****



**研 究 生 毕 业 论 文**

**（申请工程硕士学位）**

|  |  |
| --- | --- |
| **论文题目** | 基于用户历史行为的详情页个性化  推荐的方法研究 |
| **作者姓名** | 何鑫 |
| **学科、专业名称** | 软件工程（软件工程） |
| **研究方向** | 软件工程 |
| **指导教师** | 葛季栋　副教授 |

**2019年 5月 24 日**

**学 号： MF1732046**

**论文答辩日期： 2019 年 5 月 24 日**

**指 导 教 师： （签字）**

**基于用户历史行为的详情页个性化推荐的方法研究**

|  |  |
| --- | --- |
| **作 者:** | **何鑫** |
| 指导教师: | **葛季栋　副教授** |

|  |
| --- |
| **南京大学研究生毕业论文** |
| **(申请工程硕士学位)** |

|  |
| --- |
| **南京大学软件学院** |
| **20**19**年**05**月** |

**Research on Personalized Recommendation Method of Detail Page Based on**

**User’s History Behavior**

**He, Xin**

**Submitted in partial fulfillment of the requirements for the degree of Master of Engineering**

Supervised by

Associate Professor **Ge, Jidong**

Software Institute

**NANJING UNIVERSITY**

Nanjing, China

May, 2019

# 摘 要

近年来，随着互联网的普及，以及移动终端的普及。论坛作为集发布信息、聊天和讨论为一体的平台，已随着互联网和移动终端的普及，而变得越来越重要了。公路物流运输APP是一个能够提供货主线上发货，司机线上找货提供服务。使用互联网技术，促进货主和司机更快捷、更高效、更公平的交易。而论坛作为公路物流运输APP的一个服务，它能够给司机用户展示咨询信息，促进司机行业的交流，把每个孤立的司机在网络中连接起来。

本文针对传统论坛推荐的不能够很好的利用用户信息做个性化推荐问题。提出了一种能够利用用户行为序列的来做详情页推荐的方法。

本文首先分析了详情页推荐问题的本质，概括了详情页推荐问题的特点，描述了详情页推荐的问题以及面临的困难。其次本文对论坛那推荐的中数据进行了分析，在现有可用的数据设计推荐方法。并提出了帖子的预处理方法。接着设计了两种基于用户浏览帖子的历史做推荐的方法，分别提出了基于概率矩阵分解的详情页个性化推荐的方法和基于用户行为序列的详情页个性化推荐的方法，概括如下：

（一）基于概率矩阵分解的详情页个性化推荐方法：由于详情页个性化推荐本质衡量两个帖子在用户这一变量限制下相似度的问题。该方法在概率矩阵分解中引入帖子内容特征和用户历史特征。使用CNN来提取帖子内容的文本特征；把用户在一段时间浏览的帖子看作用户特征，把用户浏览记录中的帖子对应的向量进行对应维度求和取平均的方法来简单构造用户特征。然后使用概率矩阵分解方法来计算用户特征和帖子特征的相似度进行详情页的个性化推荐。

（二）基于用户行为序列的详情页个性化推荐的方法：是在基于概率矩阵分解的基础上，重新对用户特征进行了设计，通过把用户在一段时间内浏览的帖子作为一个Session，使用GRU来处理这个Session中的帖子作为用户行为序列特征。然后使用概率矩阵分解方法来计算用户行为序列特征和帖子特征的相似度来进行详情页的个性化推荐。

在实验验证阶段，本文使用公路物流运输领域的论坛帖子和用户浏览的帖子的真实数据进行效果评估，通过多组实验，设置了基于最热门帖子推荐方法、基于文档相似度的推荐方法作为对照组和两种基于用户历史行为数据的推荐方法进行对比；并比较了基于概率矩阵分解的详情页个性化推荐方法与基于用户行为序列的详情页个性化推荐的方法的效果，验证了本文提出基于用户行为序列的详情页个性化推荐方法的有效性，并对比了两个方法的优缺点，阐述了未来可能的工作方向。

**关键词**：论坛推荐、详情页推荐、个性化推荐、概率矩阵分解、用户行为序列

# Abstract

In recent years, with the popularity of the Internet, as well as the popularity of mobile terminals. BBS, as a platform for publishing information, chatting and discussion, has become more and more important with the popularity of the Internet and mobile terminals. Highway logistics transportation APP is a service that realizes online delivery of goods by the owner and offline delivery by the driver. Use the Internet to facilitate faster, more efficient and fair transactions between shippers and drivers. Similarly, as a service of road logistics transportation APP, BBS is able to show consulting information to drivers and users, promote communication among drivers and connect each isolated driver in the network.

This thesis aims at the problem that traditional BBS recommendation cannot make good use of user information to make personalized recommendation. A method of detail page recommendation based on user behavior sequence is proposed.

This thesis first analyzes the nature of detail page recommendation problem, summarizes the characteristics of detail page recommendation problem, and describes the problems and difficulties of detail page recommendation. Secondly, this paper analyzes the recommended data of BBS, and designs the recommended methods in the available data. And the pretreatment method of the post is put forward. Then, two recommendation methods based on the history of users' browsing posts are designed. The personalized recommendation method of detail page based on probability matrix decomposition and the personalized recommendation method of detail page based on user behavior sequence are proposed respectively, which are summarized as follows:

(I) detailed page personalized recommendation method based on probability matrix decomposition: because the personalized recommendation of detailed page essentially measures the similarity of two posts under the limitation of the user variable. In this method, the features of post content and user interest are introduced into the probability matrix decomposition. Use CNN to extract the text features of the content of posts; The posts viewed by users in a period of time are regarded as the characteristics of users' interests, and the user characteristics are simply constructed by summing the corresponding dimensions of the “? post vectors and averaging them. Then the probability matrix decomposition method is used to calculate the similarity of users' interests and posts for personalized recommendation of detail pages.

Details page (2) based on user behavior sequence of personalized recommendation methods: is based on probability matrix decomposition, on the basis of the design characteristics of the user is done again, by the user for a period of time to browse posts as a Session, using RNN to handle your post to this Session as user interest in dynamic characteristics. Then the probability matrix decomposition method is used to calculate the similarity of users' dynamic interests and posts to make personalized recommendations for detail pages.

In experiment phase, this thesis USES the BBS posts, and in the field of highway logistics transport users browse real data for evaluation of post, through many experiments, set up based on the most popular post recommendation method, based on the recommendations from the history of each user behavior of the most popular post, post recommendation algorithm based on simple similarity as control group and two recommendation method based on user history behavior data comparison; And compare the details page personalized recommendation method based on probability matrix decomposition and details page personalization recommendation based on user behavior sequence, the effect of the method is verified in this paper, based on user behavior sequence details page of the effectiveness of personalized recommendation method, and compared the advantages and disadvantages of the two methods, this paper expounds the direction of future work.

**Keyword**: BBS Recommendation, Detail Page Recommendation, Personalized Recommendation, Probability Matrix Decomposition, User Behavior Sequence

**目 录**

[摘 要 I](#_Toc4379549)

[Abstract II](#_Toc4379550)

[图目录 VI](#_Toc4379551)

[表目录 VII](#_Toc4379552)

[第一章 引言 1](#_Toc4379553)

[1.1 研究背景 1](#_Toc4379554)

[1.2 国内外研究现状 1](#_Toc4379555)

[1.3 本文研究内容与贡献 3](#_Toc4379556)

[1.4 本文的组织结构 4](#_Toc4379557)

[第二章 技术综述 5](#_Toc4379558)

[2.1 概率矩阵分解 5](#_Toc4379559)

[2.2 卷积神经网络 6](#_Toc4379560)

[2.3 循环神经网络 7](#_Toc4379561)

[2.4 嵌入式词向量 10](#_Toc4379562)

[2.5 Redis 10](#_Toc4379563)

[2.5.1 Redis数据结构 11](#_Toc4379564)

[2.5.1 Redis持久化 12](#_Toc4379565)

[2.6 本章小结 12](#_Toc4379566)

[第三章 基于用户行为序列的论坛推荐系统的设计 13](#_Toc4379567)

[3.1 论坛推荐问题分析 13](#_Toc4379568)

[3.1.1 详情页推荐问题描述 14](#_Toc4379569)

[3.1.2 帖子和用户数据分析 15](#_Toc4379570)

[3.2 系统总体设计 17](#_Toc4379571)

[3.3 帖子内容预处理 18](#_Toc4379572)

[3.3.1 帖子特征提取 19](#_Toc4379573)

[3.3.2 帖子中文分词 19](#_Toc4379574)

[3.3.3 构建帖子停用词库 19](#_Toc4379575)

[3.3.4 去除帖子停用词 20](#_Toc4379576)

[3.3.5 帖子向量化表示 21](#_Toc4379577)

[3.4 基于概率分解的详情页推荐 21](#_Toc4379578)

[3.4.1 具体框架 21](#_Toc4379579)

[3.4.2 构造用户特征 22](#_Toc4379580)

[3.4.3 提取帖子特征 22](#_Toc4379581)

[3.4.4 相似度计算 24](#_Toc4379582)

[3.5 基于用户行为的详情页推荐 25](#_Toc4379583)

[3.5.1 具体框架 25](#_Toc4379584)

[3.5.2 构造用户特征 26](#_Toc4379585)

[3.5.3 提取帖子特征 28](#_Toc4379586)

[3.5.4 相似度计算 28](#_Toc4379587)

[3.6 本章小结 29](#_Toc4379588)

[第四章 基于用户行为序列的论坛推荐系统的实现 30](#_Toc4379589)

[4.1 数据集 30](#_Toc4379590)

[4.2 评估指标与评估方式 30](#_Toc4379591)

[4.3 基于概率矩阵分解的详情页推荐的实现 31](#_Toc4379592)

[4.4 基于用户行为序列的详情页推荐的实现 32](#_Toc4379593)

[4.5 对比实验 33](#_Toc4379594)

[4.5.1 基于热门帖子的推荐 34](#_Toc4379595)

[4.5.2 基于文档相似度的推荐 36](#_Toc4379596)

[4.5.3 基于概率矩阵分解的推荐 38](#_Toc4379597)

[4.5.4 基于用户行为序列的推荐 40](#_Toc4379598)

[4.6 方法的优缺点分析 42](#_Toc4379599)

[4.7 本章小结 44](#_Toc4379600)

[第五章 基于用户行为序列的论坛推荐系统的测试 45](#_Toc4379601)

[5.1 总结 45](#_Toc4379602)

[5.2 进一步工作展望 46](#_Toc4379603)

[参 考 文 献 47](#_Toc4379604)

[致 谢 51](#_Toc4379605)

[版权及论文原创性说明 52](#_Toc4379606)

