**面向重复记录检测的数据清洗算法的研究**

**作者姓名 张 攀**

**学校导师姓名、职称 刘志镜 教授**

**企业导师姓名、职称 吴春苗 高工**

**申请学位类别 工程硕士**

**面向重复记录检测的数据清洗算法的研究**

**西安电子科技大学**

**硕士学位论文**

**作者姓名：**张 攀

**领　　域：**计算机技术

**学位类别：**工程硕士

**学校导师姓名、职称：** 刘志镜 教授

**企业导师姓名、职称：** 吴春苗 高工

**学　　院：**计算机学院

**提交日期：**2017年6月

**学　号　 1503121785**

**密　级　 公开**

**学校代码 10701**

**分类号 TP39**

Supervisor: Liu Zhijing

Supervisor: Wu ChunMiao

Title: Professor

Title: Senior Engineer

A thesis submitted to

XIDIAN UNIVERSITY

in partial fulfillment of the requirements

for the degree of Master

in Computer Technology

**Research of Data Cleansing Algorithm for Duplicate** **Elimination**

By

Zhang Pan

June 2017

**西安电子科技大学**

**学位论文独创性（或创新性）声明**

秉承学校严谨的学风和优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同事对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文若有不实之处，本人承担一切法律责任。

本人签名： 日 期：

**西安电子科技大学**

**关于论文使用授权的说明**

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权属于西安电子科技大学。学校有权保留送交论文的复印件，允许查阅、借阅论文；学校可以公布论文的全部或部分内容，允许采用影印、缩印或其它复制手段保存论文。同时本人保证，结合学位论文研究成果完成的论文、发明专利等成果，署名单位为西安电子科技大学。

保密的学位论文在 年解密后适用本授权书。

本人签名： 导师签名：

日 期： 日 期：

摘要

随着信息技术的发展与信息化建设的深入研究，激烈的市场竞争对于企业信息化程度的要求也越来越高。在形形色色的企业信息化系统进行不断地迭代和重构时，数据库中积累了大量的脏数据[1]，主要包括错误数据、相似重复数据和缺失数据三种类型。这些脏数据产生的原因多种多样，包括数据来源不同，存储于不同的操作系统以及硬件平台等[2]。其中多源数据的归并造成的数据重复则是最关键的热点问题[3]。

本文在对相似重复记录技术发展和研究现状进行简要介绍的基础上，首先详细地阐述了基于不同实现方式的重复检测技术，然后在借鉴前人工作成果的前提下，提出和设计了一种改进的IMPN算法。该算法主要有三个改进点：通过统计字段区分度一定程度上改善了传统的MPN[4,5]算法在选择排序关键字时过于依赖专家经验的缺点；通过动态调整滑动窗口的步长与大小以节约时间消耗；通过标记排序关键字为空的记录提高算法应对缺失字段的能力，增强了鲁棒性。

其次当数据量增大到一定量级的时候，算法的效率问题随之凸显。本文将Java中的多核并行框架Fork/Join应用到IMPN算法中，经实验对比可以验证其高效性。多核并行并没有改变算法是基于排序-归并[6,7]思想的事实，所以该解决方案依旧存在算法的检测精度对于排序关键字过于依赖的问题。针对大数据量的相似重复记录检测，本文提出一种改进的人工神经网络算法。首先通过已知小数据集作为训练数据对BP神经网络进行训练，然后利用遗传算法弥补其收敛速度慢和易陷入局部极小值的缺点，改进后的神经网络对于相似度向量输入可以迅速得到记录是否相似的判断。该算法的主要优势在于并不需要直接估算各属性的权重，而是通过神经网络学习属性之间的内在关系来完成记录是否匹配的判断，随着环境的变化具有较强的自适应能力。经对比试验证明，该算法拥有较好的检测效果。

**关键词**：数据清洗，重复记录检测，相似度向量，BP神经网络，遗传算法

ABSTRACT

With the development of information technology and information construction, fierce market competition drives the enterprises to improve their information management. During the process of iteration and reconstruction of EIS, databases have accumulated a lot of “dirty data”, which consists of error data, duplicate data and missing data. There are varies of reasons of these data, such as the different data source, the different OS, the different hardware and so on. The key issue of these data cleansing problem is deduplication caused by the merge of multi-source data, which becomes the research focus nowadays.

In this paper we introduce the development and research status of the duplicatoin detection technology firstly, including their theory and implemention. Then a promoted algorithm called IMPN is present on basis of previos work. There are three improvement, decreasing the dependence of expertise and experience by the statistic of fields’ discrimination when choosing sort key, decreasing time consumption by dynamically adjusting the size and step length of sliding window, and promoting rubust by marking those records whose sort key is null..

The efficiency becomes a fierce problem as the data grows too large. Fork/Join, the multi-core parallel computing framework, is used to improve its performance. It is the fact that this solution is based on the Sort-Merge method, which depends much on the choice of sort key and perform not so good at time consuming, although with the Fork/Join framework. This paper provides a promoted artificial neural network algorithm for the deduplication of large dataset. Firstly, a small dataset is used to be the input of a BP network. After training process, the network will evolve as the chromosome, then this network could rapidly decide weather two records are the same entity with the similarity vector being the input. This solution’s advantage is that it gets the result through the inner relation of records rather than estimates the fields’ weight directly, which enhances its self-adaption ability as the environment changes. This algorithm reaches a higher accuracy as experiments prove.

**Keywords:** Data Cleansing, Duplicate Record Detection, Similarity Vector, BP Neural Network, Genetic Algorithm

插图索引

[图2-1 Flume组成结构示意图 5](#_Toc482816758)

[图2-2 Flume数据流向示意图 5](#_Toc482816759)

[图2-3 Flume总体架构示意图 6](#_Toc482816760)

[图2-4 Flume抽象数据流向示意图 7](#_Toc482816761)

[图2-5 Kafka架构示意图 8](#_Toc482816762)

[图2-6 Kafka消费组示意图 9](#_Toc482816763)

[图2-7 Storm总体架构示意图 11](#_Toc482816764)

[图2-8 Storm工作时序图 12](#_Toc482816765)

[图2-9 HDFS整体架构示意图 13](#_Toc482816766)

[图2-10 MapReuce整体架构示意图 14](#_Toc482816767)

[图2-11 HDFS读取文件时序图 15](#_Toc482816768)

[图2-12 HDFS写入文件时序图 16](#_Toc482816769)

[图3-1 User-CF算法基本原理示意图 19](#_Toc482887542)

[图3-2 Item-CF算法基本原理示意图 21](#_Toc482887543)

[图3-3 Memory-base CF基本过程示意图 21](#_Toc482887544)

[图3-4 Model-Based CF基本过程示意图 22](#_Toc482887545)

[图3-5 SVDFeature模型原理示意图 25](#_Toc482887546)

[图4-1 系统总体需求示意图 29](#_Toc482887674)

[图4-2 日志采集系统需求示意图 30](#_Toc482887675)

[图4-3 推荐系统需求示意图 32](#_Toc482887676)

[图4-4 推荐方式组成示意图 33](#_Toc482887677)

[图4-5 个性化推荐算法流程图 35](#_Toc482887678)

[图4-6 UI需求示例图 36](#_Toc482887679)

[图5-1 实时推荐系统的总体结构示意图 39](#_Toc483579301)

[图5-2 流式计算流程图 40](#_Toc483579302)

[图5-3 实时推荐系统的数据流向示意图 41](#_Toc483579303)

[图5-4 日志采集系统框架示意图 43](#_Toc483579304)

[图5-5 日志采集数据序列图 46](#_Toc483579305)

[图5-6 数据接入模式架构示意图 47](#_Toc483579306)

[图5-7 Kafka生产者示意图 48](#_Toc483579307)

[图5-8 Flume的配置示例图 48](#_Toc483579308)

[图5-9 推荐算法计算过程流程图 51](#_Toc483579310)

[图5-10 Storm流式计算框架示意图 54](#_Toc483579311)

[图5-11 实时推荐实现流程图 57](#_Toc483579312)

[图5-12 小型电商APP界面展示图 58](#_Toc483579313)

[图5-13 推荐准确率柱状图 60](#_Toc479433389)

[图5-14 推荐召回率柱状图](#_Toc479433389) 60

表格索引

[表2-1 相似重复记录的典型例子 31](#_Toc477465047)

[表5-1 用户浏览信息表 42](#_Toc482800994)

[表5-2 用户检索信息表 42](#_Toc482800995)

[表5-3 用户收藏信息表 42](#_Toc482800996)

[表5-4 用户购买记录信息表 43](#_Toc482800997)

[表5-5 Flume配置方案表 44](#_Toc482800998)

[表5-6 数据源表 52](#_Toc482800999)

[表5-7 物品评分表 52](#_Toc482801000)

[表5-8 物品出现次数表 52](#_Toc482801001)

[表5-9 物品同现矩阵表 53](#_Toc482801002)

[表5-10 物品-用户评分矩阵表 53](#_Toc482801003)

[表5-11 推荐结果表 53](#_Toc482801004)

[表5-12 集群环境配置表 59](#_Toc482801004)

[表5-13 推荐计算时间表 61](#_Toc482801004)

[表5-13 不同算法计算时间表 61](#_Toc482801004)

符号对照表

exp 以e为底的指数函数

min 取最小值函数

∑ 求和

缩略语对照表

EIS Enterprise Information System 企业信息化系统

OS Operating System 操作系统

BP Back Propagation 反向传播

MPN Multi-Pass Sorted Neighborhood 多趟近邻排序算法

IMPN Improved Multi-pass Sorted Neighborhood 改进的多趟近邻排序算法

SNM Sorted Neighborhood Method 紧邻排序算法

目录

[摘要 I](#_Toc507624050)

[ABSTRACT III](#_Toc507624051)

[插图索引 V](#_Toc507624052)

[表格索引 VII](#_Toc507624053)

[符号对照表 IX](#_Toc507624054)

[缩略语对照表 XI](#_Toc507624055)

[第一章 绪论 1](#_Toc507624056)

[1.1 研究的背景和意义 1](#_Toc507624057)

[1.2 国内外研究现状 2](#_Toc507624058)

[1.3 论文研究的主要内容 3](#_Toc507624059)

[1.4 论文结构 4](#_Toc507624060)

[第二章 重复记录检测相关算法概述 5](#_Toc507624061)

[2.1 相似重复记录概述 5](#_Toc507624062)

[2.1.1 相似重复记录介绍 5](#_Toc507624063)

[2.1.2 相似重复记录清洗过程 6](#_Toc507624064)

[2.2 字段相似度匹配算法 8](#_Toc507624065)

[2.2.1 字段匹配算法 8](#_Toc507624066)

[2.2.2 编辑距离算法 10](#_Toc507624067)

[2.2.3 Smith-Waterman算法 11](#_Toc507624068)

[2.2.4 N-Gram算法 11](#_Toc507624069)

[2.2 相似重复记录检测算法 12](#_Toc507624070)

[2.2.1 近邻排序算法 12](#_Toc507624071)

[2.2.2 多趟近邻排序算法 14](#_Toc507624072)

[2.2.3 优先队列算法 15](#_Toc507624073)

[2.3 算法的衡量标准 15](#_Toc507624074)

[2.4 本章小结 16](#_Toc507624075)

[第三章 基于MPN改进的IMPN算法 19](#_Toc507624076)

[3.1 Memory-CF模型 19](#_Toc507624077)

[3.1.1 User-CF模型 19](#_Toc507624078)

[3.1.2 Item-CF模型 20](#_Toc507624079)

[3.2 Model-CF模型 22](#_Toc507624080)

[3.2.1 Logistic模型 22](#_Toc507624081)

[3.2.2 LFM(SVD)模型 23](#_Toc507624082)

[3.2.3 FM模型 25](#_Toc507624083)

[3.3 系统推荐算法的优化 25](#_Toc507624084)

[3.4 本章小结 27](#_Toc507624085)

[第四章 Fork/Join并行框架在算法中的应用 29](#_Toc507624086)

[4.1系统功能性需求 29](#_Toc507624087)

[4.1.1 日志收集系统 30](#_Toc507624088)

[4.1.2 推荐系统 32](#_Toc507624089)

[4.1.3 UI系统 36](#_Toc507624090)

[4.2本系统的重点需求 36](#_Toc507624091)

[4.2.1 实时性问题 36](#_Toc507624092)

[4.2.2 冷启动问题 37](#_Toc507624093)

[4.2.3 时效性和多样性问题 37](#_Toc507624094)

[4.3 本章总结 38](#_Toc507624095)

[第五章 改进的遗传神经网络算法 39](#_Toc507624096)

[5.1 整体框架 39](#_Toc507624097)

[5.2 日志采集模块 42](#_Toc507624098)

[5.2.1 日志数据类型 42](#_Toc507624099)

[5.2.2 日志采集系统框架 43](#_Toc507624100)

[5.2.3 日志采集系统连接器 44](#_Toc507624101)

[5.3 数据接入模块 47](#_Toc507624102)

[5.3.1 数据接入框架 47](#_Toc507624103)

[5.3.2 数据接入连接器 48](#_Toc507624104)

[5.4 离线计算平台模块 50](#_Toc507624105)

[5.4.1 基于物品的协同过滤算法原理 50](#_Toc507624106)

[5.4.2 算法实现过程 50](#_Toc507624107)

[5.5 流式计算平台模块 54](#_Toc507624108)

[5.5.1 Kafka与Storm的连接器 54](#_Toc507624109)

[5.5.2 实时推荐计算 55](#_Toc507624110)

[5.6 数据存储和展示模块 57](#_Toc507624111)

[5.6.1 数据存储和输出 57](#_Toc507624112)

[5.6.2 UI展示 58](#_Toc507624113)

[5.7 实验与测试 58](#_Toc507624114)

[5.7.1 评测标准 58](#_Toc507624115)

[5.7.2 实验设计 59](#_Toc507624116)

[5.7 本章小结 62](#_Toc507624117)

[第六章 总结与展望 63](#_Toc507624118)

[参考文献 65](#_Toc507624119)

[致谢 67](#_Toc507624120)

[作者简介 69](#_Toc507624121)

# 第一章 绪论

## 1.1 研究的背景和意义

信息化时代人类正在由工业化时代进入信息化时代，经济学家们普遍认为，进入21世纪后，信息将成为第一生产要素，同时将构成信息化社会的重要技术物质基础。为了在激烈的市场竞争中占据先机，各行业如保险、金融等纷纷加快了信息化的步伐。随着数据库技术的快速发展和广泛应用，形形色色的企业信息化系统应运而生。数据库的大小也与日逐增，人们已经得出如下结论，每过20个月世界上信息的数据量就会翻一番。

