统中使用 JMS 标准的消息队列客户端作为数据接收工具。要求用户提供的数据必须通过消息队列发送至系统。以 ActiveMQ 为例，用户创建生产者（Producer），创建主题并向主题发送数据；客户端绑定消息队列协议，并且订阅（Subscribe）对应的主题（Topic），这样用户数据就可以被自适应推荐系统接收。

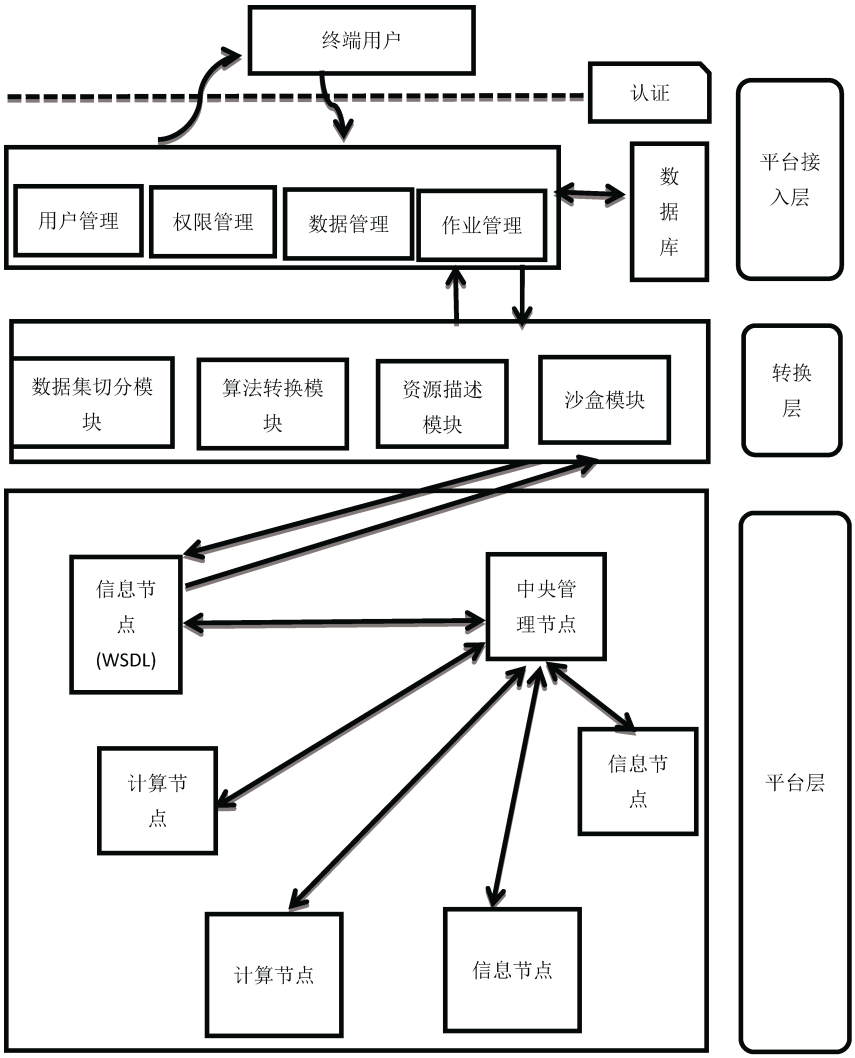

图 1 自适应推荐的系统架构

3.3 客户数据存储

客户数据存储分为关系型数据存储和非关系型数据存储。自适应推荐系统的结构化数据的存储主要包括平台接入模块的数据模型，包括用户数据模型，资源数据模型，数据管理模型，任务数据模型，资源数据模型，常见操作数据模型，算法数据模型，数据协议模型，数据获取方式数据模型等。对于结构化数据存储，自适应推荐系统采用 Mysql 作为存储方案。而对于结构不定的非结构化数据，如用户消息，日志，用户数据等，则使用非关系型数据库 MongoDB 来存储。对于用户上传的离线数据包，则使用 HDFS 进行存储。