# 图目录

[图2.1 PMF结构图 5](#_Toc4391473)

[图2.2 TextCNN结构图 7](#_Toc4391474)

[图2.3 RNN节点的结构示意图 8](#_Toc4391475)

[图2.4 LSTM节点的结构示意图 9](#_Toc4391476)

[图2.5 GRU节点的结构示意图 10](#_Toc4391477)

[图3.1详情页推荐问题的示例图 15](#_Toc4391478)

[图3.2详情页推荐问题的框架 18](#_Toc4391479)

[图3.3 基于概率矩阵分解的推荐流程 23](#_Toc4391480)

[图3.4 用户行为序列 24](#_Toc4391481)

[图3.5 CNN模型提取帖子特征流程 24](#_Toc4391482)

[图3.6 CNN模型结构 25](#_Toc4391483)

[图3.7 基于概率矩阵分解计算流程 26](#_Toc4391484)

[图3.8 基于用户行为序列的推荐流程 27](#_Toc4391485)

[图3.9 用户特征构造的流程 28](#_Toc4391486)

[图3.10 用户行为序列（一） 29](#_Toc4391487)

[图3.11 用户行为序列（二） 29](#_Toc4391488)

[图3.12 用户行为序列（三） 29](#_Toc4391489)

[图3.13 GRU模型的结构 30](#_Toc4391490)

[图4.1 ConvMF模型代码 34](#_Toc4391491)

[图4.2 GRUforRes代码 35](#_Toc4391492)

[图4.3 基于热门帖的推荐的帖子占比 37](#_Toc4391493)

[图4.4 基于热门的帖的推荐的点击率 38](#_Toc4391494)

[图4.5 基于文档相似度的推荐的帖子占比 39](#_Toc4391495)

[图4.6 基于文档相似度的推荐的点击率 40](#_Toc4391496)

[图4.7 基于概率矩阵分解的推荐的帖子占比 41](#_Toc4391497)

[图4.8 基于概率矩阵分解的推荐的点击率 42](#_Toc4391498)

[图4.9 基于用户行为序列的推荐的帖子占比 43](#_Toc4391499)

[图4.10 基于用户行为序列的推荐的点击率 43](#_Toc4391500)

[图4.11 四种模型的推荐的帖子数占比的对比 44](#_Toc4391501)

[图4.12 四种模型的点击率对比 45](#_Toc4391502)

# 表目录

[表 2.1 Redis中的五种数据结构 11](#_Toc4379205)

[表 3.1用户浏览帖子历史信息的举例 15](#_Toc4379206)

[表 3.2帖子详细信息的举例 16](#_Toc4379207)

[表 4.1四种模型的比较 34](#_Toc4379208)

# 第一章 引言

## 1.1 研究背景

公路物流APP是一个通过提供货主线上发货，司机线上找货的服务，来改善传统公路物流行业的信息不对称、货主发货慢、司机找货难的问题。论坛作为APP中的一项服务，给司机提供了一个信息交流的平台。随着APP注册用户数量的增加，线上用户逐渐的活跃，论坛的帖子呈现爆发的增长。传统以推荐热门帖和新帖的方式使推荐具有很强的同质化，这种推荐方式产生了长尾效应，使推荐给用户的帖子集中在少数的帖子上。为了能够实现根据用户历史的行为，为用户个性化的推荐不同的帖子，从而减少用户获取相关帖子信息的时间，增加用户对推荐帖子信息的满意程度。

本文的研究目的是设计一种基于用户历史行为信息构建用户特征的个性化推荐模型和算法以解决用户在浏览帖时推荐其它相关帖子的准确度的问题。该课题的意义在于通过对用户进行个性化的推荐，提升用户的使用满意度，提升论坛服务对用户的黏性；通过准确推荐不同帖子的起到间接对用户进行信息的收集作用（即帖子相当于一类标签），能够为公路物流这个行业的司机画像提供重要的数据支持；能够根据用户浏览帖子的数据分析用户类群关系，形成不同兴趣的社区，为平台其它功能服务设计开发提供重要参考依据。

## 1.2 国内外研究现状

个性化推荐一直是推荐领域的热点问题。该问题的研究可以被分为两个方向：

（1）通过对文档信息的进行建模来提高预测文档相关性的准确度。

Wang等人通过使用LDA主题模型和堆栈去噪自动编码器的方法来利用文档的上下文信息提升传统协同过滤算法的评分预测的精度[Wang et al, 2011]。提出了协作主题模型。

Wang等人考虑到传统的协同过滤算法不能够处理大量的数据。因此，把基于LDA主题模型和堆栈去噪自动编码器的方法整合到概率矩阵分解中，能够在评分预测精度方面产生更为准确的潜在模型，从而提出了协同深度学习[Wang et al, 2015]。

Kim等人考虑到像LDA这种主题模型，只是使用了词袋模型忽略句子中词的顺序特征，从而不能完全的捕获到文档的信息[Kim et al, 2016]。Kim使用卷积神经网络，这种能够捕获文本局部特征的模型行来替代LDA主题模型，同样考虑能够处理大量数据的能力，使用PMF来代替传统的协同过滤算法。

（2）通过对用户行为进行建模提升推荐信息和用户匹配程度。

协同过滤作为推荐系统经典算法，是一种基于共现关系的算法。这种算法把每个物品相互独立，不能对用户浏览过的物品进行连续偏好建模建模。

针对这个问题Linden提出了基于物品的协同过滤算法[Linden et al,2003],这种算法通过分析用户的行为记录计算物品之间的相似度，能够利用用户的历史的行为进行推荐，相对于基于用户的协同过滤算法具有更好的解释性。但是这种算法只考虑了用户最后一次浏览的物品的相似性，忽略了用户之前浏览物品的信息，没有考虑用户的浏览的整个序列信息。

针对这个问题，Shani提出了马尔可夫决策过程的算法[Shani et al,2002]，他使用了四元组<状态，动作，转移概率，奖励函数>来刻画序列信息，使用状态转移概率的计算浏览下一个动作：即浏览物品的概率。这种算法虽然能够很好的用户历史信息，但是随着物品的增加，状态转移矩阵的维度会呈现指数怎加。

由于循环神经网络具备处理序列数据的天然优势，能把前一时刻的隐藏信息和输入，作为下一时刻的隐藏层信息的输入，从而达到能够处理序列信息。Hidasi等人首次把RNN应用到推荐系统的用户历史行为记录的处理中来，这种算法把用户历史行为中浏览的物品作为一个序列，使用RNN进行刻画[Hidasi et al, 2015]。

由于Hidasi等人使用RNN处理用户历史行为取得非常好的效果，Hidasi在这个算法的基础上探究了如何把每个物品的属性信息添加到RNN算法中[Hidasi et al, 2016]。

Bogina等人考虑到用户在每个物品上停留的时间长短，能够区分出用户对物品的喜好程度[Bogina et al,2017]。因此，Bogina通过对用户历史行为记录中的物品的停留时间设定单位时间阈值，将每个物品按照但是划分长小的时间片，按照与Hidasi等人的做法进行训练，通过实验证明这种法确实有效。

Quadrana等人使用来给你个RNN对用户的历史行为进行的不同级别的刻画。提出了一种层次化的RNN模型[Quadrana et al,2017]。他使用Session-level GRU对用户Session信息进行刻画；使用User-level GRU刻画用户的历史信息。通过这种方式能够很好的对用户的兴趣变化进行刻画。并通过实验证命这种方法的有效性。同时使用用户历史行为建模和物品信息建模来进行用户和物品的评分预测，目前还没有很好的方法。

本文在前面方法的基础上提出一种能够同时使用用户历史行为建模和物品信息进行建模的方法，进行帖子推荐。

## 1.3 本文研究内容与贡献

使用用户行为序列建模，来提升用户和帖子的评分估计已经被广泛的研究和应用；同时，使用文档信息进行建模来进行推荐的方法也被广泛的研究和应用。然而，使用用户历史行为和文档信息同时来进行的推荐的方法，却鲜有人进行研究。另一方面，由于论坛中的推荐，经常被一些简单的推荐算法占据。推荐热门帖、最新的帖子，是一些论坛惯用的做法。这种推荐方法虽然也能很好的提升用户的信息匹配的满意，推荐热门、新鲜的帖子，这样会减少帖子的推荐量，让一些不够热门和时间稍微滞后的帖子得到很少或者得不到展示。这种推荐方法没有起到推荐系统挖掘信息长尾的作用。因此，我们尝试在论坛引入能够利用用户历史信息和文档信息推荐算法来提升帖子的个性化推荐能力，来达到挖掘长尾信息的能力。另一方面，本文探究了使用不同方式处理用户历史信息的推荐算法的推荐效果的差异。

本文首先对论坛详情页个性化推荐问题的本质进行剖析，总结了论坛数据和用户浏览历史数据的特点。其次本文提出了针对帖子的预处理方法，其中包括构建帖子停用词库、帖子停用词去除、帖子向量化表示等。然后我们对用户特征和帖子特征进行建模，提出两种基于用户历史行为与文档信息的详情页个性化推荐方法。

在实验验证阶段，本文使用公路物流运输领域的论坛真实的帖子数据进行效果评估，通过四组实验，对比了基于热门帖的详情页推荐方法、基于文档相似度的详情页推荐方法和两种基于用户历史行为与文档信息的详情页个性化推荐方法的效果，验证了本文提出的基于用户行为序列建模的详情页个性化推荐方法的有效性，并对比了四种方法的优缺点，阐述了未来可能的工作方向。

本文的有两个重要的贡献，一是探究个性化推荐算法在公路物流这个比较垂直领域的论坛详情页个性化推荐的实践方法；二是，提出一种能够同时融合用户历史行为信息和文档信息的个性化推荐方法；提出了使用用户行为序列建模在个性化推荐算法。

## 1.4 本文的组织结构

本文的组织结构如下：

第一章：引言部分。本章介绍了论文的研究背景、研究现状和本文的研究内容与贡献。

第二章：相关背景知识。本章介绍了论文的相关背景知识，介绍了概率矩阵分解、卷积神经网络、循环神经网络、词向量的嵌入技术和Redis数据库等知识。

第三章：基于用户历史行为的论坛详情页个性化推荐方法的设计。本章介绍了详情页个性化推荐方法，详细阐述了详情页推荐数据的特点和详情页个性化推荐需要解决的问题以及面临的难点。提出了针对帖子的通用预处理方法，并详细介绍了本文提出的两种详情页推荐方法，包括基于概率矩阵分解的详情页个性化推荐方法和基于用户行为序列建模的详情页个性化推荐方法。

