****

**《信息检索》课程大作业**

题目Title： Recommender Systems

院 系

School (Department)：数据科学与计算机学院

专 业

Major： 软件工程

学生姓名

Student Name：陈明亮 陈铭涛 陈慕远

学 号

Student No.：16340023 16340024 16340025

时 间

Submitted Time.：2020.1.9

中山大学 • 中国广州

**【摘 要】**

推荐系统，一个已历经近二十年发展的概念，其定义可以涵盖多个方面。一般来说，推荐系统可以理解为是主动向用户推荐信息、物品的系统，其推荐的事物也可以是多方面的，如音乐、书籍、活动、股票、新闻等等，依赖于它所服务的具体应用领域。简单来说，推荐系统就是根据内部的算法，得出可能对用户有帮助或是用户可能感兴趣的物品或信息。近年来随着电子商务规模的不断扩大，商品的种类与数量不断增长。这样的现状对于应用的检索和推荐服务提出了更高的要求。由于不同用户在个人经历、兴趣爱好等方面的不同，他们的购买倾向必然也有所不同。为了满足不同用户的不同推荐需求，实现对每个人定向推荐的个性化推荐系统，对推荐系统进行详尽深入的研究、改进是至关重要的。

在本文中，我们将对推荐系统的发展作各个方面的探究。首先我们将介绍推荐系统研究中常用的数据集和评判指标，以说明推荐系统所基于的数据基础及其功能优劣的判定。接着我们将探索那些较为传统的推荐算法，厘清它们的工作原理，并详述它们的优势与不足。然后我们会将目光投向现在，对当下热门的基于神经网络的深度推荐算法做介绍，包括其架构、工作原理及功能评估。最后我们会以那些著名的工程——诸如推荐系统在Youtube、Netflix、Amazon、Bytedance中的应用——为示例，阐述推荐系统在实际工程中的应用与挑战。总体来说，我们主要对推荐系统的历史进行细致的综述，对推荐系统的现在进行深入而广泛的学习。

**【关键词】：**推荐系统；协同过滤；神经网络；深度学习；

[ABSTRACT]

Over more than two decades of growth, recommender systems have been proven useful in various fields. By its definition, recommender systems is a system that, with its underlying algorithm, actively push new items, such as music, books, stock or news, to its users. With the fast growth of e-commerce in recent years, the numbers of active users and catalog items have been skyrocketing. Such status is posing a huge challenge for information retrieval and recommendation. There are a lot of differences in the preference of different users due to their diverse personal experience and interests, to fulfill which will require a more personalized recommender system, and more in-depth researches into this topic are much needed.

In this work, we will look into various aspects of the past and current research of recommender systems. Specifically, we will describe how are different recommendation methods evaluated. Then we will introduce some of the traditional recommendation methods that build up the basis of our current research. We will then discuss the principle, architecture, and evaluation of the current state-of-the-art recommendation algorithms that are based on deep learning methods. Lastly, we will look at how recommendation systems are practiced in some of the renowned companies such as YouTube, Netflix, Amazon, and ByteDance. Overall, our survey will provide a detailed overview of the history of recommender systems and an in-depth description of the current research.

**[Keywords]:** recommender system; collaborative filtering; neural network; deep learning;

目 录

第一章 概述/引言 5

1.1 研究推荐系统的背景和意义 5

1.2 问题的描述 5

1.3 本文的工作 6

1.4 论文结构简介 6

第二章 推荐系统研究的常用数据集与评判指标 7

2.1 常用数据集 7

2.2 评判指标 8

**2.2.1** **评分预测准确度（Ratings Prediction Accuracy）** 8

**2.2.2** **使用情况预测（Usage Prediction）** 9

**2.2.1** **排名评估（Ranking Measures）** 10

第三章 传统的推荐算法 12

3.1 基于人口统计学的推荐 12

3.2 基于内容的推荐 12

3.3 基于协同过滤的推荐 13

**3.3.1** **基于用户的推荐** 13

**3.3.2** **基于项目的推荐** 14

**3.3.3** **基于模型的推荐** 15

第四章 基于神经网络的深度推荐算法 16

4.1 深度社交网络推荐系统 16

4.2 基于神经网络的会话推荐系统性能评价 18

4.3 基于神经图的协同过滤算法 19

第五章 推荐系统在工程中的实践 22

5.1 Youtube 22

5.2 Netflix 24

5.3 Amazon 25

5.4 字节跳动（ByteDance） 26

第六章 总结与展望 27

参考文献 28

# 概述/引言

## 研究推荐系统的背景和意义

推荐系统，一个已历经近二十年发展的概念，其定义可以涵盖多个方面。一般来说，推荐系统可以理解为是主动向用户推荐信息、物品的系统，其推荐的事物也可以是多方面的，如音乐、书籍、活动、股票、新闻等等，依赖于它所服务的具体应用领域。推荐系统就是根据内部的算法，得出可能对用户有帮助或是用户可能感兴趣的物品或信息。近年来随着电子商务规模的不断扩大，商品的种类与数量不断增长。这样的现状对于应用的检索和推荐服务提出了更高的要求。由于不同用户在个人经历、兴趣爱好等方面的不同，他们的购买倾向必然也有所不同。为了满足不同用户的不同推荐需求，实现对每个人定向推荐的个性化推荐系统，对推荐系统进行详尽深入的研究、改进是至关重要的。

## 问题的描述

推荐系统在本质来讲是一种信息过滤系统，着重于预测用户对物品项目的评分或偏好，最终把在用户和物品之间的关联程度显式表示出来。世间万事万物都有连接，人与人之间有社交连接，人与商品之前有消费连接，人和资讯之间有阅读连接….而推荐系统主要的工作，就在于利用已经存在连接信息，去预测未来可能存在的连接，不同的推荐系统架构的工作重心都在于此，但随着技术和算法的进步，系统内部所依赖的算法正产生着日新月异的变化。