3.4 模型训练

自适应推荐系统使用 Hadoop 作为基础分布式计算平台，以 Mahout 为推荐系统 Map-Reduce 计算引擎，对推荐系统模型进行并行训练。

4 自适应推荐系统实现

自适应推荐系统的实现可以通过以下几个功能点分别进行展示：

1）首页。首页主要展示系统的相关信息，如图 2 所示。

2）登陆／注册界面。登陆注册界面提供用户登录和注册功能，如图 3 所示。

3）系统管理界面。系统管理界面是用户成功登陆之后展现给用户的后台，如图 4 示。

4）任务管理界面。如图 5 示。

5）添加任务。如图 6 所示。

6）任务查看。如图 7 所示。

7）系统使用

通过查看任务界面，可以看到任务的进行状态，如果当前状态为已完成。则使用系统中自动生成的 restful

图 2　首页

图 3　登陆/注册界面

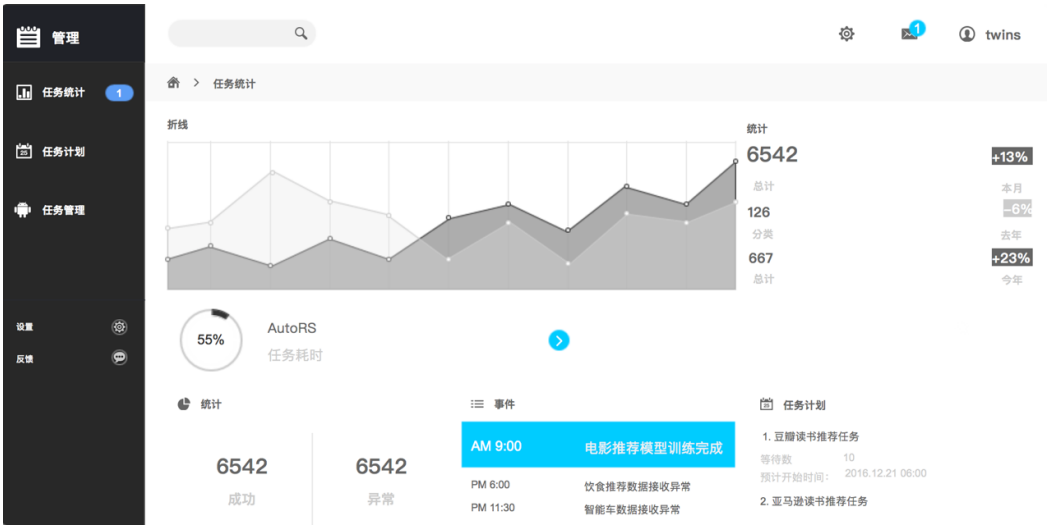

图 4　系统管理界面