从规模庞大数据库中提取重要信息，从而对企业单位的发展提供参考，为决策者提供技术支持，是近年来数据挖掘的研究重点。然而由于不可避免的人为录入错误，或者是不同的数据表示方法，抑或是从不同的数据源合并数据甚至数据存储于不同的操作系统和物理设备，都不可避免地降低了系统的数据质量，从而产生各种“脏数据”。脏数据的类型主要包括重复数据、不完整数据、错误数据等。对这些数据若不进行清洗，则会影响信息化系统的正确运行，使得数据中提取的信息不再可靠，为企业决策支持和商务应用带来负面影响。因此，为了确保数据的准确性、一致性，数据清洗显得尤为重要。

最早的数据清洗过程需要大量的人为操作，所以当遇到较大规模的数据集，就会凸显出人为操作的低准确性和低效性。所以在当前数据规模急剧加大的情况下，只有借助信息技术，数据清洗才能实现其高效性。目前的信息化清洗过程中，仍不能完全离开专家的经验、人工的操作等行为，所以研究的一个重要方向就是尽可能减少人为的参与和影响。

相似重复的记录是数据库中降低数据质量最重要的一个原因，所以如何高效地检测和去除重复数据是数据清洗研究范畴的一个热点问题[8,9]。

同一个实体在数据库中不同的展现形式是相似重复记录的本质，它主要会引发以下的问题：

(1)资源浪费：重复记录会造成数据冗余，导致存储空间的极大浪费。

(2)破坏数据一致性：相似重复记录之间的关系可能是互为补充，也可能存在部分的冗余，甚至互相矛盾。当它们共同对应的现实中的实体发生变化，极有可能导致这些记录中只有某个或者某些记录发生改变，而其余无法同步更新。

以上，相似重复记录的检测与消除，保证了数据的一致性、减少资源的浪费，是数据清洗的重要环节。

## 1.2 国内外研究现状

早在上个世纪50年代，数据清洗已经开始了相关研究。将出自不同数据源的数据集进行整合自研究开始以来就被认为是一个困难而且极为重要的问题，开始主要是从数据连接[10]、数据实体识别[11]、对象识别等问题来研究，是商业保险、医疗、等领域中的研究重心之一。美国当时清除全美社会保险号数据集中的错误数据被视为数据清洗技术研究的开端[12]。

数据清洗的研究重心在以下几点：重复记录检测、异常数据检测、缺失数据的处理。数据仓库的出现以及数据挖掘相关技术的发展和应用，造成了多源数据进行合并容易出现大量重复数据的问题。因而相似重复记录的检测与清除成了数据清洗领域的研究重点。

在重复记录清洗方面，国外展开了大量的研究，主要的工作在于两个方面——属性匹配和重复检测。属性匹配问题的解决方法主要有Smith-Waterman算法、递归属性匹配算法、和R-S-W算法[13]。

相似重复记录检测大部分的算法采用排序然后归并的思想，即先将数据连接成一整个数据集，之后按照某种规则进行排序，将相似重复的记录排列在附近，最后通过某种相似判断方法检测出重复的记录，主要的差异在于排序方法和相似检测方法。最基本的算法是Jaro[14]提出的“排序&合并（Merge/Purge）”算法。这种算法存在明显的缺点，许多研究人员在此基础上提出了各种各样的改进思路和算法实现。

Monge等[15]将数据库中的一条记录视为一个字符串，在排序和比较的时候采用优先级队列的方法，检测相似重复时则使用了基于字符串的编辑距离。Hernandez等[5,4]提出了一种名为“多趟近邻排序”的算法，即MPN（Multi-Pass Sorted Neighborhood），该算法多次执行SNM（Sorted Neighborhood Method）同的排序关键字段以及较小的滑动窗口，最后判定记录相似时候使用的是一种规则定义的判等理论，它是使用C语言重写的OPS5[16]规则编程。Newcombe等[17]则采用应用相关的键值以将重复记录聚类到相近的位置。Qiu等[18]则是首先计算每条记录的N-gram统计值，然后根据这个N-gram值对数据集进行排序，最后再用用优先级队列的方式聚类检测重复记录。Gianni Costa等[19]采用文本聚类中的增量技术将新数据划分到最近的已知重复的聚类中，目的是为了解决大文本库中的相似检测问题。Alfredo Ferro等[20]使用了基于q-grams的相似度衡量函数，可以避免许多不必要的比较和判断，提高了时间效率。

国内的相关研究主要是对已知算法的改进和创新以实现更高的精度和效率。复旦大学周傲英等[21]等比较早开始数据清理的研究工作。邱越峰等[22]提出了基于N-Gram的相似记录检测算法。算法以一条数据的N-Gram值作为排序键，在对因为拼写造成的重复记录进行聚类时表现良好。!!!1韩京宇等[23]提出了一种基于q-gram空间层次聚类的方法。不同于传统的“排序&合并”的方法，该算法将数据映射成q-gram空间中的点，形成具有相似性层次子空间序列，采用层次聚类来实现相似重复记录的检测和清洗。方法不仅能避免应外排序引起的大量 I/O，而且克服排序不能保证将相似记录排在邻近位置的缺陷，还能通过层次逐步细化聚类来检测相似重复记录的方法提高检测精度。陈伟[24]提出一种按字段等级划分并将等级转变成权重的方法检测相似重复记录，给各个属性赋予合适的权重以提高检测精度，同时提出了长度过滤方法减少不必要的编辑距离计算进行相似重复记录检测优化。李星毅、庞雄文和周丽娟等[25,26,27]对大数据集相似记录检测问题。采用某种方法确定关键属性，依据关键属性进行数据分组，然后再划分后的组内进行相似重复记录检测的研究。这两种方法不同之处是确定关键属性的方法和组内聚类比较方法。

近几年，国内外的数据清洗市场的发展很快，其中包括商业上的数据清洗软件,也有各大学和研究机构开发的数据清洗软件[28,29]。

综上所述，国内外研究人员对数据清洗尤其对相似重复记录清洗的检测的研究已经得到取得了诸多进展，但仍旧或多或少存在适用面窄或者检测效率和精度不足等问题。

## 1.3 论文研究的主要内容

从国内外的研究现状可以看出，数据清洗领域己取得发展，促进了各行各业的发展。但数据清洗领域还存在一定问题，主要体现在：

(1)尽管检测重复记录受到很大关注，采取了许多措施，但检测效率与检测精度问题并不令人满意。大数据量的相似重复检测消耗大量时间，有待于更好的检测算法。

(2)大多数数据清理工具都是针对特定领域，其应用受到一定限制，未来特定领域的数据清洗仍是应用重点，但通用的清理方案应受到越来越多的关注。!!!1

(3)传统的相似重复记录检测大多基于“排序&归并”的思想，排序的效果以及最终归并的结果受排序关键字影响较大，尤其是当数据库排序关键字对应的字段为空或者是错误数据时，部分重复记录无法被正确的检测到，从而影响数据清洗的质量。

如何高效地检测相似重复记录，进而剔除数据库中的冗余数据，一直是数据清洗研究的重点问题。本文在分析了常用相似重复记录检测算法的基础上，针对传统的基本紧邻排序算法MPN在时间消耗和检测精度的不足，提出了改进的IMPN算法。IMPN算法的改进点主要在于：

(1)通过统计字段区分度一定程度上改善了传统的MPN算法在选择排序关键字时过于依赖专家经验的缺点。

(2)通过动态调整滑动窗口的步长与大小以节约时间消耗。

(3)通过标记排序关键字为空的记录提高算法应对缺失字段的能力，增强了鲁棒性。

为了节约IMPN算法的时间消耗，本文将多核并行框架Fork/Join应用到改进后的算法中，经过充分的对比实验可以验证其高效性，从而充分地利用了多核多线程计算机的优势。

随着人工神经网络研究的兴起，越来越多的跨学科研究正在如火如荼地展开。本文将训练后的反向传播神经网络应用于相似重复记录检测中，利用两条记录对应字段间的相似度，构建基于神经网络的检测模型，利用遗传算法对网络模型的权值进行优化，使用遗传神经网络组合多个字段上的相似度来检测相似重复记录。在多个测试数据集上的测试结果表明，该方法能够提高相似重复记录检测的准确率和检测精度。

## 1.4 论文结构

论文的结构如下所示：

第一章，介绍了的数据清洗研究的背景和意义，相似重复记录的国内外的研究与发展现状，简单描述了论文的主要研究目的以及研究内容，展示了论文的组织架构。

第二章，主要介绍了相似重复记录检测的相关算法。第一部分首先简单介绍了衡量字段相似度的相似度检测有关算法，分析了它们各自的有缺点以及适用条件。然后在第二部分首先介绍了最原始的直观方法，然后对最基本的紧邻排序算法SNM，介绍了算法的基本原理、设计思路以及大致实现和算法的优缺点。然后对在SNM基础上改进的MPN算法的设计原理、实现与优缺点分析进行了描述。除此之外还介绍了常用的算法思路比如优先队列算法等。

第三章，首先介绍了IMPN算法的提出背景，然后详细介绍了算法的设计思路和其改进点，并采用SNM算法和MPN算法作为对照，进行了对比实验以验证算法的效率和检测精度等优势。

第四章，首先介绍了神经网络的基础知识以及神经网络在相似重复记录检测中的应用，接下来介绍了多种对于BP神经网络的改进思路，重点介绍了遗传算法对其的改进，并说明了如何使用改进的BP神经网络进行相似重复记录检测。通过使用不同数据量的测试数据集进行改进证明该算法的优势，并分析了它的缺点。

第五章，介绍了多核并行的基础理论，将Fork/Join并行框架应用到IMPN的设计与实现思路，并与IMPN做对比实验，验证多核并行实现对于算法在时间效率上的提高。

第六章，总结了本文的内容包括IMPN算法与遗传神经网络进行重复检测研究中遇到的问题，以及算法本身的不足之处，并对未来的研究进行展望。

# 第二章 重复记录检测相关算法概述

## 2.1 相似重复记录概述

### 2.1.1 相似重复记录介绍

!!!2由于数据输入错误、不标准的缩写词，或其它原因，数据库中可能包含关于现实世界同一实体的重复记录。虽然关系数据库系统不允许含有重复主键值的记录输入，但是，由于数据输入错误，不管主键的值是否被这些错误影响，关系数据库不能再保证不存在重复的记录。因此，在数据清理中，相似重复记录的检测与清除是一个重要问题。

为了便于给相似重复记录下定义，下面首先看一个关于的重复记录实例，如表2.1所示。

表2.1 学生信息表中的重复记录

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Num | Name | Sex | Brithday | Department |
| 20110001 | John Smith | Male | 1985/02/01 | College of Computer Science,Guilin University of Electronic Science and Technology |
| 20110002 | Harry Potter | Female | 1988/08/12 | College of Computer Science,Guilin University of Electronic Science and Technology |
| 20110003 | J. Smith | M | 1985/2/1 | College of Computer Science,Guilin University of Electronic Science and Technology |
| 20110004 | Harry Potter | Female | 1988/08/12 | College of Computer Science ,Guilin University of Electronic Science and Technology |
| 20110005 | Mr.John S | Male | 1985-02-01 | Corllge of Computer Science ,Guilin University of Electronic Science and Technology |

表2.1关于学生信息的五条记录。表2.1中学号为2011002和2011004这两条记录所有属性值完全相同，因此，这两条记录可以认为是完全重复记录。而表中学号为2011001、2011003和2011004这三条记录看起来不相同，但实际上只是姓名属性值“John Smith”、“J. Smith”、“Mr.John S”书写形式不同但都是同一个人；性别属性的值“Male”、“M”采用了全称和缩写两种表示男性，出生日期属性的值“1985/02/01”、“1985/2/1”、“1985-02-01”采用不同格式书写表达式同一个日期；学院属性值中出现了最常见的拼写错误把“College”拼写成“Corllge”；可以认为三条记录描述的是同一个学生。因此，这里的三条记录可以认为是相似重复记录。根据以上分析，数据源中的重复记录可分成完全重复记录和相似重复记录，分别描述如下：

(1)完全重复记录[30]：指在数据表中除了主键外，其它各属性完全相同的记录，或者没有主键，所有属性完全相同的记录。

(2)相似重复记录[30]：指客观上表示现实世界同一实体的，但是由于在格式、拼

写上有些差异而导致数据库系统不能正确识别的记录。

### 2.1.2 相似重复记录清洗过程

数据清洗的目的就是首先通过某种科学的方法检测出相似重复记录，然后根据合适的规则和策略自动或手动清除这些相似重复记录。分析相关研究资料可知，完全重复记录检测大多采用“排序&合并”的方法：首先将数据库中的记录进行排序，重复的记录就会重新排序在相邻位置，通过对相邻记录比较，就可以检测出完全重复记录。这一方法被扩展得到相似重复记录的清洗方法的过程：数据调入→记录排序→记录相似检测→相似重复记录合并/清除。相似重复记录清洗过程如图2.1所示。



图 2.1

如图2.1所示，相似重复记录清洗过程描述如下：

(1)数据调入。通过数据库连接接口把含有相似重复记录的数据源中的数据调到系统中；

(2)相似重复记录检测。如果数据规模较大应从算法库中调用排序算法，执行记录之间的排序；在记录已排序的基础上，记录相似检测模块从算法库中调用相似检测算法，作邻近范围内记录间的相似检测，从而计算出记录间的相似度，并根据预定义的重复识别规则，来判定是否为相似重复记录。为了能检测到更多的重复记录，一次排序不够，要采用多轮排序比较，每次排序采用不同的键，然后把检测到的所有重复记录聚类到一起，从而完成重复记录的检测；若数据规模较小可以直接进行记录相似检测；

(3)数据合并/清除。对所检测出的每一组相似重复记录根据预定义的合并/清除规则，完成相似重复记录的合并处理。

由图2.1可知，记录排序和相似重复检测是相似重复记录清洗的两个核心步骤，本文中重点讨论相似重复记录检测的技术。

## 2.2 字段相似度匹配算法

为了从数据集中检测并清除重复记录，首要的问题是如何判断两条记录是否是重复记录。一条记录由不同的字段组成，字段之间的相似度考察是衡量两条完整记录相似与否的主要方法。目前相似重复记录的识别算法(也称为匹配算法)主要有：字段匹配算法（Field Matching Algorithm）、编辑距离法（Edit Distance Algorithm）、聚类算法（Clustering Algorithm）、N-Grams算法、对XML数据的匹配算法（Matching Algorithm for XML）等。