传统的推荐算法包含基于人口统计学，基于内容和基于协同过滤算法的模型，其中以协同过滤算法CF为代表，曾是近现代使用最广泛的推荐算法。但随着深度学习方法和神经网络的发展，科研人员们认识到深度学习领域的技术可以应用于推荐领域，并且可以把握和利用经典算法没有考虑过的关联信息和特征提取方法，使得推荐系统的整体性能表现更加优越，更加适用于当前蓬勃发展的互联网领域。

## 本文的工作

本文将对推荐系统的各个方面进行介绍。首先对推荐系统研究的常用数据集和评判指标进行了解，接着探究传统的推荐算法的特征，并深入介绍当今基于神经网络的深度推荐算法，最后对实际工程中推荐系统的应用进行探究。

## 论文结构简介

本文第一章主要是阐述研究推荐系统算法的背景和意义以及本文所作大致工作内容；第二章主要介绍说明推荐系统研究的常用数据集和评判指标；第三章将会介绍在推荐系统的发展历程中诞生的那些传统而经典的推荐算法；第四章则主要论述前沿的基于神经网络的深度推荐算法；第五章介绍了在真实的工程环境中各类推荐系统的实践；第六章是对本文的总结以及对接下来研究方向的展望；最后是参考文献的列表。

# 推荐系统研究的常用数据集与评判指标

## 常用数据集

在推荐系统的学术研究中，往往难以使用真实场景的实时数据，因此会通过使用接近真实场景的数据集来对模型与算法进行评估，数据集中一般包括如下部分：

* **Item：**被推荐的对象，如商品，影片，网页等
* **User：**对推荐item 进行打分和接受新 item 推荐的主体
* **Rating：**user对 item 的 偏好程度，可以是二元（喜欢/不喜欢），离散（1-5星）或连续（一定范围内的分数数值）的。

一下将介绍一些研究中常用的数据集：

1. **MovieLens[[1]](#footnote-2)**

MovieLens 数据集由University of Minnesota收集，完整数据集包含对58000个电影的27000000个 rating，也提供了较小的10M，25M 等数据集。数据集中除了上述的三个部分外还包含了电影的标签.

1. **Jester[[2]](#footnote-3)**

由 UC Berkeley收集，3个数据集共包含了对150个笑话的6500000个 rating。

1. **Yelp Dataset[[3]](#footnote-4)**

由 Yelp Challenge提供的数据集，包含了餐厅评论数据。在结合文本评论的推荐系统研究中较常使用。

1. **Last.fm[[4]](#footnote-5)**

在 HetRec 2011上发布的包含了1892个用户对92800个艺术家的收听记录的数据集。

1. **Netflix[[5]](#footnote-6)**

来自于网飞（Netflix） 公司的用户评分预测比赛的数据集，第二版中共包含了来自17770部电影，480189个用户的数据。

## 评判指标

针对推荐系统不同的目标场景，有着不同的评判方式。根据实验设置不同主要可以分为在线（Online） 和离线（Offline） 评估。

在线评估中，常用的方法是 A/B测试（Kohavi等人（2009）[1]），将一部分的请求导向到控制变量的不同推荐系统，并记录用户交互的情况，获得的大量数据可以使用数据挖掘的方法来对用户界面、用户意图、用户行为等进行分析。A/B 测试对于流量资源有着一定的需求，而且一般会产生大量的数据，流媒体公司网飞（Netflix）首先使用了 Interleaving 方法（Chapelle 等人（2012）[2]）来对需要测试的推荐算法进行快速筛选，然后再对筛选出的推荐算法进行 A/B 测试，以快速测试大量的算法，提升学习效率（Aurisset 等人（2016）[3]）。A/B测试也存在着无法在展示给用户前进行测试的问题，因此还需要合适的离线评估进行补充。

离线评估的目的是较低成本地对算法效果进行测试，常用的指标主要分为准确性（Accuracy） 和可用性（Usefulness）两种，以下根据预测目标分类进行介绍（Gunawardana等人（2011）[8]）。

### **评分预测准确度（Ratings Prediction Accuracy）**

评分预测准确度主要用于预测用户评分的场景中，较多使用的指标包括根均方差（Root Mean Squared Error） 和平均绝对误差（Mean Average Error），其方法是对测试集 中的每一个 User-item 对生成一个预测评分，并与真实值计算出评估指标值。

NRMSE （Normalized RMSE）和NMAE（Normalized MAE） 则是在RMSE 和 MAE 的基础上，根据范围对评分进行正规化（）。

Average RMSE 和 Average MAE则可对分布不均的测试集进行调整，对每个 item 或每个 user单独计算 RMSE 或 MAE 后取平均值，可以避免不均衡的item 或user数据对评估结果造成的影响。

### **使用情况预测（Usage Prediction）**

使用情况预测对用户可能选择的 item 进行预测，比如网飞或 YouTube根据用户观看和收藏列表向用户推荐用户可能感兴趣的影片。在离线评估中，数据集一般包括了每个用户曾经使用过的 item，对于每个用户，测试时隐藏一部分选择，并让推荐系统预测用户可能会选择的 item，并列出如下的confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | **Recommended** | **Not Recommended** |
| **Used** | **True-Positive(tp)** | **False-Negative(fn)** |
| **Not Used** | **False-Positive(fp)** | **True-Negative(tn)** |

根据该 confusion matrix 可以计算如下的常用指标：

该种评估方法存在的问题是其假设了用户未选择 item 被推荐会必定不会被选择，但这种假设可能是不正确的，因此计算的False-Positive可能被过高地估计。随着推荐项目数量的增加，一般存在着 Recall上升，Precision 下降的趋势。在推荐项目数量预定的场景中，最常用的评估指标为**Precision At N**，即推荐的前 N 项的 Precision 值。在推荐数量不确定的场景中，则一般对会对一定范围内不同的推荐数量进行评估，计算不同的 Precision 和 Recall 等指标，即绘制Precision-Recall 曲线或 ROC 曲线，然后可以使用 F-measure 或 Area Under Curve(AUC)等方法对曲线进行分析。