第四章：基于用户历史行为的论坛详情页个性化推荐方法的实现。本章描述了实验数据集、实验评估指标和评估方式、相关实验设计、实验过程和实验结果，通过一系列实验验证了本文提出的详情页个性化推荐方法的有效性，并对比了本文提出的两种详情页推荐推荐方法的优缺点。

第五章：总结与展望。本章对本文的研究工作进行了总结，并对未来的工作方向做了进一步展望。

# 第二章 技术综述

本章将介绍论文的相关理论知识，首先在2.1节对概率矩阵分解进行介绍，PMF是现代推荐系统方法，在此基础衍生了很多算法，本文在其中一种衍生算法ConvMF基础上，进行了改进。在2.2.节将对卷积神经网络进行介绍，本文将使用CNN来提取帖子文本的特征。在2.3将对集中循环神经网络进行介绍，本文将使用其中一种循环神经网络GRU来对用户行为序列进行建模。在2.4节将对嵌入式词向量进行介绍，嵌入词向量技术是一种常用的文本表示方法，本文将使用嵌入式词向量来表示帖子内容。在2.5节将对Redis数据库进行介绍，本文将使用Redis数据库来存储中间数据。最后，是对本章的小结。

## 2.1 概率矩阵分解

概率矩阵分解[Mnih et al,2007]（Probabilistic Matrix Factorization，简称PMF）是由多伦多大学Mnih等人在2007年提出来的。PMF很好的了MF不能够对大数据和稀疏数据进行处理的缺陷，PMF算法是现代推荐系统的基础算法之一。由于PMF的这个特型，ConMF[Kim et al,2016]在PMF基础上使用CNN获取文档的上下文信息，实现更精准对相似分数进行估计。结构如图2.1所示：



图2.1 PMF结构图

在推荐系统中，用户和物品的关系矩阵往往是一个稀疏的，PMF是基于low-dimensional factor模型来处理这个问题。即，认为用户和物品的关系可以由较少的几个因素的线性组合决定。

在PMF中有两个基本假设：

（1）观测噪声为高斯分布；

（2）用户向量和物品向量均为高斯分布。

基于第一假设，我们计算出观测矩阵的概率密度函数：

其中是观测噪声的方差，需要指定。

基于第二个假设，我们能够计算出用户、物品的概率密度函数：

其中是先验噪声的方差，也需要指定。

综合上面两个概率密度函数，利用后验概率推到，得：

## 2.2 卷积神经网络

卷积神经网络(Convolutional Neural Networks,简称CNN)[Krizhevsky et al,2012]是Krizhevsky等人在2012年对ImageNet进行分类提出来的，Krizhevsky使用AlexNet在ImageNet比赛中取得冠军的成绩，这使得基于卷积的神经网络成为图像处理的经典算法。在此之后，出现了大量CNN模型。

在此之后Kim把CNN应用到文本分类任务，提出了TextCNN[Kim et al,2014].TextCNN利用多个不同大小的卷积核来提取句子中的关键信息，这类似于多个窗口大小的ngram，从而能够很好捕捉文本局部相关性。TextCNN结构如图2.2所示：



图2.2 TextCNN结构图

（1）Embedding层：该层的输入通过词向量化把文档表示成一个文档矩阵。首先要对文档进行预处理，要去除文档中的非汉字和字母的字符。为了能够精确提取文档特征，要去除高频且没有具体含义的停用词。然后把所有文档的词进行预训练。然后把文档装换成文档矩阵进行输入。

（2）卷积层：卷积层是对文本进行特征抽取。对于卷积核，其中是卷积核滑动的窗口大小，是词向量的维度，卷积核相当于从文档开头不断向后滑动，通过神经网络的非线性变换嫁给你这个窗口内的输入值转换成文档的特征值，随着窗口不断地向后滑动，持续的产生文档特征值，最终形成了这个卷积核对应的特征向量。此过程相当于获取文档k-gram特征。

（3）池化层：池化层通过对卷积层的特征进行降维处理，产生最终的特征。池化层有多种不同的池化操作，Max-Pooling是一种最常用的降维操作，这种操作能够提取卷积层提取的特征中的最强的特征。

（4）输出层：在池化层之后通过连接全连接层神经网络，最终形成分类的过程。

## 2.3 循环神经网络

循环神经网络(Recurrent Neural Network,简称RNN)，由Rumehart等人提出来的一类以序列数据为输入,在序列的演进方向进行递归且所有节点按链式连接的递归神经网络[Rumelhart et al, 1986]。

循环神经网络具有记忆性、参数共享，因此能以很高的效率对序列的非线性特征进行学习。

### 2.3.1 RNN

RNN是一种能够处理序列数据的神经网络。如图2.2所示，是RNN单个节点的机构示意图。



图2.3 RNN节点的结构示意图

RNN某一节点的输入有两个，分别是上一层的输出的隐状态和当前序列的值,下一隐状态和的计算过过程为：

其中为激活函数。

由于RNN的神经网络中的激活函数通常都是Sigmoid函数，这种函数具有饱和的特型，当输入的值满足一定范围，输出的变化就不是很明显。因此，随着序列的增加会出现梯度消失和梯度爆炸的问题。

### 2.3.2 LSTM

长短期记忆神经网络（long short-term memory,简称LSTM）由Hochreiter等于人在1997年提出来，是一种门控RNN[Hochreiter et al, 1997 ]。LSTM通过引入了遗忘门、输入门和输出门三种门控结构来解决RNN神经网络梯度消失和梯度爆炸的问题。LSTM的结构如图2.3所示：



图2.4 LSTM节点的结构示意图

LSTM的输入门、遗忘门、细胞态和输出门更新公式如下所示：

（1）遗忘门更新：

（2）输入门更新：

（3）细胞态更新：

（4）输出门更新：

### 2.3.3 GRU

门控循环单元（GRU）[Cho et al, 2014],通过使用单个门控制单元来同时控制输入门和遗忘门，GRU简化了LSTM内部结构，参数要比LSTM少，计算成本要比LSTM少。GRU结构如图2.4所示：



图2.5 GRU节点的结构示意图

GRU的更新：

（1）更新门更新：

（2）重置门更新：

（3）隐状态更新：

## 2.4 嵌入式词向量

自然语言处理中，词有两种表示方式，一种基于词袋法，把每个词看成是无序的，通过独热编码的编码把词表示成一个高维稀疏的向量。另一种基于词共现关系把词汇表示成一个低维稠密的向量，这种方法也被称为词嵌入方法。

由于嵌入式法能够捕获到词袋法不能捕获词的顺序和周围词的关系，所以嵌入式法作为一种更有效的词向量的表示方法。嵌入式法有两个比较经典的代表，Word2Vec和Glove。

Word2Vec的网络结构很简单，包括一个输入层、一个隐藏层、一个输出层。其中，输入层对应某个（上下文）单词的独热编码向量（共有V个词汇），输出层为与输入单词同时出现的单词的概率分布，换句话说，词汇表中的每个单词，出现在这一上下文中的概率分别是多少。隐藏层由N个神经元组成。经过训练之后，我们使用输入层和隐藏层之间的连接权重矩阵WVxN表示单词之间的关系。矩阵W共有V行，每一行都是一个N维向量，每个N维向量分别对应一个单词。这样，词向量的维度就从V降到了N。

而Glove的主要直觉是，相比单词同时出现的概率，单词同时出现的概率的比率能够更好地区分单词。Word2Vec中隐藏层没有使用激活函数，这就意味着，隐藏层学习的其实是线性关系。基于以上两点想法，Glove提出了一个加权最小二乘回归模型，输入为单词-上下文同时出现频次矩阵。在某些场景下，Glove的表现优于Word2Vec。

## 2.5 Redis

Redis是一内存数据库数据库。相对于传统的关系型数据库，Redis是一种非关系的数据库，它可以存储键和五种不同类型的值之间的映射，可以将存储在内存的剑指对数据持久化到硬盘，可以使用复制特性来扩展读性能，还可以使用客户端分片来扩展写性能[Carison et al, 2015]。下面我们将对Redis的数据结构和持久化进行介绍。

### 2.5.1 Redis数据结构

Redis数据库有五种数据结构，这五种数据结构分别是：STRING（字符串）、LIST（无序集合）、SET（集合）、HASH（哈希表）和ZSET（有序集合）。我们将对这五种数据结构的存储的值类型和读写能力进行介绍，如表2.1所示：

表 2.1 Redis中的五种数据结构

|  |  |  |
| --- | --- | --- |
| 结构类型 | 结构存储的值 | 结构的读写能力 |
| STRIGN | 其值可以为字符串、证书或者浮点数 | 对整个字符串或者二字符串的其中一部分执行操作；对整数和浮点数执行自增或者自减的操作 |
| LIST | 一个链表，链表上的每一个节点都包含了一个字符串 | 从链表的两端插入或者淡出元素；根据偏移量对链表进行修建；读取每个或者多个元素；根据值查找或者删除元素 |
| SET | 包含字符串的无序收集器，并且被包含的每个字符串都得独一无二、各不相同的 | 田间、获取、删除单个元素；检查一个元素是否存在于集合中；计算交集、并集和差集；从集合里随机获取元素 |
| HASH | 包含键值对的无序散列表 | 添加、获取和删除单个剑指对；获取所有键值对 |
| ZSET(有序集合) | 字符串成员于浮点数分值之间的有序映射，元素的排列顺序由分支的大小决定 | 添加、获取和删除单个元素；格局分值范围或者成员来获取元素 |

### 2.5.1 Redis持久化

Redis的持久化可以让用户的数据存储到硬盘上，是的数在Redis重启之后不会消失。Redis提供了两种不同的持久化方法来将数据存储到本地硬盘上，分别是：快照和只追加文件。

快照持久化(snapshotting)是Redis通过创建快照的方式来获取存某个时间的上的内存里的副本。快照创建完毕之后，用户就可以把快照进行备份，把快照复制到其它服务器上，来达到捐建具有相同数据的服务求副本；另外，还可以将快照存储在本地，当服务器重启可以使用该快照进行数据恢复。

只追加文件持久化(append-only file,简称AOF持久化)，AOF持久化将被执行的写命令写道AOF文件的末尾，一次来记录数据发生变化。因此，Redis只要从头到尾重新执行一次AOF文件包含的所有写命令，就可以恢复AOF文件所记录的数据集。

## 2.6 本章小结

本章介绍了论文的相关理论知识，首先在2.1节对概率矩阵分解进行了介绍。在2.2.节对卷积神经网络进行了介绍，本文使用CNN来提取帖子文本的特征。在2.3对几种循环神经网络进行介绍，本文使用其中一种循环神经网络GRU来对用户行为序列进行建模。在2.4节对嵌入式词向量进行了介绍，嵌入词向量技术是一种常用的文本表示方法，本文使用嵌入式词向量来表示帖子内容。在2.5节对Redis数据库进行了介绍，本文使用了Redis数据库来存储中间数据。