### 2.2.1 字段匹配算法

字段匹配是用来确定两个字段值是否表示同一个语义实体的句法上的可替换者, 是记录匹配的基础。基本的字段匹配算法和递归的字段匹配算法是两种最基础的字段匹配算法。

基本的字段匹配算法是基于传统的相似度匹配算法发展而来的[31]，能够有效的识别出包含前缀在内的相似重复记录，下面给出几个关于基本字段匹配算法的定义：

(1)元字符串：元字符串是指利用字符序列中的标点符号，空格以及其他分割符号，将字符序列分割成一段段以单词词组为中心的字符串[32]。

(2)字符序列的匹配度计算方法如下：

 公式2-1

基本的字段匹配算法的算法思想是将一条记录看作是一个字符序列，然后按照某些定界符将这个字符序列分割成一段段的字符串，也就是元字符串，然后通过基本的字段匹配算法公式，计算其匹配度。在这个过程中，元字符串相互匹配的标准主要有两种：一种是两者之间完全相同，每一个字符及其在其中的位置都相同；另一种情况则较为特殊，其中一条字符串的前几位和另一条字符串完全相同。只有在这两种情况下，才能计入匹配的数目中。两条记录间相似的元字符个数比上两条记录总的字符序列，得到的结果便是两条记录的基本字段匹配算法的相似度。

表2.2相似重复记录的典型例子

|  |  |  |  |
| --- | --- | --- | --- |
| Num | Name | Address | Tel |
| A | Join Kow | Comput Sci & Eng Dept, University of California, San Diego | 888-1234 |
| B | K. Join | Department of Computer Science, Univ. Calif, San Diego | 8881234 |

如表2.1中两个地址字段，A=“Department of Computer Science, Univ. CaliL, san Diego”，B=“Comput. sci. &Eng. Dept. university of California, san Diego”将这两条记录按照标点符号分割成小的字符串，其中，由于英文的阅读习惯，英文中的of在字段中作为标点符号使用，因此，经过分割操作，A和B分别变成如下两种情况：

A={Department，Computer，Science，Univ，Calif，san，Diego}

B={Comput，sci，Eng，Dept，university，California，san，Diego}

对他们两者进行分析可知，两条字段共有6条相似字符序列，分别为A的Computer，Science，Univ，Calif，san，Diego和B字段的Comput，sci，university，California，san，Diego。其中sci是Science的前缀。则两条记录的相似度为6/(|A|+|B|)/2)=0.8。

基本字段匹配算法有着很大的缺陷，根据它的算法思想，当一条字符串是另一条字符串的前缀缩写时，则将两者看作相似，但是，名词缩写并不仅仅局限于前缀或后缀缩写，更多的情况是类似于英文姓名缩写，取有代表性、并且和已存在的名词不重复的字符来代替整条字符串，当出现这种情况的时候，该算法并不能很好的识别出来。如上述A字符串中的Department的缩写是B中的Dept，相似重复记录的数据清洗系统中，这两个字符串被认为是相似数据，然而在基本的字段匹配算法却不能处理。但是，基本字段匹配算法却很直观，采用先分割后计算的方式，首先截取原记录，然后根据两者之间原子匹配数来计算相似度，其效率及准确度较其他匹配算法，有着很大的优势，为了保留这种优势，Minton等人对基本字段匹配算法进行了改进，提出了递归字段匹配算法。

递归字段匹配算法是一种较为准确的算法，能够有效的识别出含有缩写等情形的相似重复数据。它是基本字段匹配算法的升级，通过给一个字符串中的分割符以及标点设置优先级，并按照优先级的不同，从高到低，来对字符串进行分割，并且运用嵌套功能，不断重复这个过程，直到分割出来的子串不能被分割为止。利用这种递归的方式，将一条字符串分割成许多的原子字符串，然后再采用递归的方式，从上到下，不断匹配每一对的原子字符串，从而得到他们的相似度[33]。每两条字符串之间的相似度都是他们的子串的相似度的平均值，递归字段匹配算法的计算公式如公式2-2：

 公式2-2

其中，Ai、Bj分别是分割AB后得到的子串，、分别表示字符串A，B中的子串的个数。该算法适用于名词缩写和顺序颠倒等环境，能够有效的识别出含有这些情况的相似重复数据，例如上一个例子中的两条字符串匹配问题中，Department与B中原子字符串的相似度的最大值是与B中的Dept相比的4/7=相同字符个数/两者总共个数，computer最大值是6/7，science的最大值是3/5，univ的最大值是4/7,calif的最大值是2/3，san的最大值是1，Diego的最大值是1，则字符串A和字符串B相匹配，两者的相似度为0.75238,可以认为两者为相似重复记录。但是，递归字符匹配算法的缺点也很明显，与基本字段匹配算法相比，大量的使用嵌套递归功能，使得该算法的时间复杂度变高，是子串个数的平方；当字符串较长是，匹配每一个原子字符串都需要遍历另一个字符串，效率低下，而且对于包含大量错误字符的字符串记录，该算法并不能很好的识别。

### 2.2.2 编辑距离算法

Levenshtein等人在1965年提出一种基于编辑距离的相似重复记录匹配算法，又被称为L-距离算法[34]。编辑距离算法是Levenshtein等人所提出的一种用于发现相似重复数据的一种先进算法，他的主要思想是通过比较两条记录A和B，对他们进行分析，找到两者在字符上的不同，从而计算从记录A变化到记录B所需要经过的插入、删除、替换等操作的次数，该次数被称为编辑距离。所需要的操作次数越少，就代表这两条记录之间的编辑距离越近，两者之间的相似度就越大。因此该算法也被叫做L-距离算法。自从L-距离算法在1965年被提出以后，由于他的高效和高准，就得到了普及。例如两条记录children和child，如图2.2所示：



图2.2

记录children转换到记录child需要删除后面的三个字母r、e和n，共需要三个删除操作，则这两条记录的编辑距离为3。计算记录间的编辑距离算法在计算不同字符串的相似度上有着广泛的应用，它能够有效的识别出例如拼写错误、名词缩写等问题所产生的相似重复记录。由于编辑距离算法的高效性和广泛性，国内许多专家学者在它的基础上开发出更多的算法：刘旭辉等人在2008年在编辑距离算法的基础上，提出了一种基于NFA的编辑距离方法[35]，它将需要进行相似度计算的两条记录看作是一个二叉排序树，采用某种遍历方法，得到这个二叉数的有序序列，按照这个有序序列查找树进行索引，能够很大的提高识别效率和准确率；赵作鹏等人为了解决编辑距离算法无法解决单词顺颠倒等问题，在其基础上提出了一种改进的非相邻位置的编辑距离算法，能够有效的检测出单词替换和顺序颠倒等字符串之间的相似度[36,37]。

### 2.2.3 Smith-Waterman算法

S-W算法最开始被用于生物领域，用于匹配蛋白质和DNA序列，后来便逐渐被应用在相似重复记录的数据清洗中。S-W算法是一种动态编程技术，通过罚值和间隙计算不同记录之间的相似度。其算法的基本思想主要通过一个矩阵E来计算相似度，矩阵E是由a，b，c三个参数来确定的，其中，参数a为给定匹配字母表β的|β|\*|β|矩阵，字母表中第i个字符和第j个字符的匹配度放置在由字母表β组成的矩阵a的(i,j)位置上；b则为间隙开始的罚值；c为间隙间隙继续的罚值。矩阵E的组成元素(i,j),与字符串X和字符串Y的子串相关，他们的前i个和前j个字符组成他们的两个子串，这两个子串之间的最佳匹配值作为矩阵E的每一个元素，其计算公式为：

 公式2-3

当的值高于标准时，则认为X和Y这两者有着相同或相似的属性，互为相似重复数据。S-W算法可以有效解决字段匹配算法所包含的不足，对于包含不正确值的相似重复记录，能够非常准确的将他们辨别出来。但是对于名词缩写和前后顺序颠倒情况的相似重复记录，该算法未能取得较好的结果[38]。

### 2.2.4 N-Gram算法

基于N-Gram聚类算法的基本思想是：给每个记录赋一个N-Gram值，以该值为键来对记录聚类[39]。该算法能适应常见的拼写错误，从而较好地聚类相似重复记录。文献[22]提出了N-Gram标记算法和N-Gram聚类算法，用基于域的重复矩阵代替全局的重复矩阵，将相似记录聚到一个cluster中，并对每个cluster中的记录做改进的Pair-wise比较，从而提高了检测精度。文献[53]提出了一种基于N-Gram层次空间的聚类算法DGHS，定义记录r1、r2的N-Gram相似性，如下所示：

 公式2-4

其中表示记录r的所有属性组成的多重集合。通过将记录字符串根据依次加长的N-Gram映射到不同的子空间，采用层次聚类进行归并，从而实现相似重复记录的检测。

表2.3给出了几种N-Gram算法及其与传统排序-合并算法中常用的Merge/Purge算法[40]比较的情况。

表2.3几种N-Gram算法的比较

|  |  |  |  |
| --- | --- | --- | --- |
| 算法 | 时间复杂度 | 优点 | 缺点 |
| 基本的N-Gram算法 | O(n) | 能适应常见的拼写、交换、替换错误,是单词位置无关的 | 对插入、删除错误敏感，识别效果差 |
| 改进的N-Gram算法 | O(n\*c)c为常数，时间复杂度略高于基本的N-Gram算法 | 能自动校正单词的插入、删除错误，提高检测精度 | 时间开销大，对错误分布均匀的数据精度改进效果不明显 |
| 基于N-Gram层次空间的聚类算法 | O(h\*k)，k为簇的个数，h为聚类的趟数 | 能自动校正单词的插入、删除错误，提高检测精度避免外排序操作引起的大量I/O，对大数据量的处理具有良好的伸缩性 | 算法细节还需改进 |
| Merge/Perge算法 | O(nlgn) | 解决了如何检测数据库中完全重复记录的问题 | 对字符位置敏感，随着数据量增大算法精度会下降 |

## 相似重复记录检测算法

相似重复记录清除是在重复记录识别后进行的数据的合并或删除操作[41]，是影响数据质量的关键步骤。最简单直白的方法是对每条记录进行互相之间的比较，该方法的识别精度非常高，但是在数据量较大的情况下，其处理时间会让用户难以忍受[42]。目前的清除算法主要是采用排序-合并的思想，先将数据库中的记录排序，后通过比较临近记录是否相似来检测记录是否重复。常用的算法有近邻排序法（Sorted-Nighborhood Method，SNM）、多趟近邻排序法（Multi-Pass Sorted-Neighborhood，MPN）、优先队列算法（Priority Queue Strategy，PQS）、Delphi算法（Delphi Algorithm）等。

### 2.2.1 近邻排序算法

目前解决海量数据的记录相似匹配检测中，采用的较多的是近邻排序算法（Sorted-Neighborhood Method，SNM算法）。算法的基本思想是：将数据集R中的记录按指定的关键字（key）排序后，然后在排序后的数据集上移动一个固定大小的窗口，只检测窗口内的记录，判定它们是否匹配。由此来减少记录的比次数。SNM算法步骤为：

(1)生成关键字段：抽取表中的相关属性，生成关键字段；

(2)排序：按照生成的关键字排序对数据集中的记录排序；

(3)记录匹配：在排序后的数据集上滑动一个固定大小的窗口，数据集中的每条记录看做新记录仅与窗口内的记录进行比较，直到数据集的最后一条记录。

如果窗口的大小是包含N个记录，当窗口移动时，原来窗口中的第一条记录移出窗口，新进入窗口的记录与窗口中的旧的N-1记录相比较，判定是否匹配，如图2.3所示。



图2.3

算法执行时，首先要将数据级中的记录进行归一化处理，然后再进行记录的比较检测，下面给出算法实现，见算法2.1。

算法2.1最近邻算法SNM

输入：DB={t1,t2,…,tn} //数据集DB

A //表示成员之间距离的邻接矩阵

输出：K //簇集

K1={t1};

K={K1};

K=1;

For i=2 to n do

find the tm in some cluster Km in K such that dis(ti, tm) is the smallest;

IF dis(ti,tm)≤t Then //t 为阀值，由经验值确定

Km=Km∪ti

Else

k=k+1

Kk={ti}

SNM算法采用滑动窗口的方法，每次可以只比较窗口中的N条记录，提高了匹配效率；采用滑动窗口也极大的提高了比较速度，只需要进行N×n次比较，显然，N与n相比是小的多。但是SNM方法存在这样的两个缺陷[43,44]：

（1）对排序关键字的依赖性太大。SNM方法检测重复记录的精度很大程度上依赖所创建的排序的关键字，如果选取的关键字不当，邻近的位置不是相似重复记录，而使本来是重复的记录物理位置相距很远。

（2）滑动窗口的大小N的选择很难控制。当N值较大时，需要比较的次数较多，这样就增加了时间复杂度，N若较小的话，许多相似的记录可能就被漏配。当所有记录中各个重复记录聚类数差别较大时，则N的选取无论多大都不适合，除非N等于n。

### 2.2.2 多趟近邻排序算法

针对SNM算法存在的缺陷，Hernandez[4]等人提出了多趟近邻排序算法（Multi-Pass Sorted Neighbothood，MPN），该算法的基本思想是独立地执行多趟SNM算法，每趟创建不同的排序关键字和使用相对较小的滑动窗口。然后采用基于规则的知识库来生成一个等价原理，作为合并记录的判定标准，将每趟扫描识别出的重复记录合并为一组，在合并时假定记录的重复是具有传递性的，即计算其传递闭包(transitive closure)。传递闭包是指若记录Rl与R2互为重复记录，R2与R3互为重复记录，则Rl与R3互为重复记录。通过将每趟扫描识别出的重复记录计算传递闭包的方法，可以得到较完全的重复记录集合，能部分解决漏匹配问题。

但是MPN算法使用传递闭包容易引起误识别，针对这一缺点，文献[45]做了两个方面的改进：采用窗口大小可在Min和Max之间变化的可变窗口，通过相似度与阈值的比较及时调整窗口值，减少遗漏；不采用传递闭包，而引入有效权值，以消除字段缺失造成的负面影响。对于进行比较的两条记录A、B，假定参与比较的字段有n个，记录A、B的相似度计算方法如下所示：