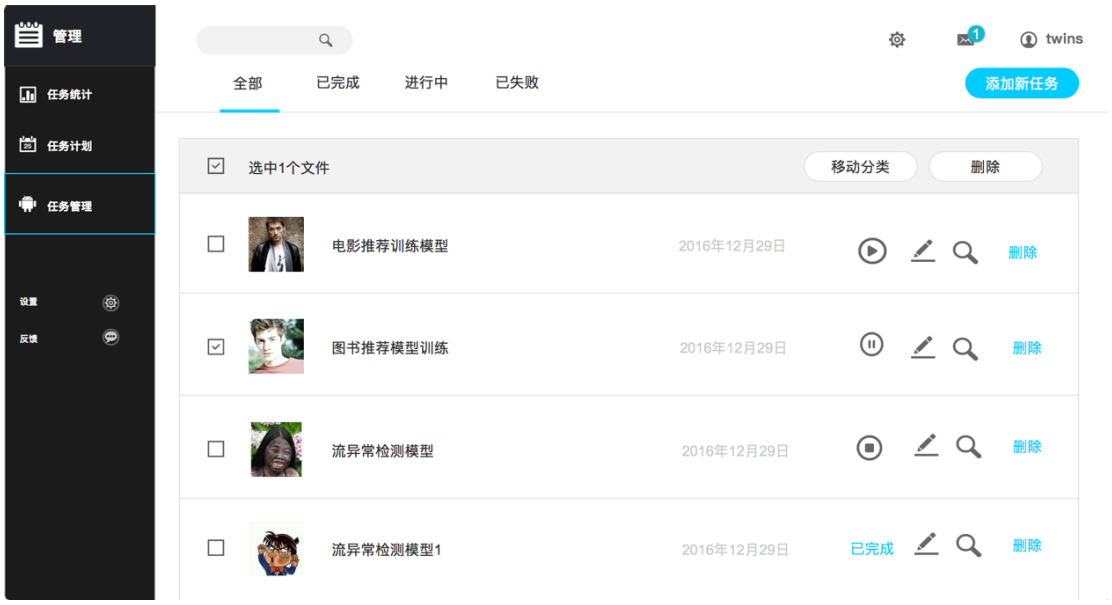

图 5　任务管理界面

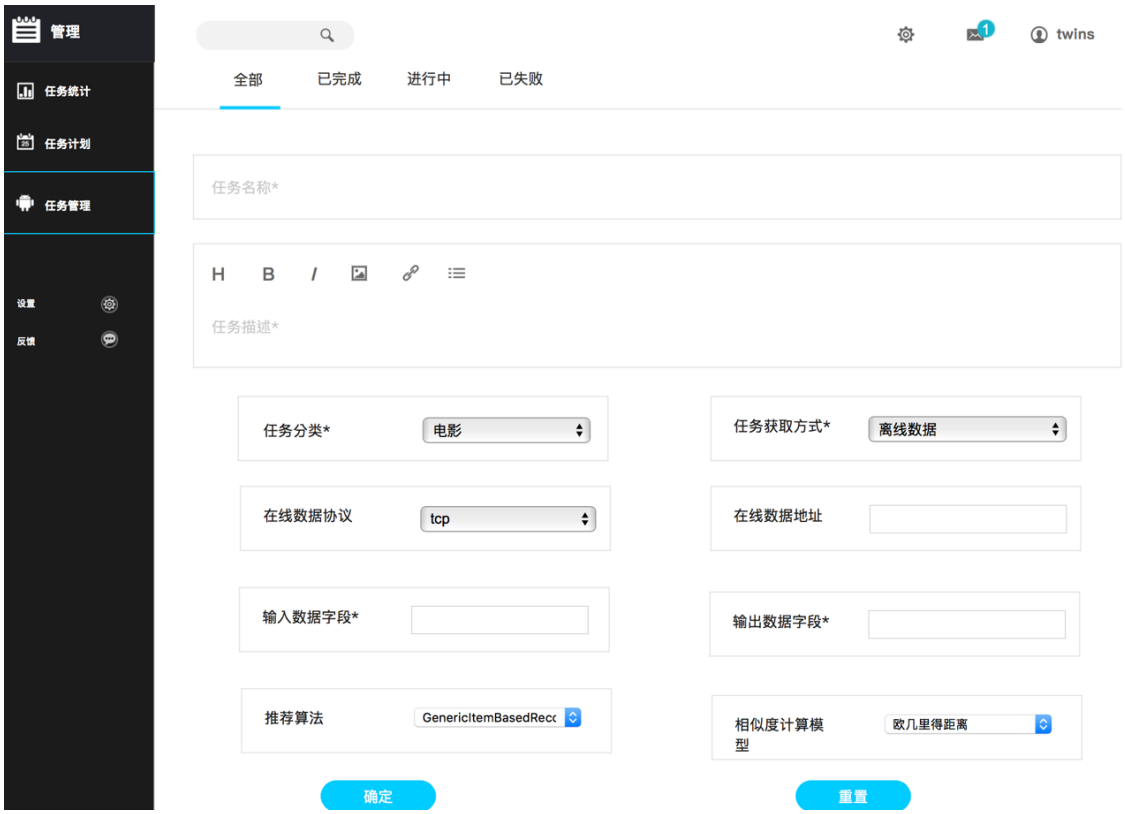

图 6　添加任务

调用接口，使用 HttpClient 进行请求获取推荐结果。

5　结束语

本文对自适应推荐系统进行了需求分析，系统设计和系统实现。本文对现有的推荐算法进行分类和概要描述，对系统进行了详细设计。文中对数据的存储，训练集数据绑定，模块间通信以及模型训练作简要描述。但是自适应推荐系统需要用户自己选择模型训练的方法和相似度计算方法，使得整个模型训练不具有完全的智能化。如何设计自适应推荐算法，如何能够通过评测集，系统自动训练出用户评价最好的推荐模型需要进一步研究。

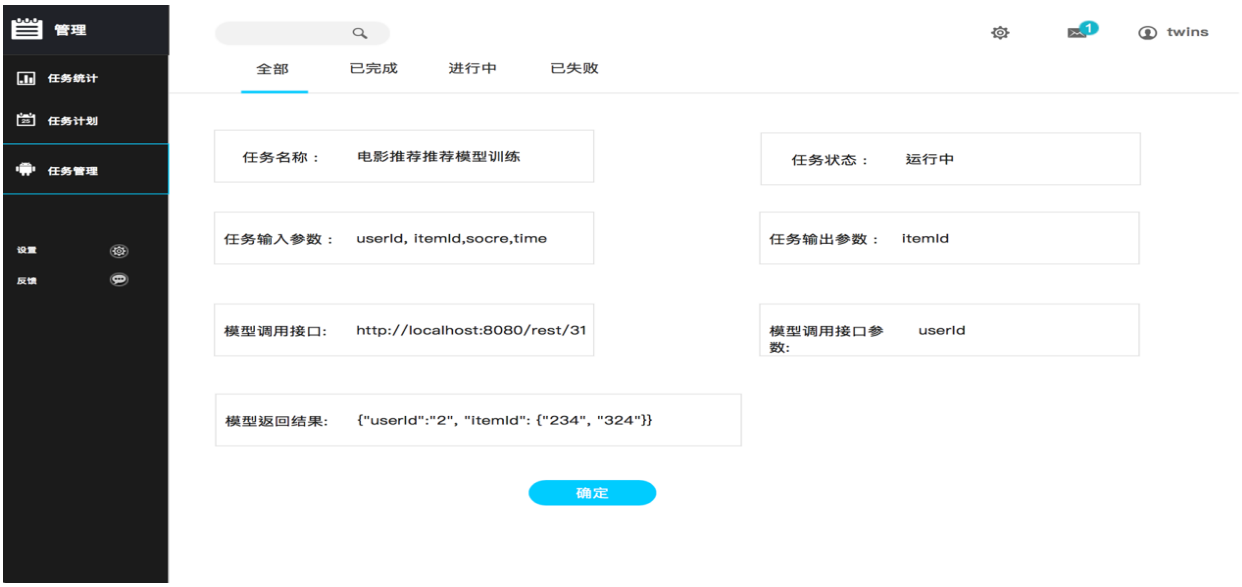

图 7　任务查看

参考文献

[1] BAWDEN D, HOLTHAM C, COURTNEY N. Perspectives on information overload[J]. Aslib Proceedings, 1999, 51(8): 249-255

[2] 李善涛，肖波．基于社交网络的信息推荐系统[J]．软件，2013, 12(5): 41-42.

[3] Ken Goldberg, Theresa Roeder, Dhruv Gupta, Chris Perkins. Eigentaste: A Constant Time Collaborative Filtering Algorithm[J]. Information Retrieval. 2001(2): 34-36.

[4] 王春才，邢晖，李英韬．推荐系统的推荐解释研究[J]．现代计算机(专业版). 2016(02): 41-42.

[5] 许海玲，吴潇，李晓东，阎保平．互联网推荐系统比较研究[J]．软件学报. 2009(02): 351-353.

[6] 陈雅茜．系统及相关技术研究[J]．计算机工程与应用. 2012(18): 9-12.