# 第三章 基于用户历史行为的详情页个性化推荐方法的设计

本章将会介绍详情页推荐的方法。3.1节介绍详情页推荐的问题的分析，分析论坛推荐的两个问题，对详情页推荐问题进行定义；分析详情页推荐中帖子的特点，以及用于推荐数据的特点；3.2节介绍详情页推荐的框架，设计从数据处理到推荐的流程；3.3节介绍帖子数据的处理，用户特征的构造；3.4节将会介绍基于概率矩阵分解的详情页推荐算法；3.5节将会介绍基于用户行为序列建模的详情页推荐算法；3.6节对本章进行总结。

## 3.1 论坛推荐问题分析

公路物流运输的APP通过提供货主线上发货，司机线上找货的功能，来提高公路物流运输行业效率。论坛作为该APP中的其中一项服务，是司机进行交流沟通的平台。每个司机都可以把自己个人的想法和运输过程经历写成帖子和其他司机进行交流。尽管帖子的主题不同，但是帖子的话题关于公路运输行业的。

论坛的推荐推荐有两部分组成：论坛的首页推荐和论坛的详情页推荐。

首页主要功能主要是把用户吸引到论坛中，所以首页采用展示能够吸引绝大多数用户的热门的和最新的帖子。这种推荐策略，越简单越好，主要是通过统计每个帖子的浏览量和帖子的发布时间的两个统计量来进行推荐。这两个统计量之间存在着冲突，即发布越早的帖子浏览量越多，发布的新帖浏览量在短时间会很少。在做首页推荐的时候，需要根据推荐策略来来设定这个统计量的权重。论坛倾向于推荐热门帖，会把浏览量的权重设置大些；相反，论坛倾向于推荐新帖，会把帖子的发布时间的权重设置大些。

详情页作为帖子被点击后的帖子详细信息展示的页面，同时在详情页下面会有相关性推荐，详情页推荐的目的是想把用户留在这个论坛里，即通过不断推荐符合用户兴趣的帖子，来增加用户的在论坛中停留的时间。

本文研究的是详情页推荐，即根据用户当前浏览的帖子推荐相关的帖子。在用户浏览当前的帖子给用户推荐与当前帖子相关的5个用户感兴趣的帖子。为了区分首页帖子推荐，我们把这种基于当前帖子的推荐相似帖子的问题称为详情页推荐，这是因为这种推荐的相似帖子展示在该帖子详情页页面里。

在接下来的小节中，本文将着重描述详情页推荐推荐的问题，并分析帖子和用户数据的特点。

### 3.1.1 详情页推荐问题描述

详情页推荐是在用户点击某个帖子进入详情页的情况下，根据用户兴趣和详情页帖子内容推荐相关的帖子。如图3.1所示，这个是APP详情页的页面。页面的上班部分是用户点击浏览的帖子，包括：详情页帖子标题和详情页帖子内容；下半部分是相关推荐，是根据用户上下文信息和当前详情页帖子内容推荐5个最为相关的帖子。



图3.1详情页推荐问题的示例图

### 3.1.2 帖子和用户数据分析

本小节，我们将对公路物流运输领域的论坛详情页推荐的可用数据进行分析。由于公司对用户数据较为敏感，司机的信息属于一级保密的数据，导致我们在做论坛推荐只能用到司机ID和司机浏览帖子的信息。下面我们将对推荐场景用到的数据进行举例和分析。

用户浏览帖子的信息包括用户ID、浏览的帖子ID、浏览时间和帖子创建时间。用户浏览帖子的历史信息如表3.1所示：

表 3.1用户浏览帖子历史信息的举例

|  |  |  |  |
| --- | --- | --- | --- |
| 用户ID | 浏览的帖子ID | 浏览时间 | 帖子创建时间 |
| 96436391389912947 | 514655 | 2019-01-07 23:12:12 | 20181126 |
| 96436391948879316 | 514673 | 2018-12-26 06:23:33 | 20181126 |
| 600608457526745925 | 514673 | 2018-12-26 00:00:26 | 20181126 |
| 600608457554287653 | 514673 | 2018-12-26 03:32:35 | 20181126 |
| 600608457600401852 | 514673 | 2018-12-26 00:46:28 | 20181126 |

通过对用户浏览帖子的历史信息，能够构建一个用户字典。对于没有在论坛中产生行为的用户，我们把他们标记为未登录用户，推荐这么部分用户的推荐需要考虑冷启动问题。

通过对两个月（12月26日到2月25日）的用户浏览帖子的信息，我们发现，在这两个月中有551990个用户（注册用户数为5000000个）最近两个月在论坛中浏览过帖子，只占总体用户的10%。这其中的原因有：

（1）用户注册平台账户后就没有使用过APP，所以就不可能会使用论坛服务；

（2）有一部分用户，很少使用论坛服务，由于论坛不是APP的主要功能，该APP的主要功能是司机线上找货；

（3）有一部分用户使用论坛的间隔较长，要远远大于两个月。

通过对近两个月用户浏览帖子的历史记录，可得出以下结论：

（1）用户浏览的帖子中近期发的占比较大，其中近7天的占88%，近30天99%。

（2）用户浏览的帖子栏目满足资讯、曝光、卡嫂、卡友心路、买车卖车的帖子占绝大部分，约为90%。

（3）特殊栏目，虽然话题较少，但是帖子质量较高，用户浏览量较大，如卡嫂栏目，每天发帖量不到10条，但是浏览量非常大。

（4）超过一半的用户（约占比54%）每天只会浏览一篇帖子。

（5）每天浏览帖子数大于等于3的用户占比为22.7%

帖子的信息包括帖子ID、帖子内容、板块ID、创建时间、浏览数、回复数和点赞数。帖子ID用户标识每个帖子；帖子内容是呈现给用户内容，是以文字主，其中也会包含html标签和表情符号等其他非文字的数据；板块ID用于标识每个帖子所属板块；创建时间是帖子发布时间的。具体数据示例如表3.2所示：

表 3.2帖子详细信息的举例

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 帖子ID | 帖子内容 | 板块ID | 创建时间 | 浏览数 | 回复数 | 点赞数 |
| 514655 | 中国三千万驾驶员… | 37 | 2018-11-26 03:46:50 | 135 | 0 | 2 |
| 514673 | 现在到处抓4米2… | 35 | 2018-11-26 06:47:54 | 14661 | 102 | 22 |
| 389429 | 现在在哪到电子签单… | 46 | 2018-05-31 15:26:11 | 2 | 0 | 0 |
| 47507 | 端午节到了，我在酝酿… | 35 | 2016-06-08 17:55:44 | 882 | 66 | 51 |
| 550381 | 荆州这个地方再也不… | 35 | 2019-02-10 19:52:43 | 364 | 2 | 2 |

我们对论坛中所有帖子（2019年2月25日的之前创建的帖子总数为297115）进行分析，总结帖子信息的以下特点：

（1）帖子长短不同：论坛对发布的帖子格式要求不严，用户可用按照每个人不同的风格创建的帖子，尤其对字数没有限制。有些帖子只有几个字；有些帖子长达数百个字 。帖子文本处理需要考虑帖子这个特性。

（2）帖子创建者不同：每个用户都可以创建帖子，所以帖子的作者可以是每个司机；另外，作为平台，会每天发布一些优质的新闻资讯。

（3）口语化：由于论坛主要面向的是从事公路物流运输的司机用户，这些用户倾向于使用口语化来表达自己的想法。

（4）帖子中存在方言：由于平台注册的司机来自于全国34省市，地区不同造成语言差异，不同用户在创建帖子是会夹杂着方言。

## 3.2 系统总体设计

相对其他推荐系统，我们设计一套详情页推荐框架。把详情页的推荐划分为两个阶段，一个是详情页推荐模型构建阶段，二是使用详情页推荐模型进行相关帖子推荐阶段。如图3.2所示，展示的是详情页推荐的框架流程图，其中详情页推荐模型构建阶段主要包含四个步骤，分别：



图3.2详情页推荐问题的框架

（1）对用户历史记录进行处理。由于用户历史记录数据可能存在不完整信息或者需要对部分字段进行过滤、格式转换处理。需要对样本数据进行训练集和测试集划分。

（2）帖子数据预处理阶段，由于帖子的内容作为帖子一个非常重要特征。需要使用自然语言处理对文本数据处理的流程处理这部分数据进行处理。这部分我们将在3.3节进行详细介绍。

（3）特征构造阶段，由于不同模型的输入的特征，需要构造的特征也不同。特征构造要根据模型的设计来进行构造，我们把特征的构造放在模型设计里来介绍。

（4）模型训练阶段，为了实现详情页的个性化推荐，我们使用两种不同方法对用户历史行为进行处理，提出了两种的推荐方法：

第一种模型的设计，我们把详情页推荐定义为一个推荐问题，根据用户的历史行为记录分析用户兴趣来给用户推荐感兴趣的帖子。所以在这种方法中，我们把当前浏览帖子作为用户的历史行为记录来进行建模的。该方法将会在3.4节进行详细的介绍。

第二种模型的设计，我们从一个全新角度来定义详情页推荐这个问题。把整个推荐系统看成一个搜索引擎，当用户点击浏览某个帖子是，可以把这个帖子看成用户输入的“关键词”，不过这个关键词是一个文档，然后再搜索引擎推荐输出用户查询的结果。我们目标是实现个性化查询，其实在这个输入过程中还有一个隐变量要被输入，这个隐变量就是用户上下文信息。该方法将在3.5节进行具体介绍。

（5）模型评估阶段。本文使用推荐系统最关心两个指标模型评估。一个是推荐贴的点击量，来衡量帖子的推荐的准确性；另一个指标是帖子的推荐量，衡量的是帖子推荐的个性化程度，这个指标稍微有点抽象，举个例子说，我们的推荐系统不同总推荐热门帖，热门帖子点击率是很大，但是这不满足推荐另一个作用，推荐系统作为挖掘长尾的技术手段，让每个帖子都能够有得到展示机会。

使用详情页推荐模型进行帖子推荐阶段主要包含五个步骤，分别是：

（1）用户点击浏览一个帖子。

（2）根据帖子预处理方式获取帖子处理后的特征。

（3）根据处理后帖子特征和用户特征进行特征构造。

（4）将构造的特征输入到推荐系统模型计算法相似度计算。

（5）输出相似度较高的5个帖子。

## 3.3 帖子内容预处理

为了能更好的利用帖子内容信息进行相似性推荐，我们要把帖子处理成可以进行计算的表达形式。为了能够充分利用帖子的内容，我们需要对帖子的内容进行替换、删减的操作。本节将详细帖子的预处理方法，分别包括帖子的语料提取，帖子中文分词、构造公路物流运输垂直领域的停用词库、对帖子进行停用词去除、以及对帖子进行词向量化的五个步骤，各个步骤的具体方法将在各个小节进行描述。