 公式2-5

其中，。只有当两条记录在第i个属性上对应的值都不为空时，才进行字段比较，此时Valid[i]=1，对应的权值Wi为有效权值；否则Valid[i]=0。将总相似度与阈值比较从而判断重复记录，提高了识别的准确性，减少了误识别。其准确率几乎不受数据量的影响，一直保持在99%以上，取得了较好的识别效果。但参数的确定依赖领域知识，缺少特定的标准。

### 2.2.3 优先队列算法

优先队列法是由Monge等提出的[46]领域无关的处理相似重复记录的方法。用一个重复记录的优先队列来代替固定大小的窗口，优先队列的每一项代表的是一个重复记录聚类，而不是一条记录。算法顺序扫描记录，如果当前记录Rj本来就是优先队列对应簇中的成员，那么扫描下一条记录；如果不是，与优先队列包含的项比较，如在优先队列中有重复记录，则将这条记录合并入匹配记录所在的簇中，包含有这个记录的集合进入优先队列并有最高的优先级，如果扫描整个优先队列后发现，Rj不属于任何一个簇，则将Rj本身所在的聚类加入到优先队列中，并使其具有最高优先级，成为该聚类的第一个代表记录。优先队列越大，算法的检测精度越大，但是运行时间也越长。采用优先队列策略识别重复记录的精度很大程度上依赖于排序所选择的关键字。为了解决一个关键字不足以将所有重复记录聚集在一起的问题，可以每趟使用不同的关键字独立执行多趟优先队列算法，最后合并每趟扫描的结果。在合并时假定记录的重复具有传递性。图5描述了一个两趟式的优先队列扫描法[47]。

图2.4优先队列扫描算法

基于优先队列的算法以重复记录聚类为元素的优先权队列结合特征记录，通过适当的设置阈值，可以大大减少不必要的记录比较次数。而且该算法几乎不受数据规模的影响，能很好的适应数据规模的变化，但是检测结果的精度与这阈值的设定有很大的关系。而且排序算法的缺点在优先队列这里依然存在。

## 2.3 算法的衡量标准

衡量重复记录检测算法效率的标准，应该是算法是否能把数据集中存在的所有重复记录都检测出来。常用的标准主要有召回率，误识别率一和精确度。下面分别来介绍这几个度量的标准[48]。

(1)召回率（Recall）

也被称为百分比采样数。它定义为被重复记录检测算法正确识别出的重复记录占数据集实际包含的重复记录的百分比。其计算公式为：

召回率 = ( 正确识别的重复记录数 / 实际包含的重复记录数 ) \* 100%

假设我们有7条记录，A1，A2，A3，B1，B2，B3，C1，其中{ A1，A2，A3}与{ B1，B2，B3}分别是记录A和B的重复记录。通过一个数据清洗的过程识别出{A1，A2，C1}和{B1，B2}是重复记录，则其Recall=4/6\*100%=66.7%。

(2)误识别率（False-Positives）

有时被称为false merge，它的定义是被重复记录检测算法错误地识别为重复记录的数目占被算法识别为重复记录总数的百分比。误识别率越低表明算法结果的置信度就越高。其计算公式为：

误识别率 = (被错误地识别为重复记录的数目 / 被识别为重复记录的总数)\*100%

在上面的例子中，检测算法错误地把C1识别为一条重复记录，则其误识别率为1/5\*100% = 20%。

(3)精确度（Precision）

精确度是指识别出的重复记录表示的是否都是同一个实体，即是否存在误识别的情况。其计算公式为

精确度 = 1 –误识别率

一般而言，当召回率从0增加到100%时，精确率从100%减少到0。!!!2

## 2.4 本章小结

本章首先介绍了相似重复记录产生的原因和大部分去重算法的主要清洗过程，然后介绍了四种不同的相似度匹配算法：字段匹配算法、编辑距离算法、SW算法以及N-Gram算法。

其次本章重点介绍了几种不同的相似重复记录检测算法。从算法的设计原理，实现步骤，主要优缺点等方向对SNM算法、MPN算法、优先队列算法进行了介绍。

最后本章介绍了衡量算法检测效率的几个重要指标。

# 第三章 改进的基于字段区分度提取键值的IMPN算法

多趟近邻排序算法（Multi-Pass Sorted Neighbothood，MPN）在传统的近邻排序算法（Sorted Neighbothood Method,，SNM）的基础上，通过计算每趟近邻排序算法重复检测结果的传递闭包，实现了更好的检测效果。它能够以更小的滑动窗口进行重复检测，并且可以检测到一些人工都难以发现的重复记录。但是MPN并没有克服SNM算法对于排序关键字的选择标准过于依赖领域专家的经验，并且当记录的排序关键字字段为空时，检测效果会大打折扣。所以本文提出了基于统计规律的提取键值的IMPN算法，通过统计不同字段对于数据集记录的区分度，以及对空字段情况下的标记处理，可以一定程度上克服MPN对于专家知识的依赖，并提高了算法的鲁棒性。

## 3.1 基于字段区分度提取键值的方法

### 3.1.1 传统的提取键值的方法

基于排序/合并思想的重复记录检测算法，在排序之前需要提取和生成记录的排序键值，然后再对记录按照键值的顺序进行排序。完全理想状态下相似重复记录的键值也相等，所以排序过后相似重复记录则汇聚到邻近的位置，进而容易检测出相似重复记录对。

键值的定义是，从记录中提取出来的不同属性组成的序列或者属性的字符串子集[5]。键值的选择十分关键，因为只有选择恰当的排序键值才能够使得数据集在经过排序之后，相似重复记录聚集在排序后的数据集合中相邻的位置，这样在之后的滑动窗口中才能覆盖到相似重复的记录，合适的排序键值还可以减小之后排序使用的滑动窗口的大小。另外，不同的记录经过排序后应当距离较远，这样可以避免无意义的判等比较。综合来看，排序键值的选择不仅影响了算法的整体重复检测效果，还影响了算法的时间运行效率。

表3.1 四条相似重复记录的例子

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Record | First Name | Last Name | Address | ID |
| R1 | Jack | Stolfo | 123 First Street | 12345678 |
| R2 | Jack | Stolfo | 123 First Street | 12345673 |
| R3 | Jack | Stolpho | 123 First Street | 12345678 |
| R4 | Jacon | Stiles | 123 Forest Street | 12345432 |

举例来说，表3.1是四条相似重复记录R1，R2，R3，R4。它们由四个字段组成，分别是First Name、Last Name、Address和ID。可以看出R2和R1是完全相同的两条记录，因为姓名和地址完全一致并且ID可能由于印刷或者人为录入的错误只相差了一位；R3和R1对应的现实实体也是一致的，因为两者只有Last Name字段不一致，而且两者从发音上几乎一致；然而R4和R1则是对应着现实世界的两个不同的实体，因为Last Name字段完全不一致，而且两者的地址字段Address也相差甚远。所以可以得出R1=R2 =R3R4。

在下表3.1中，假设【方案一】的键值设计者选择的排序键值由以下几个部分组成：

(1)Last Name的所有部分

(2)Address的数字部分加字母部分的前三个辅音字母

(3)ID的前三个连续数字

那么它们对应的排序键值分别如表3.2所示：

表3.2 相似重复记录及其生成的键值

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Record | First Name | Last Name | Address | ID | Key |
| R1 | Jack | Stolfo | 123 First Street | 12345678 | STOLFO123FRT123 |
| R2 | Jack | Stolfo | 123 First Street | 12345673 | STOLFO123FRT123 |
| R3 | Jack | Stolpho | 123 First Street | 12345678 | STOLPHO123FRT123 |
| R4 | Jacon | Stiles | 123 Forest Street | 12345432 | STILES123FRT123 |

则可以看出该方案下，R1，R2，对应的键值相同均为“STOLFO123FRT123”，R1和R4对应的键值不同，经过排序后R1和R2聚集在邻近位置，R1和R4则不会处于邻近的位置，只看这两条记录是不能发现问题的。R1和R3原本是相同的记录，但是由于发音的问题导致在改方案情况下两者的排序关键字并不相同，所以按照【方案一】提取出来的排序关键字进行排序，很可能会导致R1=R4无法被正确的检测出来。

接下来再看另外一种方案。假设【方案二】的键值设计者选择的排序键值由以下几个部分组成：

(1)Last Name的前三个辅音字母

(2)First Name的前三个连续字母

(3)Address的数字部分加字母部分的前三个辅音字母

(4)ID的前三个连续数字

那么它们对应的排序键值分别如表3.3所示：

表3.3 相似重复记录及其生成的键值

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Record | First Name | Last Name | Address | ID | Key |
| R1 | Jack | Stolfo | 123 First Street | 12345678 | STLJAC123FRT213 |
| R2 | Jack | Stolfo | 123 First Street | 12345673 | STLJAC123FRT123 |
| R3 | Jack | Stolpho | 123 First Street | 12345678 | STLJAC123FRT123 |
| R4 | Jacon | Stiles | 123 Forest Street | 12345432 | STLJAC123FRT123 |

则可以看出该方案下这四条记录的排序键值均为“STLJAC123FRT213”。然而从表格中可以看出： R1和R2的排序键值的相同是合理的；R3和R1对应的现实实体也是一致的，选择辅音字母可以一定程度上客服发音上的错误，R1和R3的排序键值的相同也是合理的；然而R4和R1则是对应着现实世界的两个不同的实体，但是恰巧“Stolfo”和“Stiles”连续的三个辅音字母都是“STL”，“Forest”和“First”的前三个连续的辅音字母都是“FRT”，同理关键字段中对应的ID也是巧合，这种情况下可以看出【方案二】会导致R4和与其本身并不相同三条记录R1、R2、R3处于近邻的位置，从而会导致额外的判等计算，甚至如果判等方法不合适的话则可能错误地将R4与R1、R2、R3识别为重复记录，从而影响算法的准确性。

接下来再看另外一种方案。假设【方案三】的键值设计者选择的排序键值由以下几个部分组成：

(1)Last Name的前三个辅音字母

(2)First Name的前三个辅音字母

(3)Address的数字部分加字母部分的前三个辅音字母

(4)ID的前三个连续数字

那么它们对应的排序键值分别如表3.3所示：

表3.4 相似重复记录及其生成的键值

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Record | First Name | Last Name | Address | ID | Key |
| R1 | Jack | Stolfo | 123 First Street | 12345678 | STLJCK123FRT213 |
| R2 | Jack | Stolfo | 123 First Street | 12345673 | STLJCK123FRT123 |
| R3 | Jack | Stolpho | 123 First Street | 12345678 | STLJCK123FRT123 |
| R4 | Jacon | Stiles | 123 Forest Street | 12345432 | STLJCN123FRT123 |

则可以看出该方案下, R1、R2、R3的排序键值相同，它们对应的现实实体也是一致的。R4和R1则对应不同的排序关键字。这种情况下可以看出【方案三】是最理想的一种键值选择。

综合以上的三种情况不难发现：对于特定的数据集，生成排序键值的方式优劣取决于该数据所在领域的专家经验。比较合适的排序键值才能达到较为理想的检测效果和检测效率，这就需要大量的人工操作，工作量大且影响了算法的普适性。本文3.1.2提出了一种较少专家知识依赖的方法。

### 3.1.2 字段区分度方法

衡量排序键值优劣的一个原则即为不同的记录对应的键值不同，相同的记录对应的键值理应相似或一致。对于相似重复的记录，无论排序键值以何种方式选取，理论上生成的key值是相同或者相似的，所以衡量排序键值最关键的标准是在区分不同记录时的表现，即对应现实世界中不同实体的两条记录在合适的排序键值方案下生成的key值应当不一致。所以在选择排序键值时，应当选择有足够区分度的字段来提取键值。

如何衡量字段的区分不同记录的能力，这里提出“字段区分度”的概念。若数据库中一共有N条记录，每条记录都由M个字段组成，即，对于第i个字段fieldi，它的区分度如下公式：

 公式3-1

其中，代表数据集中在fieldi字段一共有种取值，也就是将数据集按照fieldi进行聚类一共有簇。的取值介于0到1之间，值越高对应字段对于整体数据集的区分能力越大。提取排序键值之前首先计算所有字段的“字段区分度”值然后对其排序，排序键值每次选择值最大的一个字段的部分或者整体来作为键值的一部分。这里的“字段区分度”偏向于是一个相对的概念，因为相似重复记录在每个字段的内容理论上是相等的，即便是实际情况中由于印刷错误、格式不一致、人工录入等导致部分相似重复记录在某个字段表现不一致，这也是极少见的情况。因而无论数据集中的重复数据占到多大或者多小的比例，拥有较大区分度的字段的“字段区分度”值也较大。

在没有人工参与或者较少参与选取排序键值的情况下，利用“字段区分度”去选择排序键值的生成方式，有利于在排序后将对应不同实体的记录区分开来，将相似重复记录聚集到近邻位置，继而提高了算法的普适性。

## 3.2 对MPN算法中排序方式的改进

Model-based CF算法是需要学习训练的一类推荐算法的统称[23]，其中包括LFM模型[24]、Logistic回归、FM模型、贝叶斯网络算法[25]等。图3-4展示了model-based CF实现的简单流程，其中包含了和业务关联的数据源、各种特征的构造、离线的模型训练、离线和在线的推荐物品打分。



图3-4 Model-Based CF基本过程示意图

### 3.2.1 MPN的排序实现方法

Logistic模型将用户的消费行为进行一个分类，分为正样本和负样本（即二分类思想）。正样本可以是用户消费该物品，包括对物品的收藏、点击或者购买，这些对于推测用户的兴趣起到正面的作用，而针对负样本可以从用户从来没有关注或者浏览过的物品中选择，负样本的构造方式则是多种多样。构建完成之后需要提取前面得到的样本对应的特征向量，再将得到的样本放入Logistic分类器中进行学习，通过学习可以得到一个消费模型，该消费模型可以得到一个评分列表，如果评分越高就说明用户对应物品的偏好程度越高。

针对Logistic模型，假设正负样本满足如公式(3-3)的概率分布，其中y代表样本x的一个标签，可以取值正1或负1，而w、b表示对应的分布参数。

|  |  |  |
| --- | --- | --- |
|  |  | (3-3) |

通过最大似然估计方法可以得到如下的优化目标表达式：

|  |  |  |
| --- | --- | --- |
|  |  | (3-4) |

公式(3-4)中的项表示正则项， C是一个可调的常数，L表示样本数。实现上述目标可以可以采用牛顿法、拟牛顿法、共轭梯度法等优化算法，获取权重向量w。

Logistic模型主要针对特征维度数目不过高的场景，在这种场景下，模型训练得到的结果还是比较得当。但是该模型无法刻画非线性的关系，如：假设取得x维度的用户的行为特征，y维度的物品基础特征，该模型只能对x+y维度的线性特征进行描述关联，而无法得出特征之间的非线性关系。为了使Logistic模型能够在推荐系统中起到显著的作用，需要获取用户和物品之间cross的特征，对他们的基础特征进行互相关联。