### **排名评估（Ranking Measures）**

在一些应用，比如网飞的推荐影片列表中，向用户推荐的项目较多，需要根据用户偏好对推荐结果进行排序。对于排名效果的评估有两种常用的方法：

第一种方法是获取推荐列表的正确排名后将推荐系统的排名与之对比，正确的排名可以来源于用户对项目的评分（如影片评分），或用户对项目的使用情况（如音乐软件中将用户完整听完的歌排于用户跳过的歌之前）。以上两种方式的问题在于可能存在并列的排名，可以使用**Normalized Distance-based Performance Measure(NDPM)**方法（Yao（1995）[9]）来解决问题。假设对用户的个项目有真实排名和系统排名，则有：

对于结果与正确排名相同的系统，**NDPM**会给出0的分数，而对结果完全不正确的系统，则会给出1的分数。

另一种方法是评估系统的排名对于用户的可用性。常用的方法是考虑推荐列表的“上瘾”程度，计算排名可用性的方法是排名中所有项目根据对应排名削减后的可用性。在一些系统中用户只会关注前几位的推荐项目，此时可以给排名后的项目赋予较大的可用性削减，可用的指标有 **R-Score** （Breese等人（1998）[10]），对于每个用户有：

其中是排在第位的项目，是用户对项目的评分，是用户对无看法的项目的评分，控制了列表可用性的指数削减程度。

对于用户可能关注数量较多的推荐项目的场景中，排名的可用性削减则较小，常用的方法是**Normalized Cumulative Discounted Gain(NCDG)**（Javelin 等人（2002）[11]），假设用户对被推荐的项目有，则对于被推荐的 个项目有：

使用理想 CDG 正规化后可得 NCDG：

# 传统的推荐算法

在本章中，我们将介绍一些较为传统又极为典型的推荐算法，内容包括他们的方法详述以及优缺点分析。

## 基于人口统计学的推荐

基于人口统计学的推荐（Demographic-Based Recommendation）（Pazzani 等人（1999）[5]）是一种易于实现的推荐方法，现如今这种方法已经很少被单独使用，但是研究它对理解推荐系统的工作原理有很大的帮助。

简而言之，这种方法就是根据系统用户的基本信息找出用户的相关程度，再将相似用户的喜爱推荐给当前用户。它首先需要进行用户信息标签化（User Profiling），其中收集记录了消费者性别、年龄、活跃时间等主要信息（基本信息），当进行个性化推荐时，就根据User Profile计算出用户的相似度并进行排序，取最相似的前K个用户，再根据这些用户的购买、浏览、评价信息进行推荐。

优点：

* User Profile相对固定，相似度计算简单，故推荐可以实现实时响应。
* 不依赖物品本身的数据，在任一领域都可用。

缺点：

* 可信度低，基本属性相同的用户可能有着截然不同的偏好，且这种方法没有使用当前用户对物品的喜好的历史数据。

## 基于内容的推荐

基于内容的推荐（Content-Based Recommendation）（Gunawardana 等人（2009）[4]）是在推荐引擎出现之初应用最为广泛的一种推荐机制，简单来说，它是根据这样一个假设：“一个用户可能会喜欢他曾喜欢过的物品以及相似的一类物品”。

刚才提到在基于人口统计学的推荐中，使用到User Profile来进行推荐，在基于内容的推荐中，也用到了一种不同的User Profile。在这里，User Profile需要统计该名用户曾经的评价信息，一种方法是对用户曾经评价过的物品的Item Profile做一个加权平均（也可以添加时间因素进行加权，考虑用户偏好的历史变化）来作为User Profile。之后根据User Profile进行推荐，可以是计算其他未使用过的物品与User Profile间的相似度，也可以使用更复杂的算法。

优点:

- 相比基于人口统计学的推荐，这种推荐结果更加精确、可信，因为参考了用户自身的偏好历史。

缺点：

- 为了得到物品的特征需要进行预处理（分析、建模），推荐的质量依赖于对物品建模的完整和全面程度。

- 可扩展度不佳，在新领域进行推荐时需要重新进行建模。

- 无法挖掘那些用户没用接触过但存在潜在喜好可能的物品进行推荐。

### **三.3基于协同过滤的推荐**

基于协同过滤的推荐（Collaborative Filtering-Based Recommendation）是指收集用户过去的行为（对物品或信息的偏好），发现物品或者内容本身的相关性，再基于这些相关性进行推荐。它主要可以分为三种：基于用户的推荐（User-Based Recommendation）（Resnick 等人（1994）[6]）、基于项目的推荐（Item-Based Recommendation）（Sarwar 等人（2001）[7]）、基于模型的推荐（Model-Based Recommendation）。

### **三.3.1 基于用户的推荐**

基于用户的推荐（User-Based Recommendation）是很早就出现的一种基于协同过滤的推荐算法，它和基于人口统计学的推荐实际上非常类似，其原理是根据所有用户对物品或信息的偏好，找出可能有相似偏好的其余用户。不同的是，这里的相似用户并非使用User Profile（用户本身信息）计算得出，而是根据用户的评价信息或者说历史偏好得出。

User-Based方法中，常使用两种推荐方法：Top-K推荐与关联推荐。Top-K推荐是在前K个用户中进行统计，将那些出现频率最高且未在当前用户的列表中出现的物品作为输出对用户进行推荐。关联推荐是根据前K个用户的购买、评价行为挖掘关联规则，再结合关联规则与用户的购买记录进行推荐。