### 3.3.1 帖子特征提取

由于帖子通常都是表达一个观点，且帖子以短文本为主，帖子中每个文字，每个符号都是那么的重要，为了充分利用帖子内容的信息。我们对帖子进行的特征的提取。

通过分析我们发现，帖子中存在表情符号，对于自然语言处理中，通常是把表情符号进行过滤处理；我们希望能更加充分应用。因为这表情符号，能够很明确帖子中的情感。所以我们对帖子中表情符号进行的转义处理。

### 3.3.2 帖子中文分词

不同于英文，中文以字作为书写单位，各个单词之前没有明确的分割标记，在进行中文文本处理时，需要首先进行词法分析。中文分词将待处理的中文文本作为输入，使用自然语言处理技术进行成词判断、词语分割、词性判断等处理，最终输出切分好的中文单词、外文单词、阿拉伯数字等一系列字符串。

作为帖子预处理的重要步骤，中文分词的效果直接影响帖子训练数据的质量。目前常用的中文分词工具主要有Jieba、ICTCLAS、SCWS、LTP、NLPIR等，他们使用不同的分词方法，但大部分都可以取得较高的准确度。

本文选取Jieba分词工具对从裁判文书中提取出的案件基本情况进行分词和词性标注。在中文分词后，本文通过词性筛选的方式只保留词性为名词、动词和形容词的词，同时去除了数字、符号和字母。

### 3.3.3 构建帖子停用词库

我们使用推荐数据是公路物流运输领域的数据，其中存在大量的公路物流运输专有名词，如卡友、平板车、一装一卸、福佑、十二米五等。这些词语不但对模型训练没有意义，反而会提高模型训练的复杂度，影响训练结果。因此，我们需要构造一个专有停用词库来去除这些法律专有停用词。如何帖子中的一个词语是否是停用词，是否对标识一份文档有意义，本文从词语的文档频率（Document Frequency, DF）、词语的类别频率（Category Frequency, CF）和类别信息熵（Category Information Entropy, CIE）这三个角度出发进行研究，其中文档频率DF指包含一个词语的文档集合在所有文档中所占的比例，类别频率CF值指在一个特定类别中，包含一个词语的文档集合在该类文档集合中所占的比例。判断是否该去除一个词可以分三种情况考虑：

（1）如果一个词语的文档频率DF极大，则应该将该词语去除。

（2）如果一个词语的文档频率DF较小，但是对于某一种类别的类别频率CF较大，意味着该词对标识某一种类别很有帮助，应该保留。

（3）如果一个词语的文档频率DF很小且对于每个类别中的类别频率CF也很小，意味着该词为低频词，应该去除。

进一步，对于情况（2），在考虑多种案由的情况下，判断某个词语对标识某种或某些类别的裁判文书是否有用，本文参考信息熵的思想，通过计算词语的类别信息熵来判断词语对类别信息的表现能力。如果一个词语的信息熵越大，意味着该词对表现类别的信息越不确定，则意味着应该去除。词语w的类别信息熵计算公式为：

其中表示词语w在类别中出现的次数，表示词语w在所有文书中出现的次数。

本文选取了上万帖子作为构造专有停用词语料，采用上述方法来计算每个词语的类别信息熵，再结合文档频率筛选出信息熵大于某一阈值且文档频率大于某一预设值的词语作为帖子专有停用词库。

### 3.3.4 去除帖子停用词

尽可能地去帖子中的停用词能极大剔除噪声数据，降低训练模型的复杂度，提高训练模型的准确度。针对不同类型的停用词，本文以数据流的方式按照以下顺序和方法去除裁判文书停用词。

（1）去除词语长度小于2的词语。

（2）采用匹配帖子专有停用词库的方式去除帖子专有停用词如卡友、平板车、一装一卸、福佑、十二米五等词语。

（3）采用文档频率统计的方法去除人名等其他低频噪声词汇。本文通过控制最小DF阈值来去除DF小于20的低频噪声词汇。

### 3.3.5 帖子向量化表示

对帖子进行向量化的处理是后面模型能否使用帖子的前提。因此，我们需要对帖子进行向量化处。通过2.3分析可知嵌入式的词向量要比词袋法要好。本文为了尽可能提高推荐准确性，我们使用了两种嵌入式词向量表达的方法，即Glove和Word2Vec。

## 3.4 基于概率分解的详情页推荐

详情页的推荐，要尽可能充分的利用用户和帖子的数据。由于对用户浏览帖子的历史数据来看，数据是一种稀疏的数据。这种稀疏性降低了传统协同过滤技术的评分的预测的精确度，基于文档建模的方法通过利用文本数据来提高准确性。Kim D等人提出了一种用CNN提取文档信息，然后将卷积神经网络（CNN）集成到概率矩阵分解（PMF）中，Kim D在他的论文证明了这个方法的有效性。由于Kim D，没有考虑把用户上下文信息集成到PMF，中，他只是做了一个假设，假设用户向量是一个满足零均值球面高斯分布。本节按照Kim D的算法，设计了一套用于论坛详情页推荐的算法。

### 3.4.1 具体框架

由于论坛详情页数据是稀疏的。我们计算用户和帖子之间的分数（即用户之间的相似性）放弃了传统协同过滤方法，采用了PMF方法。

为实现详情页的精准推荐，我们在使用RNN提取用户一段时间内浏览帖子的记录来反映用户兴趣变化向量，把这个向量作为用户向量。然后使用CNN提取文本向量，通过把RNN和CNN提取的用户向量和帖子向量集成到PMF中，最终生成用户和帖子之间的相似度。



图3.3 基于概率矩阵分解的推荐流程

该推荐方法的流程如图3.3所示，主要步骤包括：

（1）按照3.3节所述方法对帖子进行预处理，提取帖子的特征；

（2）构造用户特征：我们将使用用户浏览帖子的历史数据来构造用户的特征；

（3）构造帖子特征的：我们僵尸CNN来提取帖子的特征；

（4）输入用户兴征向量和帖子向量到概率矩阵分解模型里，计算相似度分数；

各个步骤中的具体细节将在下面小节中分别进行详细介绍，其中3.4.2节介绍用户特征提取，3.4.3节介绍使用CNN对帖子特征进行提取，3.4.4节介绍基于PMF计算相似分数集体细节。

### 3.4.2 构造用户特征

在基于概率矩阵分解的推荐算法种，我们在ConvMF的基础上，增加用户特征。我们使用用户在一段时间浏览的帖子作为用户的特征。用户特征构建过程如图3.5所示：



图3.4 用户行为序列

首先根据3.3节对帖子进行预处理，把帖子表示成一个文档向量。然后把用户当前点击的帖子作为放在用户行为序列最后端，再从用户浏览的帖子中，把用户最近浏览的帖子按照浏览的时间顺序排列在用户的行为序列中。把帖子的文档对应维度求和取平均，最终得到一个向量来作为用户的特征。当用户新登陆到论坛时，由于用户没有浏览过其它帖子，此时，直接用当前帖子的文档向量作为帖子的向量。

### 3.4.3 提取帖子特征

论坛详情页推荐的需要利用帖子内容这个特征。通过帖子的预处理生成了一个低维稠密的词向量。我们把帖子看成由多个词向量组成的矩阵。对帖子进行特征，我们使用CNN来构造文本向量。CNN在图像领域已经取得很好的效果，CNN能够有效的提取局部特征，并通过权值共享和下采样，能够减少计算复杂度。由于CNN在图像里取得很好的应用，很多学者尝试着将CNN用在自然语言领域中，如Yoon Kim将CNN应用在文本分类中，提出了Text CNN，并且取得很好的效果。本家我们将介绍使用CNN来提取帖子的特征，具体流程如下图所示：



图3.5 CNN模型提取帖子特征流程

具体步骤，如下所述：

（1）使用3.3节的方法对帖子进行预处理；

（2）然后把每个帖子转化成文档向量；

（3）使用CNN进行文本特征提取。

由于我们在3.3节已经对帖子的预处理进行了详细的介绍。接下来我们将对使用CNN来进行帖子向量提取的过程进行详细的介绍，我们设计CNN的结构，如图3.6所示：



图3.6 CNN模型结构

CNN网络具体细节如下所述：

（1）输入层：输入层输入的是文档向量，我们通过前面的步骤对文档进行的预处理，然后把帖子转化成一个稠密的文档矩阵。假设前面的步骤把词表示成d维的词向量W=[]，预处理后的帖子由n个词组成，然后文档举证可以表示成D=[]。

（2）卷积层：CNN中使用卷积层进行文本特征提取，通过设置卷积核的滑动窗口的大小来世小对词汇进行k-gram片段特征的获取。使用非线性激活函数，对卷积来激活。我们选在ReLU函数来作为激活的函数，这是因为ReLU函数可以避免梯度消失的问题。

（3）池化层：通过使用池化层，来选取有代表性的特征。我们使用了Max-Pooling方法来进行池化层的特征选择。

（4）输出层：我们最终获取去一个固定维度的文档向量，来表示这个文档，这个向量的维度要和用户兴趣向量的维度保持一致。

### 3.4.4 相似度计算

在用户向量和帖子向量都构造完毕之后了，我们就可以使用PMF来计算相似分数。相似分数的计算过程如图3.7所示：



图3.7 基于概率矩阵分解计算流程

计算相似分数步骤如下：

（１）构造用户兴趣向量；

（２）构造帖子向量，帖子向量使用CNN方法来进行构造；

（３）使用PMF来计算相似分数。

PMF具体计算过程如下：

假设用N个用户和M个帖子，我们计算的相似分数是一个的举证，计算公式：

其中是具有均值为，方差为的高斯正太分布的概率密度函数，是只是韩式当用户浏览了帖子值为1，否则为0。

## 3.5 基于用户行为的详情页推荐

基于用户行为的推荐设在3.4节的基于概率矩阵分解的基础上，引入RNN对用户的浏览帖子的历史数据进行处理。本文在Hidasid等人使用RNN来对用户行为序列进行建模的基础上。把这个方法应用相似贴推荐问题上。同时，在训练用户行为序列模型是，我们借鉴了Word2Vec模型中负采样的思想，来解决正负样本不均衡的问题。