### 3.2.2 IMPN算法的排序方法

LFM(Latent Factor Model)隐含因素模型利用之前已经存在的评分数据，通过分析用户对应的隐含因素（如物品包含各种因素的程度），最后得到分析结果进行预测，来填补用户的评分矩阵。补全矩阵要尽量减少对原本矩阵影响，即原本的矩阵和处理后的矩阵之间的特征值相差不能太大。为了获取用户和物品之间的隐含量，通常采用SVD的算法来实现，主要是将用户和物品的评分矩阵进行分解得到。我们可以假设有m个用户和n个物品，那么 对应的是用户的评价矩阵，则SVD算法将补全后的矩阵进行分解得到如下形式：

|  |  |  |
| --- | --- | --- |
|  |  | (3-5) |

公式(3-5)中，项和为项表示对应的正交矩阵，表示一个对角矩阵，为了得到用户和物品之间存在k个因素的关联，需要获得k个特征值（特征值越大越好）表示左右奇异向量，可以得到近似得公式如下所示：

|  |  |  |
| --- | --- | --- |
|  |  | (3-6) |

公式(3-6)中的 项表示对应的m个user的隐向量矩阵，表示对应的n个item的隐向量矩阵，通过计算向量内积可以得到用户对物品的评分。

最基本的SVD的算法具有两个缺点，一个是由于在真实的平台上一般用户只会对自己感兴趣的东西进行操作，而对于大多数的产品是不会产生关联的。所以这样就会导致在获得的评分矩阵的具有很大的稀疏性，那么针对矩阵补全就需要巨大的存储空间。另外一个缺点是该算法的复杂度太高，达到O()这样的级别，这是一个可怕的复杂度，一旦数据量太大，该算法就无法进行计算。针对这种情况，Simon Funk提出一种可以快速实现矩阵分解算法，将评分矩阵分解为两个低维度矩阵的乘积如公式(3-7)所示：

|  |  |  |
| --- | --- | --- |
|  |  | (3-7) |

Simon Funk的算法主要是通过学习这两个矩阵，获得最小化损失函数，如公式(3-8)所示：

|  |  |  |
| --- | --- | --- |
|  |  | (3-8) |

公式(3-8)中的代表用户u对于物品i的一个评价，其中 分别表示用户u对应的隐向量和物品i对应的隐向量，表示可调参数，控制正则项影响，隐向量特征取值表示为用户u对第j类物品的喜爱值，隐向量特取值表示物品i隶属于第j类物品的置信度。

后来出现的BiasSVD、SVD++和SVDFeature算法都是基于Simon Funk的算法的基础上进行改造的。而这里面SVDFeature算法可以作为LFM系列中相对具有代表性算法，该算法模型了涵盖了基本的SVD、BiasSVD和SVD++模型，公式如下所示：

(3-9)

公式(3-9)中的表示求得所有的评分的一个平均数，表示的其他的所有特征取值，项代表用户特征取值，项表示为物品特征取值，为全局特征，是用户特征、表示物品特征的偏置权重，、为用户、物品特征对应隐向量。该算法的模型如图3-5所示。

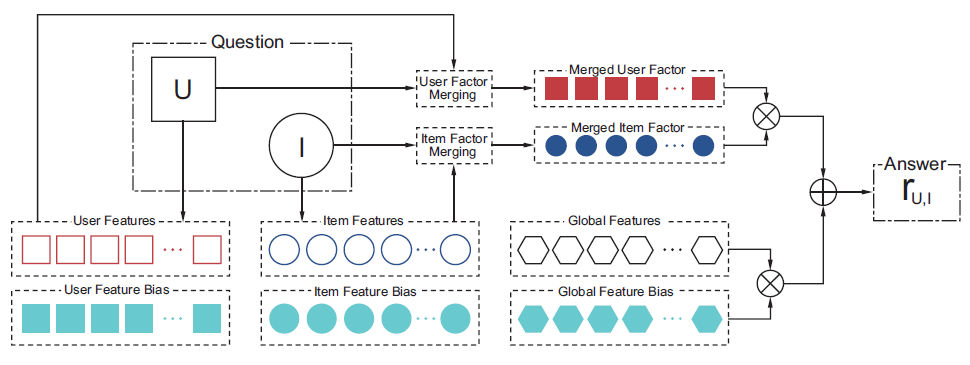


图3-5 SVDFeature模型原理示意图

SVDFeature的目标函数可以通过SGD(Stochastic Gradient Descent)随机梯度下降算法或者ALS(Alternating Least-Squares)交替最小均方误差来实现。但是SVDFeature算法具有一个缺陷，即通过学习得到的全局特征偏置项、用户特征偏置项和物品特征偏置项，他们的权重都会比较高，而他们之间隐向量的内积值很低，这很不利于进行个性化的推荐。

## 3.3 IMPN整体设计思路

由于本系统是基于物品的推荐，通过上述的介绍和分析，因此本系统采用的是基于物品的协同过滤算法。系统中采用的相似度计算公式如(3-11)所示。

|  |  |  |
| --- | --- | --- |
|  |  | (3-11) |

公式中表示喜欢物品x的用户数目，分子表示为同时喜欢物品x和物品y的用户数目。因此上面的公式可以理解为喜欢物品x的用户中有多少的比例用户喜欢物品y。

根据上述得到的物品相似度计算用户对物品的预测评分值，系统中采用的如下公式进行计算：

|  |  |  |
| --- | --- | --- |
|  |  | (3-12) |

公式里面R(u)表示用户已经评分的集合，S(j,k)表示和物品j最相似的K个物品，而表示了物品j和物品i的相似度，表示用户u对物品i的评分。通过该公式计算得到评分列表再进行排序，即最终得到的推荐结果列表。本文中采用了增量更新物品相似度和利用时间窗口来优化本系统采用的算法。

（1）增量更新物品相似度

本系统计算物品相似度是放在离线平台上进行的，当面对庞大的数据量想要通过实时计算并快速得出推荐结果是不现实的。所以将重量级的计算模块放在离线平台进行周期的批量处理，这个周期可根据数据量情况而定。同时流式计算平台实时的获得近期用户的行为，对用户行为兴趣程度进行分析。系统主要是对物品相似度进行更新计算，因为该计算公式比较简单，所以该计算规模比较小，计算更新速度比较快，采用的更新公式为（3-11）。

针对上述的公式进行分析，假设用户M 对物品x有过评分，那么相对而言物品相似度如何更新：

a. 如果用户M对物品x有过评分，即对该物品有产生行为信息，那么我们不需要更新任何物品相似度。

b. 如果用户M首次对物品X产生事件行为，对其进行了评分。那么相对的就会增加一。而通过公式中的，如果出现用户M对物品y也有过事件行为，则需要增加一；如果用户对物品y没有过事件行为，则保持不变。

（2）实时推荐的算法

实时推荐算法和离线处理的推荐算法不同之处在于，实时推荐要求其计算量要小，需要避免过多的计算导致用户体验下降。离线计算推荐物品按照公式(3-12)进行计算，主要是根据其过往的评分记录进行计算。而在线实时推荐需要保证响应速度，并根据用户当前的一段时间内最能反映用户的兴趣的原理。在实时推荐计算阶段，不会使用用户全部的评分记录，而是使用最近一段时间的行为记录通过公式进行计算，同样是根据公式(3-12)进行计算。

实时推荐算法里面是根据用户最近的行为进行推荐，那么就需要进行一个时间上的更新定义，对用户早期不能对当前行为产生影响的行为记录进行清洗。

1. 根据时间的窗口，只有在一定的时间范围内的用户行为才会被保存起来，用

来推荐计算，而在该范围外的数据将被清洗掉。

1. 根据操作行为的个数，保存一个用户的操作行为队列，如果该队列满了，则

将清洗队列最前面的事件。

本系统的实时推荐算法，主要是采用的是根据时间窗口的方法。该算法采用就是基于物品的协同过滤算法，而算法时间复杂度取决于相似度量的方法。算法采用了Jaccard的余弦公式（即：公式3-11），其时间复杂度为O（|X|\*|X|），X表示物品的数目，这适用于物品数目相对稳定的电商的平台上。而如果采用基于用户的推荐算法，那么X则对应就是用户数目，因为在电商平台上，用户的数目可能会急速得增加导致时间复杂度过大。而如果采用基于模型的协同过滤的算法，那么在实时更新计算上，几乎无法通过相应的公式计算得到推荐结果。

系统中采用相对简单的更新相似度方法，目的是为了让系统能在短时间内，对离线的物品相似度矩阵进行快速更新。对推荐列表的计算本文根据时间窗口来减少对全部用户行为计算，这样可以很好得保证了计算速率和推荐效果。所以和上述复杂的推荐算法相比，采用这种算法和公式，可以更加容易满足实时推荐的效果。

## 3.4 实验设计与结果分析

### 3.4.1 实验数据介绍

本章主要是对推荐算法中的协同过滤算法及算法流程、适应场景进行介绍。其中

### 3.4.2 SNM、MPN、IMPN综合对比实验

算法主要以相似性为核心，依赖用户的行为数据。基于模型的协同过滤算法主要

### 3.4.3 IMPN中对空键值处理方式的验证实验

Logistic模型、LFM模型和FM模型。这些模型需要对数据进行特征训练，基本过程包括：数据源的准备、特征构造、模型训练、最后得到推荐的物品，依据用户需求将物品推荐给对应的用户。最后简述了本系统采用的基于物品的协同过滤推荐算法，并利用公式对推荐结果进行实时的更新。

# 第四章 改进的遗传神经网络算法

许多的互联网用户在网上都会体验到推荐系统带来的福利。当你的一个朋友推荐给你一件商品之后，你在淘宝或者京东上面搜索这件商品，它会出现一系列返回的结果。同时，这些结果会出现推荐的类似商品，或者是列出“买了这件商品的用户可能还买了哪些产品”。如果你经常在京东或者淘宝上，那么这些平台会根据你的购买记录和操作行为为你推荐一组商品。而我们所探讨的推荐系统，就是决定把哪些商品推荐给特定的用户。

在电商平台上由于每个用户对物品的兴趣需求不同，对商品的要求不同，那么推荐给不同用户，所对应的推荐列表必然不一样，为了实现这种效果就需要运用到个性化推荐的技术。相对而言，部分现有的产品只是将最近最热门的或者最畅销的商品推荐给每一个用户，导致对不同的用户推荐并没有任何区别。这当然也可以理解为一种推荐，毕竟是一种客观的大众思维。但实际上，热门和畅销的并不一定就适合每个用户，所以推荐热销产品的作用度有限。因此，个性化推荐更加具有价值性。

而在个性化推荐的基础上，如何实时的将这些信息推荐给用户，将具有更大的价值。在市场上的如“今日头条”就是一款号称会根据用户的每次操作，在短时间内展示给用户最新的推荐结果。

## 4.1系统功能性需求

本系统为了实现电商的实时推荐系统，采用了目前比较流行的流计算框架Storm进行开发，那么要构建实时的推荐系统需要以下几个模块系统。如图4-1的功能图所示：



图4-1系统总体需求示意图

（1）日志采集系统：构建该模块的目的是为了收集大量的实时数据，这些数据主要是用户在平台上的操作信息，例如：浏览商品信息，购买商品信息，收藏商品信息等。这些数据需要被收集并且进行存储，存采集的数据需要随时提供给后面的推荐系统使用。

（2）推荐系统：这部分是本系统的核心内容，解决如何能够实时推荐的问题。该部分包含了离线平台和在线平台两个部分，离线平台主要是将存储在HDFS的大量数据进行一个离线的计算，通过基于物品的协同过滤算法周期性的（一天或者一周，根据数据量情况而定）处理过去一段时间的数据，并获得离线的物品相似度信息和物品画像信息。在流式计算平台中，主要计算实时的数据流，针对用户实时的操作，通过在线引擎进行增量计算得出最后的推荐结果。采用双平台架构实现实时推荐系统，主要是由于互联网信息的数据量的过于庞大，那么用户之前积累了大量的行为信息需要通过离线平台进行预处理，处理得出基本的物品的模型特征。而在线平台则只负责对这些特征进行增量更新计算即可得到最后的推荐结果。

（3）UI系统：本部分是将推荐结果展现出来，但是该部分具有很大的讲究，UI系统需要解决如何将推荐结果展示让用户觉得不突兀，甚至让用户觉得惊喜的问题。

### 4.1.1 日志收集系统

日志采集系统，从Web端和移动端采集数据，采集之后将数据进行存储。



图4-2日志采集系统需求示意图

现今，许多的公司的平台每天都会产生大量的信息日志，这些日志基本都是流式数据如：搜索信息、查询信息、点击信息等。而针对自己的业务，每个公司都会设计一个日志采集系统实时的收集平台上实时产生的数据，日志采集系统一般需要具有如下特点：

（1）需要建立应用系统（Web、APP等）和分析处理系统（如：推荐系统，日志分析系统）之间的通道。让实时产生的数据可以及时的被采集并可控制的传输给分析系统进行处理，这也很好的解耦了应用系统与分析处理系统的直接关联。