优点：

* 准确率较高，数据集丰富的条件下，能够根据用户行为发现物品间的隐式关联性。

缺点：

- 随着用户数量增加，计算时间会显著增长，限制了系统的可扩展性。

- 加入新用户时，由于其历史记录较少，难以准确计算相关性。

### **三.3.2 基于项目的推荐**

基于项目的推荐（Item-Based Recommendation）是Amazon的专利算法，其基本假设与基于内容的推荐类似，其原理是根据所有用户对物品或信息的偏好，找出物品与物品之间的相似度，然后根据用户的历史偏好信息，将类似的物品推荐给用户。与基于内容的推荐不同的是，这里通过物品被打分的历史进行计算，而非前面提到的Item Profile。

在推荐方法上，Item-Based也需要执行最近邻搜索（与User-Based相同），得出Item之间的两两相似度，再进行推荐。

优点：

- 计算简单，易于实时响应，可采用离线完成、定期更新的方式（Item的打分通常不会过于剧烈变化）。

缺点：

- 更多关注物品，对于用户之间的差别考虑较少，精度稍差于User-Based方法。

推荐策略的选择与具体的应用场景有很大关系，在大部分的web站点中，物品的数量都比用户的数量要少很多，因此物品的相关度较为稳定，此时Item-Based就比User-Based在实时性方面更好些。但在其他一些物品数量更多的系统里，如更新很快的新闻推荐系统中，物品的相关度不稳定，使用User-Based的效果可能会更好。

### **三.3.3 基于模型的推荐**

User-Based与Item-Based都存在一个缺点：在大数据量下表现不够好，难以处理大量即时结果。因此发展出了基于模型的推荐（Model-Based Recommendation），其原理是根据样本的历史偏好信息训练一个推荐模型，再根据实时的用户喜好信息进行预测、 推荐。

优点：

* 快速响应，得到模型后计算非常快速。
* 领域无关，模型训练得当可以很好的发现用户的潜在偏好。

存在问题：

* 推荐的效果依赖于用户历史偏好数据的多少与准确性，还需要将近期新

增的偏好数据反馈给训练好的模型，实现增量训练。

# 基于神经网络的深度推荐算法

本章主要介绍与深度神经网络相关技术结合的推荐算法，以及基于此类算法所构建的各种可应用的推荐系统。

## 深度社交网络推荐系统

深度社交网络推荐系统（Deep Social Network Recommendation System）基于深度社交协同过滤算法（Deep Social Collaborative Filtering）（Wenqi Fan等人（2019）[12]），该算法基于原始的传统协同过滤思想，结合深度神经网络，构建社交推荐系统，不仅可应用于用户项目互动（User-Item Iteration），还能提供由社会理论和社交关系链所提供的不同用户的偏好信息（User-Individual Information）。

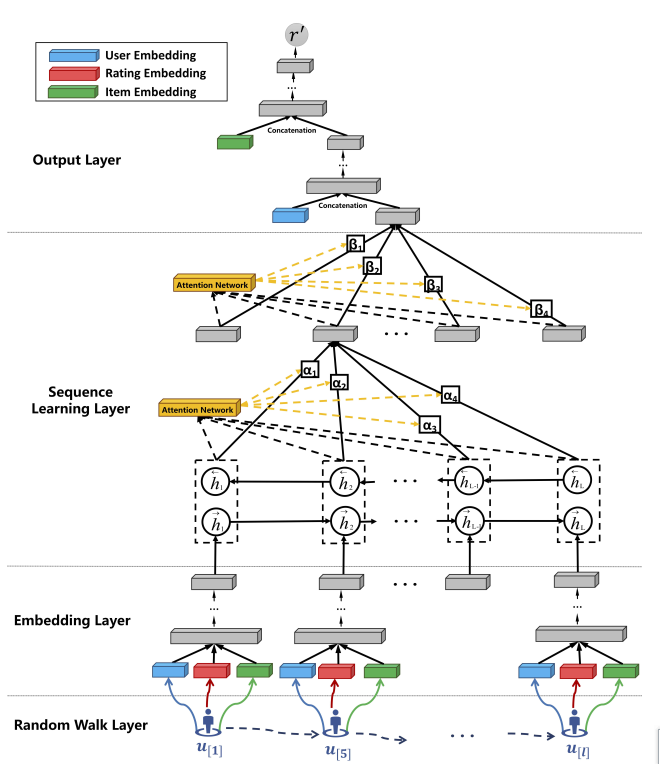
但是，当前非神经网络构建的社交网络模型往往不能充分利用社交网络信息，大多数传统架构只使用了近邻的关系信息，而忽略了距离过远的邻居之间的联系。同时，大多数模型都不存在对信息权重的分配，而是平等地对待所有近邻的信息。此外，特定情况下的推荐案例所需的额外信息也是社交网络系统的重点部分，模型也需要充分考虑初始信息，以及邻居与邻居之间的意见捕获等内容。

为了解决上述挑战，该深度社交网络推荐系统采用深度社会协作筛选框架，简称DSCF（Deep Social Collaborative Filtering framework），基于深度神经网络，利用不同方面的社会关系信息，构建比传统社交网络系统更加有效可行的深度社交网络推荐系统。

首先介绍的是该深度社交网络所基于的主题神经网络构建部分。整体的思想有：

1. 不仅考虑来自直接邻居的信息，还考虑来自遥远邻居的信息
2. 选择每个邻居的相关信息以推荐特定项目
3. 在对用户与物品的互动进行建模时，捕捉邻居对于物品的意见

如图1，DSCF框架由四层神经层组成，其中随机游走层（Random Walk Layer）用于解决远邻信息的获取，嵌入层（Embedding Layer）、序列学习层（Sequence Learning Layer）用于解决用户项目推荐问题，同时记录和学习用户意见信息。输出层（Output Layer）获取用户推荐结果。



**图1. DSCF神经网络框架架构图[12]**

其中，随机游走层采用了计算机网络中探索本地邻居的经典算法，即基于距离和最大跳数的随机探索算法 – 设定每次随机探索的最大跳数，并且总是往未探索过的距离最远的邻居方向随机进行，如此往返多次便可以获取包含近邻与远邻的全面信息。