### 3.5.1 具体框架

为了够好的利用用户历史的行为数据，我们在3.4节概率矩阵分解的基础上，重新对用户的行为特征进行设计，我们在使用GRU提取用户一段时间内浏览帖子的记录来反映用户兴趣变化向量，把这个向量作为用户向量。然后使用CNN提取文本向量，通过把GRU和CNN提取的用户向量和帖子向量集成到PMF中，最终生成用户和帖子之间的相似度。模型结构设计如图3.3所示：



图3.8 基于用户行为序列的推荐流程

该推荐方法的流程如图3.10所示，主要步骤包括：

（1）按照3.3节所述方法对帖子进行预处理，提取帖子的特征；

（2）构造用户特征： 我们将使用GRU对用户浏览帖子的历史纪录进行行为序列建模；

（3）构造帖子特征的：使用CNN来提取帖子的特征。

### 3.5.2 构造用户特征

构造用户特征需要对可使用的数据进行正负样本的设计和模型的设计，具体过程如图3.9所示：



图3.9 用户特征构造的流程

用户特征的构造线下阶段主要是根据用户历史数据训练出一个用户行为寻猎模型，具体步骤如下：

（1）训练集设计：需要对正负样本进行设计。

（2）模型设计：对比了RNN、LSTM和GRU，由于GRU既没有RNN梯度弥漫的担忧，又没有LSTM的训练慢的特点。所以我们最终选择了GRU。

（3）训练：使用构造好的正负样本和模型进行训练

（4）模型：把训练的模型进行持久化，供线上预测使用。

线上部分是根据用户的最近浏览的记录通过加载线下训练好的模型，要预测用户的行为特征，具体步骤如下：

（1）用户最近浏览的记录的构造：当用户点就某个帖子时，我们用户最近一段时间的点的帖子和当前帖子组合成一个序列。

（2）预测输出用户行为特征：加载线下训练好的模型，把构造的序列输入进行预测，输出用户序列特征。

对于样本的构造，我们根据用户浏览帖子的数，把用户浏览帖子数大于等于3的作为可用数据，替他用户浏览记录将被踢出。我们对用户浏览的帖子数作如下讨论：

（1）浏览帖子数等于3个，如图3.10所示：



图3.10 用户行为序列（一）

按浏览的时间顺序把，把帖子排序。把将前两个作为输入，把最后一个作为标签。即，{(A, B), C}

（2）帖子数大于3个，如图3.11和3.12所示：



图3.11 用户行为序列（二）



图3.12 用户行为序列（三）

首先把前两个帖子A，B作为输入，把C作为标签。即{(A, B), C}。然后通过滑动窗口，把前三个帖子A，B，C作为输入，把D作为标签。即{(A, B, C), D}。然后使用同样的方式进行窗口化，可以设置一个阈值，当窗口大于一个阈值时，就不再增加窗口的大小。

对于模型的设计，首先需要考虑的是模型的输入数据的类型，这里有两个输入方式：

（1）把帖子进行One Hot编码构造一个高维的稀疏的向量，对于高维稀疏的向量，需要在模型输入后亦曾加入一个Embedding层，用来把高维稀疏的向量映射为低维稠密的向量，方便后续的处理。（2）通过把词向量化后的帖子作为输入，此时的输入是一个低维稠密的向量，而且这个向量代表着帖子的内容。

综合考虑输入格式，由于我们在使用CNN提取帖子是，需要对帖子进行词向量化表示，所以使用第二种方式，即不需要带来额外的计算，又能降低用户序列模型的复杂程度，所以，我们使用第二种方式进行输入。

模型结构如图3.13所示：



图3.13 GRU模型的结构

### 3.5.3 提取帖子特征

基于用户行为序列的详情页推荐是在基于概率矩阵分解的基础上的进行改进的。我们改进用户的行为特征，帖子的特征与基于概率矩阵分解模型时一样的。使用CNN来对帖子特征进行提取。

### 3.5.4 相似度计算

通过3.5.2和3.5.3两节的方法获取的用户行为特征和帖子特征。计算相似度分数的过程和3.4.4节计算过程是一样的，只是在用户行为特征处理会有区别。3.4.4节是通过对用户最近浏览的帖子的向量对应维度进行求和取平均的方法，来构造用户行为特征。本小节，使用的是GRU模型来构造的用户行为特征。

## 3.6 本章小结

本章将会介绍了详情页推荐的方法。3.1节介绍了详情页推荐的问题的分析，分析论坛推荐的两个问题，对详情页推荐问题进行定义；分析了详情页推荐中帖子的特点，以及用于推荐数据的特点；3.2节介绍了详情页推荐的框架，设计从数据处理到推荐的流程；3.3节介绍了帖子数据的处理，用户特征的构造；3.4节将会介绍了基于概率矩阵分解的详情页推荐算法；3.5节将会介绍了基于用户行为序列建模的推荐算法。

# 第四章 基于用户历史行为的详情页个性化推荐方法的实现

本章主要描述基于热门帖的推荐、基于文档相似度的推荐、基于概率矩阵分解的推荐和基于用户行为序列的推荐的四种算法实验过程与实验结果。4.1节将介绍详情页推荐实验所用的数据集；4.2节将介绍详情页推荐结果的评估标准和评估方式；4.3节介绍基于概率矩阵分解的推荐算法的实现细节；4.4节将介绍基于用户行为序列的推荐算法的实现细节；4.5节将介绍四组实验和实验结果；4.6节将会对四种推荐算法的实验结果进行对比；4.7节对本章进行了总结。

## 4.1 数据集

本文实验使用的数据是公路物流运输行业领域的论坛的真实数据。可用的数据包括论坛帖子的信息和论坛司机浏览的帖子的历史记录。数据格式和内容，我们在3.1.2节对数据分析时进行了介绍。本节将从实验设置来对数据进行介绍。

本次实验我们选取的用户浏览帖子的历史记录时间跨度为两个月（2018年12月26日到2019年2月25日），数据总共有7525969条记录，其中有551990个用户，在这两个月浏览过帖子。实验中，我们把2018年12月26日到2019年2月21日作为训练集，把2019年2月22日到2月25日作为测试集。

本次实验使用的帖子数据为论坛的所有的帖，帖子的总数为297115条。

## 4.2 评估指标与评估方式

推荐系统常用的评估指标有：准确度、满意度、覆盖率、多样性、新颖性、惊喜度、信任度、实时性、鲁棒性、可扩展性、商业目标、用户留存等指标[项亮等，2012]。

推荐系统评估方式分为：问卷调查、离线评估和在线评估。问卷调查，制定相关问卷调查表耗时耗力，调查活动开展又是非常麻烦，很难找到适合的调查对象，因此成本非常高，不被采用。离线评估可以在不损害用户体验的情况下，进行快速的评估。但是离线评估只能在用户看到过的候选集上做评估，且和线上真实效果存在偏差，同时也只能评估少数指标。在线测试，常用方式AB Testing，通过划分不同策略，把线上的用户分为不同实验组。离线评估和在线评估各有利弊，在实践中可以采取离线评估和在线评估相结合的方式进行评估方式。

本文主要对推荐算法的准确性和覆盖率两个指标进行考察。下面将对这两个指标进行具体的介绍。

准确度度量的是推荐算法或者推荐系统对用户行为的预测能力，是推荐系统最重要的离线评估指标。 该指标的计算需要离线数据集包含用户的历史行为记录信息。通过将离线的数据从时间维度上划分为训练机和测试集，然后使用在训练集上训练的模型对测试集上的用户行为进行预测。最后，计算模型在测试集上预测用户行为和测试集上用户真实行为的准确度。由于不同的推荐任务对准确度的定义不一样。我们根据详情页推荐任务，重新定义了这个指标。

其中，N为测试集当天用户点击帖子的记录数。

覆盖率度量的是推荐算法或者推荐系统对帖子挖掘长尾的能力。同样不同的任务，会对应不同的定义。我们把覆盖率定义为：在训练集上训练的模型，在测试集上进行预测是推荐的不同的帖子数占真个帖子的总数。

## 4.3 基于概率矩阵分解的详情页推荐的实现

|  |
| --- |
| def PMF(res\_dir, train\_user, train\_item, valid\_user, test\_user,  R, CNN\_X, vocab\_size, init\_W=None, give\_item\_weight=True,  max\_iter=50, lambda\_u=1, lambda\_v=100, dimension=50,  dropout\_rate=0.2, emb\_dim=200, max\_len=300, num\_kernel\_per\_ws=100):  for iteration in xrange(max\_iter):  loss = 0  tic = time.time()  print "%d iteration\t(patience: %d)" % (iteration, count)  VV = b \* (V.T.dot(V)) + lambda\_u \* np.eye(dimension)  sub\_loss = np.zeros(num\_user)  for i in xrange(num\_user):  idx\_item = train\_user[0][i]  V\_i = V[idx\_item]  R\_i = Train\_R\_I[i]  A = VV + (a - b) \* (V\_i.T.dot(V\_i))  B = (a \* V\_i \* (np.tile(R\_i, (dimension, 1)).T)).sum(0)  U[i] = np.linalg.solve(A, B)  sub\_loss[i] = -0.5 \* lambda\_u \* np.dot(U[i], U[i])  loss = loss + np.sum(sub\_loss)  sub\_loss = np.zeros(num\_item)  UU = b \* (U.T.dot(U))  for j in xrange(num\_item):  idx\_user = train\_item[0][j]  U\_j = U[idx\_user]  R\_j = Train\_R\_J[j]  tmp\_A = UU + (a - b) \* (U\_j.T.dot(U\_j))  A = tmp\_A + lambda\_v \* item\_weight[j] \* np.eye(dimension)  B = (a \* U\_j \* (np.tile(R\_j, (dimension, 1)).T)  ).sum(0) + lambda\_v \* item\_weight[j] \* theta[j]  V[j] = np.linalg.solve(A, B)  sub\_loss[j] = -0.5 \* np.square(R\_j \* a).sum()  sub\_loss[j] = sub\_loss[j] + a \* np.sum((U\_j.dot(V[j])) \* R\_j)  sub\_loss[j] = sub\_loss[j] - 0.5 \* np.dot(V[j].dot(tmp\_A), V[j])  loss = loss + np.sum(sub\_loss)  seed = np.random.randint(100000)  history = cnn\_module.train(CNN\_X, V, item\_weight, seed)  theta = cnn\_module.get\_projection\_layer(CNN\_X)  cnn\_loss = history.history['loss'][-1]  loss = loss - 0.5 \* lambda\_v \* cnn\_loss \* num\_item  tr\_eval = eval\_RMSE(Train\_R\_I, U, V, train\_user[0])  val\_eval = eval\_RMSE(Valid\_R, U, V, valid\_user[0])  te\_eval = eval\_RMSE(Test\_R, U, V, test\_user[0])  toc = time.time()  elapsed = toc - tic |