（2）需要满足实时在线分析系统的要求，通过实时采集数据并将这些数据实时的发送给在线分析系统进行解析。

（3）需要扩展性强，当数据量突然变大的时候（典型如：淘宝的双11），可以

通过增加节数来满足数据的涌入。

（4）需要容错性好，当出现故障的时候，需要很好的容错，避免出现问题。

（5）需要采用合适的框架来实现，这样可以方便开发并缩短开发时间，针对不同的框架需要考虑各种条件包括：开发简易性、容错性、负载均衡、可扩展性等。

根据采集系统的需求，本文对目前的一些开源的采集框架进行了探讨和研究，得到采集框架比较表如下：

表4-1采集框架比较表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 框架 | Scirbe[26] | ChuKwa[27] | Kafka | Flume |
| 公司 | Facebook | Apache/yahoo | LinkedIn | Cloudera |
| 开源时间 | 2008.10 | 2009.11 | 2010.12 | 2009. 7 |
| 实现语言 | C/C++ | JAVA | SCALA | JAVA |
| 框架模式 | Push/Push模式 | Push/Push模式 | Push/Pull模式 | Push/Push模式 |
| 容错性 | 具有一定的容错机制，不过容错机制部分需要用户自己实现。 | Agent会记录数据偏移量。一旦出现问题，可根据记录进行恢复。 | 该框架的容错机制和ChuWa类似，也是记录了数据偏移量。 | 该框架组件提供了三种级别的机制来保证容错性。 |
| 负载均衡 | 无 | 无 | Zookeeper | Zookeeper |
| 可扩展性 | 好 | 好 | 好 | 好 |
| Agent | 需要开发者自己实现 | 本身已经自带一些agent。 | 用户需根据实现Kafka提供的组件，进行对应的开发。 | 提供了各种非常丰富的agent，可以直接使用。 |
| Collector | 可以理解为是一个简单的存储服务。 | -- | 使用了各种技术提高性能能。 | 系统本身提供很多的存储层，且可直接使用。 |
| Store | 支持HDFS | 支持HDFS | 支持HDFS | 支持HDFS |
| 评价 | 该框架设计相对比较简单，且也容易使用，但是没有实现负载均衡，自身的容错性不够好。 | 该框架主要是针对比较大型的采集系统。 | 该框架的设计非常的巧妙，采用了push和push模型。 | 在采集和输出方面都有组件可以直接使用，开发相对方便简单。而且本身容错性好。 |

日志采集系统的开发可以利用已经存在的框架进行开发，相对而言自己去造轮子，这种基于成熟的框架来实现开发是比较简单的。而经过比较之后，本论文更倾向使用Flume架构来进行开发。

### 4.1.2 推荐系统

随着推荐系统的日益发展，人们对它的要求也日益增长，那么实现推荐系统需要考虑哪些方面才能满足用户的体验。如图4-3，我们来逐一讨论推荐系统的需求。



图4-3推荐系统需求示意图

（1）满意度：这是推荐系统中最重要的一个评判标准，因为推荐系统本身就是为了推荐给用户最满意的物品，所以推荐系统里面的推荐准确率是衡量推荐系统最重要的一个因素。

（2）实时性：本文需要实现实时的推荐系统，所以实时性需要重点考虑。随着技术的发展，实时性的推荐已经越来越成为主流。其中实时性体现在短时间内，系统可以根据上下文信息，或者用户的操作行为来快速得出用户模型，然后针对不同的用户模型对推荐列表进行更新。毕竟有些用户的爱好和兴趣点总是容易变化，那么推荐内容就需要实时的更新来满足用户的需要。

（3）新颖性：推荐内容如何让用户觉得特别新鲜，一种推荐方法是推荐给用户从来没有见过的物品，当然这种物品又需要让用户有点兴趣。如：用户A在电商平台上面见到 “购买过该物品的用户也购买这些产品”，那么如果用户B也购买过同一个产品（或者多个相同的产品），可以得到用户A和用户B的具有共同的爱好，用户B所购买的物品很有可能也符合用户A的需求，那么再去掉用户A之前购买或者浏览过的物品，这样推荐给用户A的效果就是就会非常具有新颖性。

（4）惊喜度：和上述的新颖性的基本上是类似。相对新颖性，惊喜度的概念更加简单，只要推荐给用户想要的产品，让用户觉得系统可以洞悉自己的内心的需求，惊喜度和新颖度的区别是推荐集合不需要排除掉自己之前浏览过的物品。

（5）多样性：推荐系统中，推荐的多样性作为一种兼顾的需求。在满足用户满意度的前提下，推荐的物品越丰富多彩，那么给予用户选择也就更大。推荐多样性的物品，需要根据推荐物品权重比值进行推荐。其中推荐排序，都和该权重有关系，权重越大，推荐顺序越靠前，反之越靠后。

（6）透明度：这个需求是一种友好的用户体验，让用户知道为什么推荐这些物品给他们。如：我们经常在电商平台上看到“买过这件衣服的人曾经也买过如下产品”、“你现在购买的鞋子和如下的产品类似…”。这样给予用户一种透明推荐的感觉，让用户明白推荐的物品的依据，让用户体验得到提升。

（7）覆盖率：推荐系统推荐的内容需要覆盖大量的内容信息，而这些信息就需要进行挖掘物品的长尾内容。一般评判标准是推荐概率分布的物品占总物品的比例和所有物品被推荐的概率进行计算。

关于推荐系统的方式，推荐方式多种多样，为了达到上诉的推荐效果，在理论上可以归结为以下几种推荐方式，如图4-4所示。



图4-4 推荐方式组成示意图

（1）热门推荐：顾名思义是热门排行，最典型的莫过于微博热搜，根据用户点击量来判断微博的热度，进行一个热门排行展示。如：音乐热门排行、书籍热门排行，这种方式在生活中无处不在。作为一种大众的推荐方式，按曝光率排序物品展现给用户，大多数从众心理的用户会被吸引去浏览这些物品。这种推荐也具有良好的效果。

（2）人工推荐：这是一种最原始的方式，但也很有效。在没办法建立特定模型，就需要依靠人工的方式进行推荐，这种方式很依靠推荐人对该领域积累的经验。还有关于热门实时产生的新闻，内容信息一开始需要人工加入推荐列表，这样才会引起更多的关注度。

（3）个性化推荐：这种推荐方式具有独特性，主要是根据用户之前做过的行为来建立用户兴趣模型和其他的关系模型，计算得到各种不同的推荐列表。因此每个用户的推荐就会都不一样，这和热门推荐从众心理完全不一样。

（4）相关推荐：通过用户目前的行为状态和兴趣爱好关联相关的内容的进行推荐的方式。如：用户浏览一本书的某章节内容的时候，那么系统会提示目前用户浏览与此书籍相关的内容。这种推荐和个性化推荐的关联规则基本上是一致的。

本文讨论的系统主要是针对个性化推荐的基础上的应用，因为个性化的推荐具有独特性，同时可以解决信息过载问题，让用户感受到推荐系统的强大性。目前个性化推荐的算法主要有：

（1）基于内容的推荐[28][29]：通过推荐内容本身的信息来进行推荐，如内容关于历史信息的，那么就会推荐关于历史的一些内容。

（2）基于关联规则的推荐：该推荐方法需要进行大量的数据挖掘得到这种推荐物品的关联，最出名的就是“尿布和啤酒”的案例，看似没有什么关联的两个物品，通过关联挖掘得到联系。

（3）协同过滤推荐：上述的章节对此算法已经做了较详细的研究，主要是Item-CF、User-CF、Mode-CF等几种推荐算法，通过该算法会向不同的用户推荐可能感兴趣的商品[30]。该算法主要是通过获取用户的行为数据并进行长期的积累，然后进行各种模型分析得到用户的个性化的特点，最后得出推荐结果的一种特色推荐。当然也存在各种复合的算法，如协调过滤和基于内容的结合算法[31]。

那么，本文重点讨论关于个性化的推荐算法的基本需求，那么下面具体的阐述个性化推荐算法的流程的各个步骤。

（1）获取日志信息：关于用户的行为信息，如：浏览某个物品，点击某个操作等。这些大量的行为信息都会被存储起来，主要是依靠日志采集系统来进行收集，并进行存储，一般存储在HDFS中。

（2）前期的ETL（Extract-Transform-Load）数据：采集用户最原始的行为数据之后，需要对原始的数据进行清洗，主要是将后期不需要的信息进行过滤掉，这样可以减少大量的数据，同时也避免脏数据或者无用数据对后面计算存在干扰。可以将数据按照后面要求的格式进行加工如：字段要求、属性要求、格式化等，这样可以满足算法对数据的直接调用。

（3）推荐算法：选用什么推荐算法，要根据不同的业务要求来进行选择和设计，目前已经存在各种各样的算法可以运用。第三章节已经对协同过滤算法进行了仔细的探讨和研究，这边不再重复阐述。

（4）后期的ETL数据：该数据是经过推荐算法处理之后的信息，不过需要将这些数据进行分离处理，因为在算法计算过程并不只是产生我们要的推荐内容，还会产生很多关于用户或者物品的偏好矩阵，如：用户相似度、物品相似度、内容信息的关联等。

（5）画像信息：该画像信息包含算法计算过程中得到的用户兴趣和操作的行为信息。这些信息可以重复利用，不需要每次都重新计算得到。当然这些爱好行为会随着时间的向前，画像信息也需要进行局部的更新或者完全更新。

（6）推荐列表：按照要求存储得到的推荐结果，如：用户对应推荐列表，物品对应的推荐列表。系统中对于这些数据的存储的容量和读写性能要求都比较高，可以选择使用Redis集群作为此部分数据的存储方案。

（7）API服务接口：这个是暴露给外部的调用接口，满足开发者的调用，开发者可以调用得到的推荐结果列表，也可以调用用户或者物品的的画像信息。

通过上述的流程阐述，具体的算法流程如图4-5所示；



图4-5 个性化推荐算法流程图

个性化推荐算法的几个重要流程，包括对数据的多次清洗处理、算法的择取，画像信息和结果的存储，向外部提供接口。这些步骤在不同的业务中，采用的策略流程稍微会有所不同，不过基本都是大同小异。

### 4.1.3 UI系统

关于UI的展现，呈现的方式不同，给用户的体验也会不同。针对电商平台，在购买物品的时候，为了给用户良好的用户体验，会在用户购买的时候提供一些针对该用户的特色推荐，而这些推荐，会列出原因，让用户体验到系统推荐的透明度，如：你可能喜欢这些产品、猜你喜欢等。

虽然UI系统并不是本论文阐述的重点，但是作为一个项目产品，UI如何展现有时候会决定一个产品成败，因此在这边特别的列出来做一个简要的说明。

图4-6是京东平台推荐的UI展现。通过“猜你喜欢”的模块来进行个性化的推荐，这里面的推荐确实存在我最近关注的手机和VR眼镜，整体上的效果还是不错的。



图4-6 UI需求示例图

## 4.2本系统的重点需求

上面已经很好的阐述了关于本系统的几个大模块需求，下面再重点的讨论一下本系统需要注意到的重点需求问题。因为本论文实现的是实时的推荐系统，所以下面会讨论本系统中重点关注和需要解决的点。

### 4.2.1 实时性问题

对于本系统，实时性是至关重要的一个点。也是本系统需要解决的难点。为了实现实时性，本系统引入了Storm框架进行实时的计算，但是在数据量巨大的情况下，Storm的实时计算也很难快速得到我们需要的结果。因为即使采用常用的协同过滤算法，通过公式得到用户之间的关系矩阵、物品之间的关系矩阵。如果维度太大，则需要进行矩阵分解，这些计算都是比较复杂耗时的。一旦数据量太过庞大，要在短时间内得出推荐结果，基本是天方夜谭。

那么如何解决这种情况。本系统采用将复杂的计算放到离线系统进行计算，通过采用协同过滤算法离线批处理的计算用户相似度、物品相似度等，并这部分的内容当做画像信息进行存储，在Storm上利用画像数据进行一个增量计算得出最后的推荐结果。

### 4.2.2 冷启动问题

推荐系统中都需要面对关于冷启动的问题，冷启动的概念是一个新的用户，在系统中没有任何关于他的信息，那么这时候该用户找不到任何和他有关联的数据，这样就会导致系统无法给他推荐物品。同时由于被推荐内容也会不断的产生和更新，那么又如何将这些新的内容信息推荐给用户，也是一大问题。所以冷启动问题可以归纳为两点：新用户的推荐和新内容的推荐。

关于新用户的推荐，我们在使用任何平台都需要进行注册，才能拥有属于自己的账户。例如：我们在网易音乐上面注册一个用户，那么网易APP端会展现一个页面，上面会列出好几种音乐风格（古典、流行、轻音乐等），让你选择音乐爱好，这本身就是一种规避冷启动的方法，可以很直接的得到用户的喜爱的歌曲类型。当然有些用户会不喜欢去做这些调研，会直接忽视，点击跳过。那么最简单的方式就是采用热门推荐，这是一种最简单的处理方式，把最热门的信息推荐给用户也会取得意想不到的效果。

关于新内容的推荐，即这个内容才刚产生而且不属于那种特别爆炸的新闻（如果是比较爆炸的信息，会直接被人工推荐直接置顶），是否要将该新闻推荐给用户。这部分就需要特殊处理，可以通过时间的阈值来进行控制传播。

### 4.2.3 时效性和多样性问题

时效性主要是针对和时间有强关联的内容。例如：新闻内容、时事内容等。可想而知，如果在用户的推荐首页出现一条关于去年的新闻消息，如：2016年NBA冠军是骑士的内容。这样的推荐结果会让用户觉得莫名其妙，让用户对该推荐产生失望。所以，关于时效性的内容推荐问题需要单独处理，主要是根据“时间”来做影响因子。

多样性的概念本质就是得知用户的多种兴趣爱好，推荐结果可以包含着多种兴趣爱好，而不只是推荐一种类型，这样就不会显的单一。其实这个点很好理解也很好解决，就是在分析用户的行为的时候，可以得到多种的兴趣特征，那么推荐的时候可以加入多个用户的兴趣特征（如：该用户爱好体育，也喜欢电玩。），那么可以根据两个特征来计算来得到用户会被推荐的信息，这样的策略就可以丰富推荐列表。一般推荐的数目可以根据用户和相关度来度量，例如：用户的标签有t1、t2、t3、t4， 对应的相关度为w1、w2、w3、w4，总的推荐数目为TotalNum，则：List(u)=T1[TotalNum \* w1] + T2[TotalNum \* w2] + T3[TotalNum \* w3] +T4[TotalNum \* w4]。

## 4.3 本章总结

本章主要是描述推荐系统的需求，包括实时性、满意性、惊喜度、多样性、推荐透明和新颖性。具体的阐述了系统的整体上的需求，对需求的每个部分进行解析。同时突出了本系统中要实现的实时推荐系统里面实时行的重要性，当然本文主要解决了系统里面的实时性和冷启动问题，另外两种问题只做讨论。

# 第五章 Fork/Join并行框架在IMPN算法中的应用

本章主要是阐述推荐系统的设计和实现，主要包括系统的整体框架，及在框架的基础上系统模块的的设计和模块和模块之间的衔接，最终通过简单的例子阐述系统采用的基于物品的协同过滤算法流程及原理。

## 5.1 整体框架

 图5-1实时推荐系统的总体结构示意图

本系统的整体架构分为如下几个大部分：

（1）日志采集系统：如图5-1所示，该模块系统需要实时的收集用户大量的操作信息，如：收藏物品信息、购买物品信息等，系统将这些数据收集之后进行存储。存储之后的数据流向两个方向，一个方向是HDFS，提供给离线计算平台进行算法处理。另一个方向则是流向缓冲模块，提供给流式计算平台进行实时的推荐。