此外，嵌入层通过对用户项目交互序列（User-Item Iteration）的处理，首先建立用户对商品评分模型，以此区分好坏意见和程度分级，此外通过评级嵌入向量（Rating Embedding Vector）对用户商品评分进行离散化，但考虑到用户与商品交互信息的高度非线性，作者采用多层感知机MLP将交互信息与评级信息融合在一起。

序列学习层旨在提取每个序列的特征，组合所有序列的多维度特征以获取对应的预测结果，为了捕获模型中用户与用户之间的双向影响，作者提出一种基于双向长短期记忆网络（Bi-LSTM）的语言模型，来捕获NLP域中的单词之间的远程双向语义依赖程度，由此充分利用不同邻居之间的关系信息，构建出全方面的推荐系统。

最终，作者基于本深度社交网络模型，与传统的协同过滤算法进行基于Ciao数据集和Epinions数据集的性能测试，分别采用RMSE与MAE作为评分标准，得出结论如下：

1. DSCF架构在所有数据集，所有评分标准上的综合得分均高于其余参与测试的传统推荐算法
2. DSCF架构存在改进之处，可以通过调优参数，模型融合等方法进行改进，以及提高性能。

## 基于神经网络的会话推荐系统性能评价

由于当前基于会话推荐系统大多数都采用传统的推荐算法，其中大部分以协同过滤为核心，同时由于深度学习领域的快速发展，神经网络的应用更加广泛。

论文Performance Comparison of Neural and Non-Neural Approaches to Session-based Recommendation（Malte Ludewig等人（2019）[13]）将重点放在基于神经网络所构建的会话推荐系统上，并使用众多数据集对传统会话推荐系统与其进行各方面的性能对照比较实验，从各方面标准进行评估。

结合四种常用的神经网络会话推荐系统架构策略：GRU4REC、NARM、STAMP、NEXTITNET（注：这些算法策略都是RecSys会议中引用数排名靠前的常用策略），同时结合多种基本的传统会话推荐系统算法，如：AR、SR、S-KNN、VS-KNN等作为评价上述神经网络方法的标准参考基线方法，从评分对比中获得对照比较。

通过采用七种大规模数据集，包含电子商务、音乐收听等领域，同时对照实验的步骤为：

1. 超参数优化，多次迭代查询较优参数
2. 协议和指标制定，采用命中率HR、平均倒数秩MRR、平均精度MAP等指标对不同方法进行评分

最终结果显示：

1.传统推荐算法中最近邻算法KNN的跨折叠平均精度值最高，但四种神经网络架构方法的评分都比最佳基线KNN要高，并且是在不同指标上持续地优于KNN。

2. 在四种神经网络方法中，STAMP方法伸缩性要差于其他方法，表现为偶尔有较低的指标得分，以及在不同数据集上的表现差距较大，即可适用性有待提升。

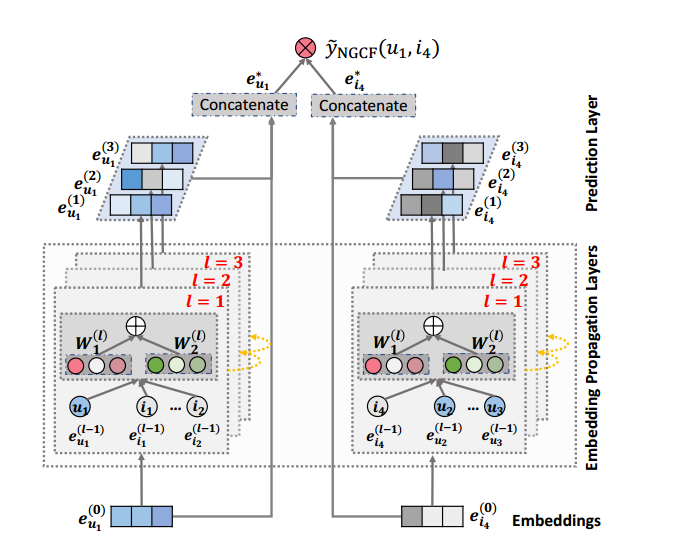
3. 总体而言，在传统推荐算法中，简单的SR方法能在小型数据集中获得较好的平均精度，而KNN在精度、查全率、命中率等方面都是传统算法中最优的，是最佳基线算法。在所评估的四种神经网络方法中，GRU4REC和NARM是性能指标评分位于前列的两种方法，同时神经网络方法对于提升推荐范围覆盖率，解决简单传统方法总是推荐更受大众欢迎的商品的偏差问题，具有十分优良的性质。

4. 除此之外，作者认为对神经网络的评判标准还要基于额外的指标，比如：可扩展性、超参数优化时间、模型训练时间（反映模型复杂度）等。

## 基于神经图的协同过滤算法

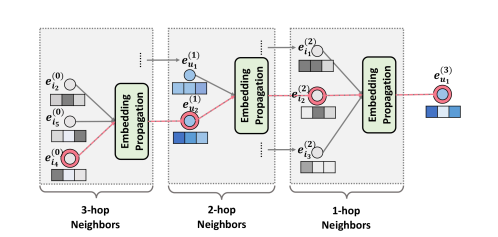
学习用户和对应项目的矢量表示的过程，又称为嵌入（Embedding），是现代推荐系统的核心。从早期的矩阵分解，到协同过滤方法，再到最近才开始流行的基于深度神经网络的方法，推荐系统的架构正逐渐发展。在原本的固有观念中，我们通常通过映射表述用户项目交互行为（User-Item Iteration）结果的不同维度特征（例如ID和属性）成为不同的离散化编码，以获得矢量表示。但在Neural Graph Collaborative Filtering一文中（Xiang Wang等人（2019）[14]），作者提出传统方法的固有缺点：在用户与项目的互动中，潜伏的协作信号未在嵌入过程中进行编码，因此所得到的矢量表示可能不足以捕获协同飘动效果（Collaborative Fltering Effect）。于是，作者团队提出将用户项目交互结果，以及该过程的全体交互信号，组成二部图结构，集成到嵌入过程中，称为新的基于神经图框架的协同过滤方法（NGCF – Nerual Graph Collborative Filtering）