图4.1 ConvMF模型代码

如图4.1所示，是概率矩阵分解的实现的代码。本段代码描述的是

## 4.4 基于用户行为序列建模的详情页推荐的实现

|  |
| --- |
| Class GRUforRec:  # 初始化模型参数  def \_\_init\_\_(self, loss, final\_act, …):  # 设置损失函数  def set\_loss\_function(self, loss):  # 设置激活函数  def set\_final\_activation(self, final\_act):  # 设置隐藏层激活函数  def set\_hidden\_activation(self, hidden\_act):  # 设置模型参数  def set\_params(self, \*\*kavrgs):  # 初始化权重  def init\_weights(self, shape, name=None):  # 设置随机失活  def drouput(self, X, drop\_p):  # 定义模型  def model(self, X, H, R=None, Y=None, drop\_p\_hidden=0.0, drop\_p\_embed=0.0, predict=False):  …  for i in range(start, len(self.layers)):  vec = T.dot(y, self.Wx[i]) + self.Bh[i]  rz = T.nnet.sigmoid(vec[:,self.layers[i]:] + T.dot(H[i], self.Wrz[i]))  h = self.hidden\_activation(T.dot(H[i] \* rz[:,:self.layers[i]], self.Wh[i]) + vec[:,:self.layers[i]])  z = rz[:,self.layers[i]:]  h = (1.0-z)\*H[i] + z\*h  h = self.dropout(h, drop\_p\_hidden)  y = h  H\_new.append(T.switch(R.dimshuffle((0, 'x')), 0, h) if not predict else h)  …  # 模型训练  def fit(self, data, sample\_store=1000000):  # 模型预测  def predict() |

图4.2 GRUforRes代码

如图4.2所示是基于用户行为序列建模的推荐方法的实现代码。

## 4.5 对比实验

本小节，设计了几组实验来验证本文提出来的基于用户行为序列的详情页推荐的算法的效果。总共设置四组实验：基于热门帖的推荐、基于文档形似度的推荐、基于概率矩阵分解的推荐和基于用户历史行为的推荐。前两种为简单的推荐策略，本文在第三章并没有介绍这两种策略的设计，后两种是本文研究的重点，比前两种要为复杂，我们在第三章的3.4和3.5两节对这两种算法的设计做了具体的介绍。两一方面，前两种方法没有使用用户浏览帖子的历史信息，而后两种则使用了用户浏览帖子的历史特征这一信息。在使用用户浏览帖子的历史信息，基于概率矩阵分解和基于用户行为序列的方法处理这一数据的方式有很大区别。基于概率矩阵分解的方法是通过简单的把用户浏览贴的向量进行对应维度相加取平均，为基于用户行为序列的方法通过使用GRU对用户浏览帖子的进行建模处理。四种模型的比较如表4.1所示：

表 4.1四种模型的比较

|  |  |  |
| --- | --- | --- |
| 使用信息  推荐方法 | 是否用到文档信息 | 是否用到用户浏览的历史 |
| 基于热门帖的推荐 | 没有 | 没有 |
| 基于文档相似度的推荐 | 有 | 没有 |
| 基于概率矩阵分解的推荐 | 有 | 有 |
| 基于用户行为序列的推荐 | 有 | 有 |

下面将对这四种方法实现和实验结果进行具体的介绍：

### 4.5.1 基于热门帖子的推荐

基于热门帖的推荐是把最热门的帖子推荐给用户，是最简单且有效的推荐方式。衡量一个帖子的热门程度有三个指标，帖子的浏览量，帖子的评论量，帖子的点赞量。

其中，，，为各个指标的权重指数；View\_Count、Replay\_Count、Praise\_Count分别为浏览量、评论量和点赞量。

热门帖的推荐常会出现哈利波特效应（即，不管用户浏览什么帖子，都给用推荐最热门的帖子），为了解决哈利波特效应，我们可以增加用户是否浏览过的帖和发布的时间帖子时间两个指标：

通过设置一个is\_View，如果用户已经浏览过，is\_View值为1，Score为0，热门度最低，达到不推荐的目的。分母是对帖子的发布时时间做一个惩罚，T为指定某个时刻，如果只推荐近一个月的帖子，可以把T设置为一个月前的第一天的时间。

图4.3 基于热门帖的推荐的帖子占比

如图4.1所示是基于热门帖的推荐在2月22日到25日第四天的推荐的帖子的数占比。推荐帖子的占比等于每天的推荐的不同帖子的数除于总体帖子的值，这个指标评价了算法的挖掘长尾的能力。从图中可以看出基于热门的推荐的方法，推荐的帖子的占比在1.30%以下。这四天的均值是1.09%。基于热门帖的推荐的只会推荐热门帖，这会导致越推越窄。

图4.4 基于热门的帖的推荐的点击率

如图4.1所示是基于热门的帖的推荐的算法的推荐帖子在2月22日到25日第四天的点击率。点击率等于被点击的帖子除于推荐的帖子，这个指标的衡量的是推荐的正确性。从图可以看出基于热门帖的推荐算法，每天推荐的帖子点击率都在1.28以上，这四天的平均值为2.04%。

### 4.5.2 基于文档相似度的推荐

基于文档相似度的方法，是通过把文档表示成一个向量，然后通过对每个帖子进行余弦值计算：

我们通过对词进行向量化表示，对帖子预处理去除了停用词。然后把每个词向量对应维度相加取平均来表示一个文档向量：

离线对每个帖子进行文档相似的的计算，把计算出的文档相似度，保存在能够高效的访问的Redis中。这种方式会存在一个风险，就是线上预测时，出现了一个新发布的帖子的。这是需要实时对这个帖子进行文本预处理，表示成文档向量，然后和离线库中的每个文档进行文档相似度的计算。

在做推荐的时，假设帖子A与帖子B最为相似。当用户点击帖子的A，会给用户推荐５个相关的帖子，这其中包含B；但是当用户点击帖子B时，由于帖子A和帖子B最为相似，所以也为把帖子推荐给用户。为了解决这个问题，我们对计算相似度公式进行了重新设计：

其中isView表示用户是否浏览，已经浏览表示为1，为浏览表示为0。

这样推荐的时需要根据每个用户浏览过的帖子实时计算Score分数。为了减少计算，我们需要离线存储用户浏览帖子的记录。

图4.5 基于文档相似度的推荐的帖子占比

如图4.1所示是基于文档相似度的推荐在2月22日到25日第四天的推荐的帖子的数占比。从图中可以看出基于文档相似度的推荐的方法，推荐的帖子的占比在3.76%以上。这四天的均值是4.56%。由于，我们设置不推荐用户已经浏览的帖子，这种推荐方法相对于基于热门帖的推荐一定程度上能减少越推越窄现象。

图4.6 基于文档相似度的推荐的点击率

如图4.1所示是基于文档相似度的推荐的算法的推荐帖子在2月22日到25日第四天的点击率。从图可以看出基于文档相似度的推荐算法，每天推荐的帖子点击率都在0.97%以下，这四天的平均值为0.67%。这种推荐的算法的点击并不高，因为只用文档的信息，没能很好的利用用户的信息。

### 4.5.3 基于概率矩阵分解的推荐

基于矩阵分解的推荐方法，使用CNN来提取文档特征，这种方式要比直接把文档的词向量对应维度相加取均值更能有效的提取文档的特征。

首先对所有的文本进行预处理，然后把文表示成一个文矩阵。使用3.4节CNN模型把文档处理成一个文档向量。对所有文档进行同样的处理，把文档向量存储在可以快速访问的数据库Redis中。

概率矩阵分解在构造用户向量是通过，把用户在近期已经浏览的帖子的文档向量对应维度相加取平均。这时候需要根据用户浏览帖子的历史纪录中取出用户最近浏览的帖子ID，通过帖子ID从Redis中读取对应的文档向量。计算用户向量：

其中为用户浏览序列中第个文档向量，k表示用户浏览序列的长度。表示的是用户当前点击的文档，即我们把用户当前点击的文档作为用户浏览序列中的最后一个文档。

图4.7 基于概率矩阵分解的推荐的帖子占比

如图4.1所示是基于概率矩阵分解的推荐在2月22日到25日第四天的推荐的帖子的数占比。从图中可以看出基于概率矩阵分解的推荐的方法，推荐的帖子的占比在4.27%以上。这四天的均值是4.65%。这种方法利用用户的历史信息，一定程度增加推荐的个性化。所以能够一定程度上能减少越推越窄现象。

图4.8 基于概率矩阵分解的推荐的点击率

如图4.1所示是基于概率矩阵分解的推荐的算法的推荐帖子在2月22日到25日第四天的点击率。从图可以看出基于概率矩阵分解的推荐算法，每天推荐的帖子点击率都在1.43%以上，这四天的平均值为2.07%。这种推荐的算法的点击要比推荐热门帖的稍微高一点。因为这种算法能够利用了文档的信息和用户的信息。通过给每个用户推荐感兴趣的帖子，一定程度提高推荐的点击率，但是提升的并不是很明显。

### 4.5.4 基于用户行为序列的推荐

基于用户历史行为的推荐的方法是在基于矩阵分解中对用户行为序列进行了改进。使用了同样的方法对文档进行处理，然后把文档向量存储在可以快速访问的数据库Redis中。基于概率矩阵分解的方法通过对用户的浏览帖子记录进行行为序列建模，使用3.5节的方法，能够得到一个反映用户动态兴趣的向量。文档向量可以表示为：

其中GRU表示的3.5设计的用户行为序列的模型，表示的时用户行为序列中的第文档，用户行为序列的长度为。

图4.9 基于用户行为序列的推荐的帖子占比

如图4.1所示是基于概率矩阵分解的推荐在2月22日到25日第四天的推荐的帖子的数占比。从图中可以看出基于概率矩阵分解的推荐的方法，推荐的帖子的占比在5.16%以上。这四天的均值是5.61%。这种方法利用用户的历史信息，一定程度增加推荐的个性化。所以能够一定程度上能减少越推越窄现象。

图4.10 基于用户行为序列的推荐的点击率

如图4.1所示是基于概率矩阵分解的推荐的算法的推荐帖子在2月22日到25日第四天的点击率。从图可以看出基于概率矩阵分解的推荐算法，每天推荐的帖子点击率都在3.67%以上，这四天的平均值为4.39%。这种推荐的算法的点击要比推荐热门帖和基于概率矩阵分解算法的都要高，超过两倍多。因为这种算法能够利用了文档的信息和用户的信息，并且通过对用户的行为进行动态建模，能够抓住用户的动态兴趣变化。通过给每个用户推荐感兴趣的帖子，大大提高推荐的点击率。