（2）推荐系统：作为核心部分，推荐系统主要包括离线平台和在线平台，离线平台主要是将大量数据进行离线的计算，该模块被设计为日处理历史的数据，并提取离线特性，即物品的相似度信息。在流式计算平台，主要计算实时的数据流，获取用户时刻的操作，通过在线引擎对物品的相似度矩阵进行更新，再实时计算得到最后的推荐结果。这种设计是比较符合实时推荐的，由于互联网信息的数据量的庞大，用户历史数据信息需要一个离线平台进行处理，计算出基本的用户模型特征。之后短时间内出现的数据，则可以通过一个流式计算快速处理，更新模型，进行最后的推荐。

（3）输出展示系统：该系统模块，也算是比较重要的模块，该系统最后通过开发的一个APP来进行展示，给予用户更加直接的体验。

本论文主要的重点在于日志采集系统和流式计算平台的设计，如图5-2简述了流式计算平台的流程图，这个流程主要是对数据进行采集，然后将采集数据传输到数据接入模块，从而起到缓冲的作用。将传输的数据存储起来等待流式计算模块来拉取数据，流式计算模块实时处理信息，处理完之后再从上个模块拉取，实现即拉即用的模式，最后将数据结果存储并暴露接口供外部调用。



图5-2流式计算流程图

如图5-3具体的阐述了整个系统的详细的数据流向过程，我们可以从该示意图将系统分为为五个模块：

（1）数据采集模块：本模块采用了Flume框架进行开发，图中可以看到日志采集模块从各个不同的平台上通过多个Agent采集实时产生的数据，接着会将数据传输给Collector，最后Flume的Sink一个是HDFS，另一个则是下一个模块Kafka。

（2）数据接入模块：本模块采用Kafka框架，图中所示，Kafka的Produce是Flume的Sink，通过编写Flume的配置文件，可以很容易实现两者之间的结合。实现该模块的目的主要起到一个数据缓冲的作用，当Flume采集的数据量太大导致Storm实时计算模块无法及时处理，则会产生数据堵塞。因为Kafka采用Push/Pull的模式，即采集的消息是推入进去的，而在下一个的流式计算模块采用拉取数据的模式，流式计算模块可以根据自己的计算情况，把目前可以承载的数据拉取过来进行计算。这种方式可以很好得避免出现由于处理不及时导致的堵塞的情况，而且采集信息会存储在Kafka的Broken中，数据会一直存储7天（默认7天，也可以自定义），保证数据不会丢失。

（3）离线计算模块：本模块是基于Hadoop平台上进行计算，可以得到物品的相似度和一个推荐列表。为了实现实时的推荐，同时也要保证推荐的准确性，系统设计就将存储在HDFS的数据采用线下的基于物品的协同过滤算法进行处理，根据数据量该过程设计为日处理。

（4） 流式计算模块：本模块采用Storm来实现，是实现实时推荐的关键，这边处理实时得到的数据，同时结合离线模块提供的画像信息，对物品的相似度进行增量更新，最后加入时间因素，对推荐列表结果进行增量更新，从而实现实时的推荐效果。

（5）数据存储和输出模块：对计算过程得到的数据采用Redis存储，为方便外部接口的调用将最后得到的推荐列表结果存储到Mysql持久化数据库中。

图5-3实时推荐系统的数据流向示意图

## 5.2 日志采集模块

在该系统上，用户通过点击和浏览会产生一定的行为。如：浏览商品、点击收藏、购买产品等。这些行为都会产生对应的数据保存起来。

### 5.2.1 日志数据类型

日志数据主要包含这几类的日志内容：用户浏览的信息记录，用户检索信息记录、用户收藏信息记录，用户购买信息记录。下述列出应用后台的表设计：

（1）用户浏览的信息日志是在用户点击产品时候产生，日志数据字段包含：用户ID信息、点击的商品ID、商品类别、浏览时间。

表5-1用户浏览信息表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 字段长度 | 是否主键 | 说明 |
| UserID | Int | 20 | 是 | 用户的ID |
| ItemID | Int | 20 | 否 | 商品的ID |
| ItemType | Int | 20 | 否 | 商品类别 |
| Time | String | 30 | 否 | 浏览时间 |

（2）用户检索信息日志是在用户输入检索关键字点击搜索的时候产生，日志数据字段包含：用户ID、检索的关键词，检索到商品类别ID子串、检索时间。

表5-2 用户检索信息表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 字段长度 | 是否主键 | 说明 |
| UserID | Int | 20 | 是 | 用户的ID |
| SearchKey | String | 100 | 否 | 检索的关键词 |
| SearchInfo | String | 1000 | 否 | 检索到商品类别ID子串，ID之间用空格隔开。 |
| SearchTime | String | 30 | 否 | 检索时间 |

（3）用户收藏信息日志是在对商品进行点击收藏操作的时候产生，日志的数据字段包含：用户ID、收藏的物品ID，收藏物品的时间。

表5- 3 用户收藏信息表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 字段长度 | 是否主键 | 说明 |
| UserID | Int | 20 | 是 | 用户的ID |
| CollectID | Int | 20 | 否 | 收藏物品的ID |
| CollectTime | String | 30 | 否 | 收藏时间 |

（4）用户购买记录信息日志是在用户点击购买操作的时候产生，日志数据字段包含：用户ID、购买的物品ID、购买物品的时间。

表5- 4用户购买记录信息表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 字段长度 | 是否主键 | 说明 |
| UserID | Int | 20 | 是 | 用户的ID |
| ConsumeID | Int | 20 | 否 | 购买物品的ID |
| ConsumeTime | String | 30 | 否 | 购买时间 |

### 5.2.2 日志采集系统框架



图5-4 日志采集系统框架示意图

系统平台上的数据通过日志采集系统来收集，如图5-4，日志采集系统采用Flume来构建，对多个Web Agent进行监听，将产生的日志数据源发送给Collector，Collector会把全部的数据都加载到存储端（Storage），存储端会将数据一份保存在HDFS，另外一份则会发送给Kafka。而HDFS部分主要的数据供离线平台使用，而发送给Kafka的则是供实时处理的流式平台的数据。

（1）Agent：收集平台上实时产生的数据并推送给Flume里面的Collector模块，这里可以支持自己定制来满足要求，当然本文只是简单得利用了该功能将数据收集。

（2）Collector：该模块主要进行存储和转发，保证数据不轻易丢失，负责将推送过来的数据发送给Kafka模块和HDFS进行存储。从图中可以看到，两部分数据流向不同平台。

（3）Master：为了保证在分布式的情况下，可以对组件进行管理，主要是针对出现故障的时候，不让数据丢失，原则上可以设置多个Master，为简化系统的开发本系统只设置了一个。

（4）ZooKeeper：维护上述几个组件，保持数据一致性，在数据发生变化时候，及时通知master。

### 5.2.3 日志采集系统连接器

这部分主要是配置Flume的环境变量同时采用Flume提供的组件进行收集数据。 （1）配置环境变量

Flume该框架已经实现了很多功能，开发过程只需要根据系统的需求来进行配置，配置主要包括了数据源（Source）、存储管道（Channels）和数据目的地（Sink）。配置文件flume-conf.properties是整个工作的核心，是满足个性化系统定制的支撑点。

数据源包括的配置方案有exec、avro、netcat、systcp、sysudp、seq等。

存储管道的目的是避免数据可能发生丢失的情况，让数据高可靠，一般设置的模式是基于内存和硬盘的。对应的配置类型memory（内存）、file（文件）、spillablememory （文件内存结合）等。当然也可以自己定义。一般为了保证足够的安全性和稳定性，最好都是基于硬盘存储，这样可以支持断点续传。

数据目的地（sink），配置方案有Logger（日志）、HDFS、HBase、AsyncHBase、MorphlineSolr、Custom ElasticSearch等。

下列表格列出不同的采集方式的配置方案：

表5-5 Flume配置方案表

|  |  |  |
| --- | --- | --- |
| 采集方式 | 配置文件 | 描述 |
| Exec | a.sources = r  a.sinks = k  a.channels = c  a.sources.r.type = exec  a.sources.r.channels = c  a.sources.r.command=tail-F /home/hadoop/flume-1.6.0-bin/log\_exec\_tail  a.sinks.k.type = logger  a.sinks.k.channel=ca.channels.c.type =memory  a.channels.c.capacity = 1000  a.channels.c.transactionCapacity = 100 | a.sources.r.type= exec  采集方式是通过命令执行（从Unix读取tail -F）  sinks.k.channel  =ca.channels.c.type = memory  上述表示将数据放在内存，可以加快速度，减少IO时间，但是也容易导致内存溢出。 |
| avro | a.channels = c  a.sources = r  a.sinks = k  a.channels.c.type = memory  a.sources.r.channels = c  a.sources.r.type = avro  a.sources.r.bind = 192.168.0.23  a.sources.r.port = 8080  a.sinks.k.channel = c  a.sinks.k.type = logger  a.channels.c.transactionCapacity = 100 | 使用Avro Netty  RPC协议，如果采用avro source可以提高线程数目，从而会提高处理的速度。 |
| Spool | a.sources = r  a.sinks = k  a.channels = c  a.sources.r.type = spool dir  a.sources.r.channels = c  a.sources.r.spoolDir= /home/hadoop/flume-1.6.0-bin/logs  a.sources.r.fileHeader = true  a.sinks.k.type = logger  a.sinks.k.channel = c  a.channels.c.type = memory  a.channels.c.capacity = 1000  a.channels.c.transactionCapacity = 100 | Source spool，适合静态文件，即文件本身不是动态变化的。 |
| Syslogtcp | a.sources = r  a.sinks = k  a.channels = c  a.sources.r.type = syslogtcp  a.sources.r.port = 9999  a.sources.r.host = localhost  a.sources.r.channels = c  a.sinks.k.type = logger  a.sinks.k.channel = c  a.channels.c.type = memory  a.channels.c.capacity = 1000  a.channels.c.transactionCapacity = 100 | Syslog 支持TCP和UDP等2种模式，而这边采用的tcp的模式。 |
| JSONHandler | a.sources = r  a.sinks = k  a.channels = c  a.sources.r.type= org.apache.flume.source.http.HTTPSource  a.sources.r.port = 8888  a.sources.r.channels = c  a.sinks.k.type = logger  a.sinks.k.channel = c  a.channels.c.type = memory  a.channels.c.capacity = 1000  a.channels.c.transactionCapacity = 100 | JSONHandler采集模式是通过Http。 |

从上述的表格中可以看到各种不同的采集方式的配置方案都大不相同，而我们这边的Source采用JSONHandler和Spool的方式，直接从Http进行响应。而Sink主要有两步，第一步数据流向Kafka，另一步数据流向HDFS进行存储。因此本系统需要设置两个数据的目的地（sink），一个是kafkaSink，另一个是hdfsSink。配置代码如下所示：

|  |
| --- |
| a.sources = r  a.sinks = kl k2  a.channels = c1 c2  a.sources.r.type=org.apache.flume.source.http.HTTPSource  a.sources.r.port = 8888  a.sources.r.channels = c  a.sinks.k1.channel = c1  a.sinks.k2.channel = c2  a.channels.c.type = memory  a.channels.c.capacity = 1000 |

（2）Flume采集的数据流向



图5-5 日志采集数据序列图

图5-5表示了Flume数据采集的序列图，从用户通过操作外部应用产生信息数据，这些信息会通过Https或者File的方式被Flume的Source进行采集，这里面采用JSONHandler的Http方式采集。为了保证数据不丢失，数据从source开始，采用了事务的机制。开启事务之后，传输的数据都会被封装为一个个事件(即：byte数组)流向channel，最后再将数据推送给Kafka进行后期处理以及HDFS进行存储。下面列出如何将数据message从Source放到Channel的代码过程：

|  |
| --- |
| ChannelProcessor chan = source.getChannelProcessor();  Event en = EventBuilder.withBody(message.getBytes());  channel.processEvent(en); |

发送到Sink（即Kafka组件）代码如下：

|  |
| --- |
| Channel chan = getChannel();  Transaction tran = chan.getTransaction();  tran.begin();//开始事务  List<KeyedMessage<byte[], byte[]>>infolist= Lists.newLinkedList();  for (int i = 0; i < infoSize; i++) {  Event en = chan.take();//获得一条日志信息  byte[] ens=event.getBody();  infolist.add(new KeyedMessage<byte[], byte[]>(topic, ens));  }  producer.send(infolist);//把信息发送给kafka  tran.commit();//提交事务 |

## 5.3 数据接入模块

### 5.3.1 数据接入框架

该系统上应用了Kafka框架来接收来至Flume采集的信息，信息数据主要是从Flume获取。而事实上Flume的Sink就是配置为Kafka，Kafka在系统中发挥着相当重要的作用，它主要包括以下的几个方面作用：



图5-6数据接入模式架构示意图

（1）将各个不同的日志进行汇总，让日志得到统一的处理。如图5-6所示从Flume Sink将采集的数据推送给Kafka，这些数据都会被集中存储到Broken节点中。所有后面模块需要处理数据只需要从Kafka的Broken中拉取出来，加入该模块方便后期对日志的处理。

（2）该模块可以减小后面流式平台中Storm的处理压力。在系统中Storm的Spout作为一个Kafka的消费模块，当数据量很大，利用传统的推送数据的方式一直发送数据过来，很容易遇到数据堵塞，让系统陷入瓶颈。而Kafka中里面的Consume（消费者）即：系统的Storm模块，采用从Broken拉取数据的模式。这样的话Storm就可以很恰当的利用自己的空闲时间去从Kafka的Broken中拉取数据。同时Broker数据默认可以缓存7天，这边直接采用默认方式，当然也可根据具体的需要进行配置。

（3）该模块设计可以很好的让数据不会丢失掉。由于Kafka本身就具有比较好的容错性的，因此当一个节点发生问题，另一个节点可以保证正常的工作，从而可避免出现各种问题。

### 5.3.2 数据接入连接器

从图5-7中可以看出Flume集群最后的数据目的地是Kafka，本论文的系统数据一部分是给HDFS，另一部分则是同时发给Kafka。那么Kafka需要接受来自Flume的数据，这个数据可以表示为Kafka的Produce，所以Flume的Sink和Kafka的Produce是连接重合的，那么系统在配置变量同时编写对应的Topic将数据合理的传输到Kafka并让Kafka消费掉，即再发送给下一个Storm框架进行实时处理。