NGCF框架中包含三个组件：一个嵌入层（Embedding Layer） – 负责用户嵌入和项目嵌入的初始化，多个嵌入传播层（Embedding Propagation Layers） - 通过注入高阶连通性关系对嵌入矢量进行引用，预测层（Prediction Layer） - 汇总来自不同传播层的参考嵌入，并输出用户项目对的关联性得分。



**图2. NGCF模型架构图[14]**

其中，框架的核心点在于多层嵌入传播层，是NGCF区别于传统协同过滤方法的关键，整体三阶嵌入传播的架构如下：

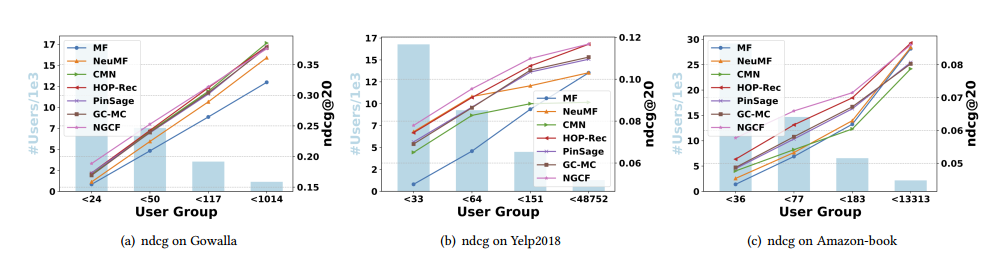


**图3. 三阶嵌入传播图示 [14]**

该架构可分为初阶传播与高阶传播两个宏过程，首先从直观上讲，交互项目提供了用户偏好的直接依据，所以我们可以反过来说，消费某个项目的用户可以称为该项目的属性，可用于区别项目之间的协作相似性。以此为基础，在连接的用户和项目之间执行嵌入传播，其中两个关键宏过程完成的功能如下：

1. 初阶传播（First-Order Propogation），显式地利用第一级连通信息，初始化并关联用户和项目表示。
2. 高阶传播（Higher-Order Propogation），通过消息构造和消息聚合（Message Construction And Aggregation），结合低阶连通性信息，堆叠更多嵌入传播层探索高阶连通性信息，以此能获得传统算法无法获取的用户项目相关性协作信号。

最终的评估实验中，采用召回率（Recall Rate）与归一化折损累计增益（NDCG – Normalized Discounted Cumulative Gain）作为评价指标，同时结合的基准线方法为：MF、NeuMF、CMN、Hop-Rec、PinSage、GC-MC等方法，结果表明NGCF方法无论是在召回率还是增益率上都超过基准方法，并且在规模不同的数据集上都持续有着最好的性能表现。



**图4. 不同数据集上的NDCG指标得分比较 [14]**

# 推荐系统在工程中的实践

本章中将介绍在真实的工程环境中各类推荐系统的实践工作。

## YouTube

Google公司在2016年的论文Deep Neural Network for YouTube Recommendation（Covinton 等人（2016）[15]）介绍了Youtube使用神经网络在推荐系统中的实践。文章中提到了YouTube 所面临的主要三大挑战：

* **大规模（Scale）**：当时存在的许多推荐算法仅能在较小的数据集上取得较好的效果，而难以在 YouTube 这样的规模下良好运行。因此需要提出更好的分布式学习算法和高效的服务系统。
* **新鲜度（Freshness）**：YouTube 每秒钟都有大量的新影片被上传，因此推荐系统需要具备能快速处理新内容和新用户操作的能力。
* **噪声（Noise）**：收集到的用户历史行为数据往往是稀疏且不完整的，另外影片的内容也难以结构化，因此对于训练数据的这些特点算法都需要有足够的鲁棒性。

论文中给出的系统整体架构如图5，包含了两个神经网络，分别用于生成推荐的 candidate 和排名。Candidate 生成的网络输入为用户的活动历史记录，基于协同过滤，输出为从一个较大的视频语料库中挑选的包含数百个视频的子集。排序网络的目标则是在candidate 列表中选出相对重要性较高的推荐项目，并提高召回率，实现的方式是根据视频和用户的特征对每一个视频赋予一个评分，评分最高的数个视频最终呈现给用户。

对于Candidate 生成网络，模型架构如图6，使用三层网络结构，输入为用户观看记录和搜索记录的 Embedding向量、地理信息、用户基本信息等特征。

对于排序网络，模型架构如图7，相对于 Candidate 生成网络，结构相似，但是使用了更加精细的特征来分析视频（item） 与用户-视频（User-item）的关系。