## 4.6 方法的优缺点分析

本节把四种算法推荐的准确性和推荐不同帖子的量进行汇总对比，并对算法在推荐准确性和挖掘长尾能力上进行对比分析。

如图4.1所示是四种推荐的算法的推荐的帖子数占比的对比，这个指标衡量的是推荐算法挖掘长尾的能力。从图4.1中可以看出：

图4.11 四种模型的推荐的帖子数占比的对比

（1）基于热门帖的推荐算法推荐的不同帖子的数要远远小于其他三种算法。这是由基于热门帖的推荐，只会推荐热门帖，推荐热门帖会导致推荐的热门帖被点击的概率更大，其他帖子没有展示的机会。因此导致越推越窄。

（2）其他三种算法推荐不同帖子的能力较基于热门帖的推荐要高很多。这是引文这三种算法不同程度上的利用了文档和用户历史行为信息。

（3）在这四种算法，基于用户行为序列的推荐算法要比其他三种推荐算法推荐不同帖子的能力要搞。这主要是这种算法能够很好的利用用户历史信息。通过对用户行为进行动态建模，能够根据用户的动态变化兴趣去推荐给用户感兴趣的帖子，提高了推荐的个性化。

如图4.1所示是四种推荐的算法的推荐的帖子点击量，这个指标衡量推荐算法准确性。从图4.1中可以看出：

图4.12 四种模型的点击率对比

（1）基于文档相似性的推荐算法推荐的帖子的点击率要远远小于其他三种算法。虽然基于文档相似性的推荐算法使用了文档这一信息，但是这种推荐没有使用用户的信息。所以在进行推荐时，给所有用户推荐的都是一样的，不是一种个性化的推荐。

（2）其他三种算法推荐不同帖子的能力较基于热门帖的推荐要高很多。基于热门帖的推荐尽管没有利用用户信息和文档信息，但是这种算法推荐的帖子都是用户最爱看的。基于概率矩阵分解和基于用户行为序列的推荐是因为通过利用用户信息和文档信息，这样既能满足个性化的推荐，又能满足相关性的推荐。

（3）在这四种算法，基于用户行为序列的推荐算法要比其他三种推荐算法推荐不同帖子的能力要搞。这是因为这种算法能够很好的利用用户历史信息。通过对用户行为进行动态建模，能够根据用户的动态变化兴趣去推荐给用户感兴趣的帖子，提高了推荐的个性化。通过利用文档信息，提高了推荐的相关性。因此推荐的帖子的准确性比较。

## 4.7 本章小结

本章主要描述基于热门帖的推荐、基于文档相似度的推荐、基于概率矩阵分解的推荐和基于用户行为序列的推荐的四种算法实验过程与实验结果。4.1节介绍了详情页推荐实验所用的数据集；4.2节介绍了详情页推荐结果的评估标准和评估方式；4.3节介绍了基于概率矩阵分解的推荐算法的实现细节；4.4节介绍了基于用户行为序列的推荐算法的实现细节；4.5节介绍了四组实验和实验结果；4.6节对四种推荐算法的实验结果进行了对比。

# 第五章 基于用户行为序列的论坛推荐系统的测试

## 5.1 总结

本文首先分析了详情页推荐问题的本质，概括了详情页推荐问题的特点，描述了详情页推荐问题可以解决的问题以及面临的困难。其次本文提出了帖子的预处理方法。本文首先在ConvMF推荐算法的基础上引入用户的特征，提出了一种基于概率矩阵分解的详情页个性化推荐方法；然后再基于概率矩阵分解的基础上，使用GRU对用行为进行序列建模，提出了基于用户行为序列建模的详情页个性化推荐方法。探究个性化推荐算法在公路物流这个比较垂直领域的论坛详情页个性化推荐的实践方法；提出一种能够同时融合用户历史行为信息和文档信息的个性化推荐方法；提出了使用用户行为序列建模在个性化推荐算法。

本文提出了两种不同方法用于详情页的推荐。基于推荐技术的详情页个性化推荐方法：由于详情页个性化推荐本质衡量两个帖子在用户这一变量限制下相似度问题。该方法在概率矩阵分解中引入帖子内容特征和用户兴趣特征，将用户在一段时间浏览的帖子看作用户一个动态兴趣的反应进行用户兴趣特征，然后通过使用概率矩阵分解方法来计算用户兴趣和帖子相似度进行详情页的个性化推荐。基于搜索技术的详情页个性化推荐方法：从另一个角度来看这个问题，详情页的个性化推荐可以看成在已知用户历史行为和当且帖子信息来搜索相关帖子的搜索匹配问题。本文通过构造用户兴趣和当前帖子的两个关键词来搜索相关帖子。然后把搜索到的相关性比较靠前的帖子推荐给用户。

在实验评估阶段，本文从两个方面对详情页推荐方法进行了评估，包括准确性和覆盖率。本文使用公路物流运输领域的论坛数据进行实验，设置四组实验进行对比。然后对实验的结果进行分析，分析四种实验的优缺点。验证了基于用户行为序列建模的详情页推荐算法的有效性。

## 5.2 进一步工作展望

本文提出的详情页个性化推荐方法还存在一系列缺陷和不足有待改进，具体如下：

（1）用户上下应用不充分：本文使用了用户浏览帖子作为用户兴趣，用户的兴趣是非常复杂，变化的也是非常复杂，需要使用更多用户上下文信息来对用户兴趣来建模。

（2）训练数据和线上有差别：推荐系统作为一种解决长尾效应的方法，能够给更多帖子的展示的机会。但是我们使用用户已经浏览的帖子来优化算法，由于帖子展现机会是非常少的，没有展示的帖子并不能代表用户和兴趣不相关。在训练模型的时候我们没有对这数据进行很好的应用。

（3）忽略展示位对点击率的影响：通常来说对于每个展示位，用户点击它们机会是不同，所以把推荐不同结果的放在不同位置，也会对推荐帖子的点击产生影响，我们将在未来的工作，对这部数据进行收集，并把Position数据作为特征用来推荐的优化。

新闻推荐、视频推荐和的购物推荐是推荐系统三个重要应用方向。随着近些年来深度学习在图像领域、自然语言处理领域和声音等领域取发展，推荐系统一个新的发展方向就引入深度学习。YouTube已经把深度学习很好的应用视频推荐，带来非常好的效果。所深度学习时推荐系统一个重要的发展方向。推荐系统的另一个发展方向就是引入强化学习，强化学习根据环境对每次行为进行反馈，这种自适应性很适合推荐领域。阿里巴巴在购物推荐系统中引入强化学习，取得非常好的效果。还有就基于Graph的推荐系统也是推荐系统发展的一个新的方向。

# 参 考 文 献

[Bogina et al, 2017] V. Bogina, and T. Kuflik, Incorporating Dwell Time in Session-Based Recommendations with Recurrent Neural Networks, RecTemp@ RecSys, pages 57-59, 2017.

[Carlson et al, 2015] J. L. Carlson, *Redis IN ACTION*, New York:Manning Publications Co. Press, 2013

[Cho et al, 2014] K. Cho, B. v. Berrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, arXiv:1406.1078 / EMNLP 2014.

[Hidasi et al, 2015] B. Hidasi, A. Karatzoglou, and L. Baltrunas, Session-based Recommendations with Recurrent Neural Networks, Computer Science, 2015.

[Hidasi et al, 2016] B. Hidasi, M. Quadrana, and A. Karatzoglou, Parallel Recurrent Neural Network Architectures for Feature-rich Session-based Recommendations, *In Proceedings of the 10th ACM Conference on Recommender Systems* (RecSys 2016). ACM, 2016.

[Hochreiter et al, 1997 ] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[Kim et al, 2014] Y. Kim, Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.

[Kim et al, 2016] D. Kim, C. Park, J. Oh, S. Lee, and H. Yu, Convolutional matrix factorization for document context-aware recommendation, *In Proceedings of the 10th ACM Conference on Recommender Systems*, ACM’16, pages 233-240, 2016.

[Krizhevsky et al, 2012] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[Linden et al, 2003] Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. IEEE Internet computing, 2003 (1): 76-80.

[Mnih et al, 2007] R. Salakhutdinov, Probabilistic matrix factorization, International Conference on Neural Information Processing Systems. Curran Associates Inc. 2007.

[Quadrana et al, 2017] M. Quadrana, A. Karatzoglou, and B. Hidasi, Personalizing Session-based Recommendations with Hierarchical Recurrent Neural Networks[J]. 2017.

[Rumelhart et al, 1986] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning interal representations by back-propagating errors, Nature, 323, pages:533-536

[Shani et al, 2002]

[Wang et al, 2011] C. Wang and D. M. Blei, Collaborative topic modeling for recommending scientific articles, *In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD’11, pages 448-456, 2011.

[Wang et al, 2015] H. Wang, and D. Y. Yeung, Collaborative deep learning for recommender systems, *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD’15, pages 1235-1244, 2015.

[郭慧丰, 2018] 郭慧丰，*基于用户行为关系挖掘的个性化推荐模型及算法*，博士论文，哈尔滨工业大学计算机科学与技术学院，2018。

[项亮, 2012] 项亮，*推荐系统实践*，北京：人民有点出版社，2012。

# 致 谢

两年的研究生求学生涯，一转眼就将要结束了；还记得自己到校报到第一天，是那么让人兴奋，我对这个学校充满无限的好奇，对自己未来两年的求学生活有着无限的憧憬。随着本文的写作的结束，我两年的研究生学习于生活也即将结束。

首先感谢我的导师葛季栋副教授在学习上带来的指导。葛季栋老师是一位治学严谨、以身作则的导师。葛季栋老师组织组内学生进行学术汇报，并且每次都能够莅临参加，能够在每次汇报中提出自己独特的见解。葛季栋老师以身作则，对学术充满激情的态度，感染了我。是我学习和工作学习的榜样。

同时，我还要感谢我实习的公司，他们给我提供了非常好的机会，让我能够接触工业界实际应用的开发，让我能够把自己所学知识在实际生活中得到应用。感谢我实习的导师，他工作经验丰富，学识渊博，能够出色的解答我在工作中遇到的困惑。

同时，我还有感谢我可爱的舍友，他们在学习和工作上给我做了很好榜样，在生活中给我带来无限乐趣。

最后我要感谢我的家人，他们尊重我每一次决定，他们给我带来一个温馨家庭。

# 版权及论文原创性说明

任何收存和保管本论文的单位和个人，未经作者本人授权，不得将本论文转借他人并复印、抄录、拍照或以任何方式传播，否则，引起有碍作者著作权益的问题，将可能承担法律责任。

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含其他个人或集体已经发表或撰写的作品成果。本文所引用的重要文献，均已在文中以明确方式标明。本声明的法律结果由本人承担。

作者签名：

日期： 年 月 日