图片包含 文字

描述已自动生成

**图5. 推荐系统架构 [15]**

图片包含 文字, 地图

描述已自动生成

**图6. Candidate Generation架构 [15]**

图片包含 文字, 地图

描述已自动生成

**图7. Ranking架构 [15]**

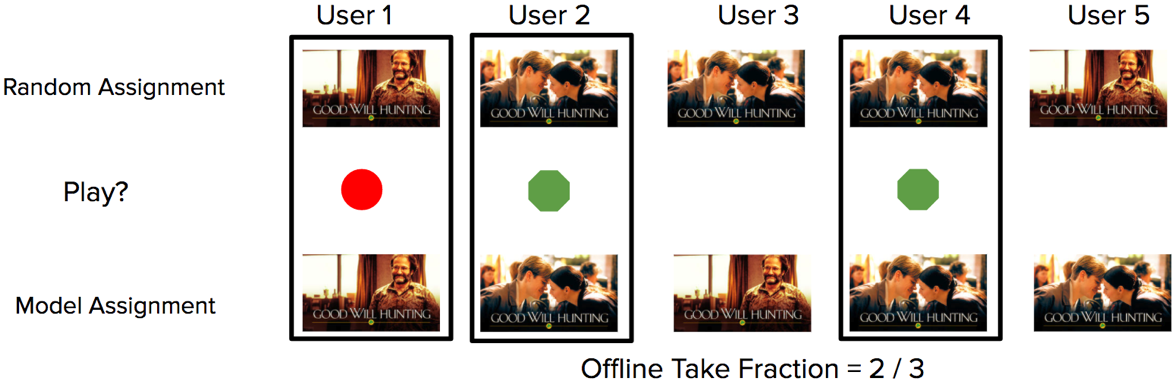
## Netflix

Netflix Research[[6]](#footnote-7) 页面中介绍了Netflix对推荐系统的应用。比如如何应用推荐算法向用户提供定制化的作品缩略图推荐（Chandrashekar等人（2017）[16]），目的是解决如下挑战：

* 同一缩略图：对于每个作品同一时间都只能使用一张缩略图进行描述，这使得收集用户对缩略图偏好的数据较为困难，因为不能了解用户点击作品时缩略图对用户行为的影响情况。
* 图片改变对推荐的影响：在两次对用户的推荐中，若对同一作品使用了不同图片来表示，可能会导致作品的可辨识度下降。
* 同一页面上的不同作品的影响：若同一页面上风格类似的缩略图较多，缩略图的特点可能会被用户忽略
* 工程上的挑战：Netflix 的服务峰值可能达到每秒2000万个请求，而该服务涉及到非常多数据量较大的图片，需要在较短时延内提供服务就要求系统具有较高的鲁棒性。

相较于传统开发流程的收集训练数据-训练机器学习模型-A/B测试的方法，Netflix 使用了Contextual Bandits的方法来避免传统流程可能造成的一部分测试用户在一段时间内无法获得更好的用户体验。这种方法会在流程中不断寻找对每一个用户的最佳推荐方式。收集的训练数据包括对每一个（用户，作品，缩略图）元组，用户是否发起了点击事件；用户行为数据包括用户播放过的作品，作品的体裁，用户所在国，用户语言偏好，设备，时间等。模型输出为用户对该作品缩略图的点击概率，系统根据该概率对图片进行排序并呈现给用户。

对模型的离线评估使用的是 **Replay** 方法（Li 等人（2011）[17]），基本思路是离线比较在历史请求下不同的算法会产生的结果。如图16，若模型的输出与用户操作时随机选择的缩略图一样，则可以得知假设使用该算法时用户点击作品的比例。



**图8. Replay 方法评估模型 [16]**

经过离线评估后的较优秀的模型最后会进行一次 A/B 测试，经过测试也发现离线测试的结果与在线测试结果呈现相关性。

## Amazon

Amazon是一家广泛应用推荐系统的科技企业，在论文Amazon.com recommendations: Item-to-item collaborative filtering（Linden 等人（2003）[18]）中介绍了亚马逊使用的 item-to-item协同过滤方法。当时算法所面临的挑战有：

* 大型的零售商可能有千万级的顾客和百万级的商品条目，数据规模极大。
* 对高质量推荐的实时性要求。
* 新用户可参考的数据较少。
* 顾客的数据变动性较大。

传统的基于用户的协同过滤方法在计算量上较为昂贵，在大规模的数据下不可行，对数据降维则会造成推荐质量的下降。基于用户的聚类模型在大数据下有更好的可扩展性，但是推荐质量较差，除非增加细粒度的用户划分，但是会导致计算复杂度接近协同过滤。基于搜索的方法容易出现推荐过于宽泛（比如推荐所有畅销 DVD）或过于狭窄（只推荐同作者的书籍）。

论文中提出的 Item-to-item 协同过滤则是基于项目的推荐，每个项目对于购买过的用户购买的其他项目计算相似性，构建相似矩阵，然后对于用户的推荐就只取决于用户购买过的项目数量，提高速度，对于购买历史较少的新用户也可以取得较好的推荐效果。

## 字节跳动（ByteDance）

字节跳动是一家新兴的独角兽公司，其包括抖音、今日头条等多个产品都是以推荐算法为技术核心。

字节跳动公司目前尚未有出版讲述其产品推荐系统细节的论文，但是今日头条的曹欢欢博士曾分享过其推荐算法原理[[7]](#footnote-8)。今日头条的推荐系统以以下三个维度的变量作为输入：

* 内容：包括视频、图文、问答等，每种类型都具有不同的特征
* 用户特征：包括兴趣标签，职业、年龄、性别等
* 环境特征：包括用户访问时的所在场景等，考虑到用户在不同环境下的信息偏好可能不同

推荐算法模型拟合的目标则可以包括：用户点击率（CTR）、内容阅读时间、点赞、转发情况等。可选的算法包括协同过滤、GBDT 等等，目前主要使用的是LR 和DNN 结合。

在内容分析方面，包括了文本分析、图片分析以及视频分析。通过对内容进行标签分析，可以对用户进行兴趣建模，以及通过语义特征、相似度特征、时空特征等为内容建立推荐策略。

对于推荐系统的评估，今日头条也是使用的主流的 A/B 测试方式，测试过程中收集用户动作数据，收集后进行日志处理、分布式统计、写入数据库，由实验平台生成数据对比、置信度、数据总结以及优化建议。

# 总结与展望

本文首先简要介绍了推荐系统算法的背景和意义以及本文所作大致工作内容，并分析了当前推荐系统的研究现状，其次阐述了主要的推荐方式和推荐结果的评估指标，进而分析了传统主流的经典推荐算法以及它们各自的优缺点，以及现代使用深度学习算法、神经网络等前沿技术所构建的更高性能推荐系统介绍，最后列举了推荐系统在众多著名公司中的工程应用实践。

推荐系统的发展一方面精确的匹配了用户与信息，降低了人们在信息过载时代获取信息的成本，但由推荐系统主导的内容分发，如新闻推荐等，也为用户带来了消极影响。可见推荐系统的发展不仅需要满足用户多元化、个性化的需求，而且需要对信息进行严格的监管和过滤，提高推荐系统的健壮性。近年来，RecSys会议上越来越多地收录了关于用户隐私、推荐引擎健壮性、信息过滤等方面的论文，这是未来推荐系统发展的一个重要研究方向。

目前，深度神经网络发展迅速，为推荐系统提供了新的特征提取、排序方法，越来越多的推荐引擎将深度神经网络与传统的推荐算法进行了结合，用于解决数据稀疏、推荐排序、社交网络等问题。工业界许多企业也已经将基于深度学习的推荐系统引入生产环境。尽管一些工作也对深度学习为推荐系统带来的提升提出质疑（Dacrema 等人（2019）[19]），但可以预见深度神经网络和推荐系统的结合将是推荐系统未来的主要研究方向。

综上所述，推荐系统是一个庞大的信息系统，它不仅仅只依赖于推荐引擎的工作，而且依赖于业务系统、日志系统等诸多方面，并结合了网络安全、数据挖掘等多个研究领域，能够为企业和用户带来价值，是一个值得深入研究的领域，更是一个具有无穷发展潜力的创新领域。

# 参考文献

1. R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. Data Mining and Knowledge Discovery, 18(1), 2009.
2. Chapelle, O., Joachims, T., Radlinski, F., and Yue, Y. 2012. Large-scale validation and analysis of interleaved search evaluation. ACM Trans. Inf. Syst. 30, 1, Article 6 (February 2012), 41 pages.
3. J. Aurisset Technology Blog, M. Ramm, and J. Parks, “Innovating Faster on Personalization Algorithms at Netflix Using Interleaving,” Medium, 30-Nov-2017. [Online]. Available: https://medium.com/netflix-techblog/interleaving-in-online-experiments-at-netflix-a04ee392ec55. [Accessed: 30-Dec-2019].
4. A. Gunawardana, G. Shani, A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. Journal of Machine Learning, Volume 10, pp.2935-2962
5. M. J. Pazzani, “A framework for Collaborative, Content-Based and Demographic Filtering”, Artificial Intelligence Review, Springer, 1999
6. P. Resnick, N. Iacovou, etc. “GroupLens: An Open Architecture for Collaborative Filtering of Netnews”, Proceedings of ACM Conference on Computer Supported Cooperative Work, CSCW 1994. pp.175-186.
7. B. Sarwar, G. Karypis, J. Konstan, J. Riedl, “Item-Based Collaborative Filtering Recommendation Algorithms”, Proceeding of the 10th international conference on World Wide Web, WWW 2001.
8. Shani, G. & Gunawardana, A. Evaluating recommendation systems Recommender systems handbook, Springer, 2011, 257-297
9. Yao, Y. Measuring retrieval effectiveness based on user preference of documents, Journal of the American Society for Information science, Wiley Online Library, 1995, 46, 133-145
10. John S. Breese, David Heckerman, and Carl Myers Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In UAI, pages 43–52, 1998.
11. Kalervo Jarvelin and Jaana Kek ¨ al¨ ainen. Cumulated gain-based evaluation of ir tech- ¨ niques. ACM Trans. Inf. Syst., 20(4):422–446, 2002. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/582415.582418.
12. Wenqi Fan, Yao Ma, Dawei Yin, Jianping Wang, Jiliang Tang, and Qing Li.  
    2019. Deep Social Collaborative Filtering. In *Thirteenth ACM Conference*  
    *on Recommender Systems (RecSys ’19), September 16–20, 2019, Copenhagen,*  
    *Denmark.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/  
    3298689.3347011
13. Malte Ludewig, Noemi Mauro, Sara Latif, and Dietmar Jannach. 2019. Performance Comparison of Neural and Non-Neural Approaches to Session-based Recommendation. In *Thirteenth ACM Conference on Recommender Systems* *(RecSys ’19), September 16–20, 2019, Copenhagen, Denmark.* ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3298689.3347041>
14. Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019.  
    Neural Graph Collaborative Filtering. In *Proceedings of the 42nd International*  
    *ACM SIGIR Conference on Research and Development in Information Retrieval*  
    *(SIGIR ’19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA,  
    10 pages. <https://doi.org/10.1145/3331184.3331267>
15. Covington, P.; Adams, J. & Sargin, E. “Deep neural networks for youtube recommendations” Proceedings of the 10th ACM conference on recommender systems, 2016, 191-198
16. A. Chandrashekar, F. Amat, J. Basilico, and T. Jebara, “Artwork Personalization at Netflix,” Medium, 07-Dec-2017. [Online]. Available: https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76. [Accessed: 02-Jan-2020].
17. L. Li, W. Chu, J. Langford, and X. Wang, “Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms,” in Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, New York, NY, USA, 2011, pp. 297–306.
18. Linden, G.; Smith, B. & York, J. “Amazon. com recommendations: Item-to-item collaborative filtering” IEEE Internet computing, IEEE, 2003 , 76-80
19. Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In Thirteenth ACM Conference on Recommender Systems (RecSys ’19), September 16–20, 2019, Copenhagen, Denmark. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3298689.3347058

1. https://grouplens.org/datasets/movielens/ [↑](#footnote-ref-2)
2. http://eigentaste.berkeley.edu/dataset/ [↑](#footnote-ref-3)
3. https://www.yelp.com/dataset/challenge [↑](#footnote-ref-4)
4. https://grouplens.org/datasets/hetrec-2011/ [↑](#footnote-ref-5)
5. https://www.kaggle.com/netflix-inc/netflix-prize-data [↑](#footnote-ref-6)
6. https://research.netflix.com/research-area/recommendations [↑](#footnote-ref-7)
7. https://www.toutiao.com/i6511211182064402951/ [↑](#footnote-ref-8)