图5-7 Kafka生产者示意图

（1）配置Flume环境变量

Flume框架里面的flume-conf.properties可以配置，将Flume里面的Sink输出设置为Kafka的Produce，一般配置就是配置flume-ng里面client的source、channels、sink。下面给出一个配置例子，里面的sink的type设置为提供的Kafkasink，如图5-8的配置文件示例。

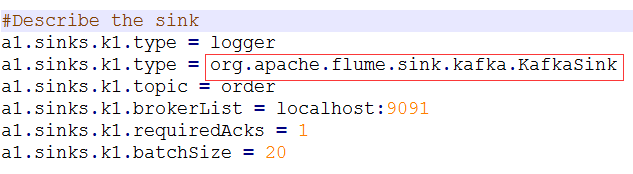


图5-8 Flume的配置示例图

（2）Kafka的连接器

系统中关于Kafka的连接器，需要编写Kafka对应的Producer，Producer本质上就是一个生产者，代表消息的生产者，也是和Flume里的Sink对应。Producer通过配置为Flume的Sink，将该消息内容推送给Broken里面进行缓存，该缓存时间使用默认为7天的配置值，并且等待着Kafka的Consumer来拉取数据，这边的消费者配置为Storm的Spout，关于生产者推送消息内容给Broken的编码如下：

|  |
| --- |
| public Producer extends Thread {  public void sendData() {  Properties props =  PropertiesParser.getProperties(PropertiesSettings.PRODUCER\_FILE\_NAME);  ProducerConfig con = newProducerConfig(props);  Producer<String, String> pd= new Producer<String, String>(con);  KeyedMessage<String, String> msginfo;  String ip = "192.168.1.11";  StringBuffer sb = newStringBuffer();  …..  pro.send(msginfo);  }  public void run() {  sendData();  }  } |

下面简要列出关于消费者从Broken结构中拉取数据的代码，主要是通过监听然后循环拉取数据。

|  |
| --- |
| public class ConsumerCluster extends Thread {  private ConsumerConnector consumer;  private String topic;  public Consumer(String topic) {  consumer =kafka.consumer.Consumer.createJavaConsumerConnector(  createConsumerConfig());  this.topic = topic;  }  private static ConsumerConfigcreateConsumerConfig() {  Properties props = new Properties();  props.put("zookeeper.connect", KafkaProperties.Connect);  return new ConsumerConfig(props);  }  public void run() {  Map<String, Integer> TopicMap = new HashMap<String, Integer>();  TopicMap.put(topic, OrderInfo);  Map<String,List<KafkaStream<byte[], byte[]>>>  CountMap =consumer.createMessageStreams(TopicMap);  KafkaStream<byte[], byte[]> st = CountMap.get(topic).get(0);  ConsumerIterator<byte[], byte[]>it = stream.iterator();  while (it.hasNext()) {  //获取数据…  }  } |

## 5.4 离线计算平台模块

离线平台主要是基于Hadoop对数据进行批处理，由于数据量的庞大，如果把协同过滤算法放在流式平台上，必然会加重流式计算的负担，无法实现实时的效果。所以，构建离线平台的目的是将之前的积累的数据进行线下处理。

### 5.4.1 基于物品的协同过滤算法原理

该小节具体阐述本系统的离线模块中采用的基于物品的协同过滤算法的实现过程。首先关于数据集的字段如下：

（1）UserId：用户的ID；

（2）ItemId：物品的ID；

（3）Score：用户对该物品的评分；

下列列出Item-CF的实现流程：

（1）先计算获得物品的同向矩阵，即计算两个物品同时出现的次数，目的为了得到物品之间的关系度。

数据格式：ItemId1:ItemId2 Num(出现次数)

（2）根据用户对物品进行一个评分（评分机制可以自己建立），得到一个评分矩阵B，这个评分矩阵的每一项就是每个用户对每个物品的评价。

数据格式：ItemId  user\_id:Score

（3）最后将第一步得到的物品同向矩阵乘以第二步的评分矩阵，就可以得到最后的推荐列表。

数据格式：userId  ItemId,Scores（推荐分值）

（4）得到推荐列表需要将用户之前已经评分过的物品（可以理解为购买过或者收藏过该物品）过滤掉，保证推荐的新颖性。

（5）最后根据推荐的分值按照倒排的取得Top（N）。

### 5.4.2 算法实现过程

根据上述的基于物品的协同过滤算法原理，通过搭建Hadoop平台之后，再编写MapReduce来实现该算法，该推荐算法实现过程如图5-9所示。



图5-9 推荐算法计算过程流程图

（1）从HDFS读取日志信息数据，然后对其进行处理。如：数据集里面没有评分的概念，那么系统进行构造评分机制，当用户点击浏览该物品评分设置1分，点击收藏则设置为3分，购买设置为5分（针对不同的情况进行评分机制的构建）。

（2）经过上述的机制处理之后，就可以获得 UserId,ItemId,Score 的数据集如表5-6所示。

表5-6 数据源表

|  |
| --- |
| 数据(UserId,ItemId,Score） |
| 1,11,5.0 |
| 1,12,3.0 |
| 1,13,2.5 |
| 2,11,2.0 |
| 2,12,2.5 |
| …… |

（3）如图5-9，编写MapReduce的计算程序，第一步读取处理得到的数据，按照用户ID分组，输出的格式如表5-7所示：

表5-7 物品评分表

|  |  |
| --- | --- |
| 用户ID | 数据（ItemID:Score） |
| 1 | 13:2.5,11:5.0,12:3.0 |
| 2 | 11:2.0,12:2.5,13:5.0,14:2.0 |
| 3 | 17:5.0,11:2.0,14:4.0,15:4.5 |
| 4 | 13:3.0,16:4.0,14:4.5,11:5.0 |
| 5 | 11:4.0,12:3.0,13:2.0,14:4.0,15:3.5,16:4.0 |
| …… | …… |

（4）通过上述的信息可以计算两个物品同时出现的次数，得到结果如表5-8所示：

表5-8 物品出现次数表

|  |
| --- |
| 数据(ItemId:ItemId Num） |
| 11:11 5 |
| 11:12 3 |
| 11:13 4 |
| 11:14 4 |
| 11:15 2 |
| …… |

（5）根据物之间出现的次数表，通过mapper1计算得到物品同现矩阵，结果如表5-9所示：

表5-9 物品同现矩阵表

|  |
| --- |
| 数据(ItemId: ItemId Num） |
| 11:11 5 |
| 11:12 3 |
| 11:13 4 |
| 11:14 4 |
| 11:15 2 |
| 12:11 3 |
| …… |

根据上述的信息可以计算得到用户评分矩阵和物品的同现矩阵，通过mapper2生成用户的评分矩阵，结果如下表5-10所示：

表 5-10 物品-用户评分矩阵表

|  |
| --- |
| 数据(ItemId UserId:Score） |
| 11 2:2.0 |
| 11 5:4.0 |
| 11 4:5.0 |
| 11 3:2.0 |
| 11 1:5.0 |
| 12 2:2.5 |
| …… |

（6）根据上面得到的两个矩阵，可以做矩阵乘法，那么就会得到推荐列表。公式是：物品同现矩阵 \* 用户的评分矩阵 = 推荐列表。推荐结果如表5-11所示：

表 5-11 推荐结果表

|  |  |
| --- | --- |
| ID | 推荐结果（ItemID,Scores） |
| 1 | 17,10.5 |
| 1 | 16,18.0 |
| 1 | 14,33.5 |
| 2 | 17,11.5 |
| 2 | 16,20.5 |
| …… | …… |

## 5.5 流式计算平台模块

该模块在实现的推荐系统中扮演一个非常重要的角色，前两节搭建日志采集系统和数据接入模块就是为这个模块服务。该模块需要处理实时的数据，主要是通过Storm框架来实现。利用Storm的性能并且结合推荐算法来实现实时推荐的效果。

### 5.5.1 Kafka与Storm的连接器

上述已经介绍了关于Kafka和Storm的原理，以及在该系统中的关系。Kafka将收集的数据日志汇集起来，等待Storm的拉取。在这个拉取的过程主要是将Kafka的Consumer和Storm的Spout两个模块连接在一起，这种方式很好的保证了两个模块之间高效的数据传输。

图5-10 Storm流式计算框架示意图

如图5-10所示，当数据存储在Kafka的Broken中，我们就可以采用Storm来主动拉取消费。由于Kafka和Storm具有良好的兼容性，所以可以采用Kafka Spout从Kafka中拉取数据，而数据源Spout需要处理特定的接口类。然后通过实现Bolt来消费从Spout发送出来的数据流，并根据具体的需求来实现对应的逻辑。如果是比较复杂的数据，可以定义多个Bolt来帮助处理，最后通过Kafka Spout和Kafka Bolt来生成Topology对象交给Storm来处理。

（1）关于Kafka Spout的实现，它们主要是处理用户的购物订单操作，当然还包括各种其他的操作信息，通过ordersAnalysisTopology来解析kafkaSpout，下面简单的列出订单事件分析的代码。

|  |
| --- |
| public class ordersAnalysisTopology {  private static String topicInfoName = "topicInfo";  private static String StormRoot = "/Storm/"+topicInfoName;  public static void KafkaSpolt(String topic) {  BrokerHosts hst = new ZkHosts("Storm1:3219,Storm2:3219,Storm3:3219");  SpoutConfig Confi = new SpoutConfig(hst,topicInfoName, StormRoot,  UUID.randomUUID().toString());  Config.scheme = new SchemeAsMultiScheme(new StringScheme());  KafkaSpout spout = new KafkaSpout(Config);  TopologyBuilder build = new TopologyBuilder();  build.setSpout("kafkaSpout",Spout);  build.setBolt("OrdersBolt", new OrdersAnalysisBolt(), 2)  .shuffleGrouping("kafkaSpout");  ….  } |

（2）关于Kafka Bolt 实现，主要是对业务进行转化，然后对消息进行处理更新。执行Storm里面的消息Tuble，将消息进行转化为可处理的信息再交给下一个Bolt对物品相似度进行更新，下面简单列出订单信息处理的代码。

|  |
| --- |
| public class OrderAnalysisBolt extends BaseRichBolt {  private OutputCollector collector;  JedisPool pool;  public void execute(Tuple tuple) {  String orderInfo = tuple.getString(0);  ordersBean order = handler.getOrdersBean(orderInfo);  Jedis dis = pool.getResource();  dis.zincrby("orderAna:Score", order.getID(), order.getScore());  }  public void prepare(Map arg0, TopologyContext arg1, OutputCollector collector) {  this.\_collector = collector;  this.pool = new JedisPool(new JedisPoolConfig(), "Storm",3219,2 \* 50000,"123");  }  } |

### 5.5.2 实时推荐计算

实时推荐模块是在Storm的框架基础上进行计算，主要是从Kafka获得实时的数据，然后将离线平台计算得到的数据（其中包括：物品的相似度信息、用户对物品评分信息以及离线的推荐列表信息），全部存储在Redis的缓存中，以供流式平台的调用。Storm流式平台主要通过时间窗口和热搜物品来进行调节。

（1）增量更新物品相似度

本系统计算物品相似度的过程是放在离线平台上进行，当面对庞大的数据量想要通过实时计算并快速得出推荐结果是不现实的。所以将重量的计算模块放在离线平台进行周期的批量处理，这个周期本系统规定为一天。同时流式计算平台实时的获得近期用户的行为，对用户行为兴趣程度进行分析。系统主要是对物品相似度进行更新计算，因为该计算公式比较简单，所有该计算规模比较小，计算更新速度比较快，采用的更新公式为(5-1)。

|  |  |  |
| --- | --- | --- |
|  |  | (5-1) |

针对上述的公式进行分析，假设用户M 对物品X有评分，那么相对而言物品相似度如何更新：

a. 如果用户M对物品x有过评分，即对该物品有产生行为信息，那么我们不需要更新任何物品相似度。

b. 如果用户M首次对物品X产生事件行为，对其进行了评分。那么相对的就会增加一。而通过公式中的，如果出现用户M对物品y也有过事件行为，则需要增加一；如果用户对物品y没有过事件行为，则不变。

（2）实时推荐方法

根据上述得到的物品相似度计算用户对物品的评分，系统中采用的评分计算公式如（5-2）所示：

|  |  |  |
| --- | --- | --- |
|  |  | (5-2) |

公式里面R(u)表示用户已经评分的集合，S(j,k)表示和物品最相似的K个物品，而表示了物品j和物品i的相似度，表示用户u对物品i的评分。

实时推荐的方法就是要根据用户最近期的产生的行为来进行推荐。人的兴趣随着时间推移会有一些变化，而近期的行为最能反映前用户的兴趣爱好。然后根据更新的物品相似度进行预测。方法主要是根据时间窗口，来得到用户近期对物品的评分。 当过了一个时间阈值过后的数据就会被过滤掉，这样的目的就是为了突出近期的用户对物品的兴趣度的影响。

对于新的用户，即冷启动问题。在离线中无法获得用户对物品的评分和推荐列表。则系统会在一开始将热门物品进行推荐，随着用户近期的浏览慢慢取代热门的物品，最终根据近期评分规则来进行推荐。

（3）实时推荐实现流程

由图5-11可以得知，在Storm上实时推荐计算的流程如下：

第一步，从Kafka得到数据信息，KafkaSpout作为一个数据源，订阅消息队列中的信息，并将消息发送给OrderAnalysisBolt进行处理。

第二步，OrderAnalysisBolt接收到数据源关于用户的事件消息之后，对消息进行一个分类评分，评分主要根据近期事件进行一个增值评分。

第三步，如果在离线平台推荐列表中没有关于该用户的推荐评分信息，那么该用户属于新用户，属于冷启动问题。则进入到ColdDealBolt处理，这个Bolt会将热门的物品列表进行推荐。

第四步，利用UpdateItemBolt对物品相似度进行一个增量更新。根据物品的ID从Redis缓存中取得离线计算的物品相似度矩阵，然后增量更新物品相似度。

第五步，得到最新的增量的相似度矩阵之后，通过RateBolt进行物品的评分计算。

第六步，对物品进行一个RateBolt降序排序，选取K个推荐物品，得到推荐列表。



图5-11 实时推荐实现流程图

## 5.6 数据存储和展示模块

### 5.6.1 数据存储和输出

数据存储部分主要存储在计算过程中产生的数据包括物品之间的相似度矩阵、物品评分矩阵以及最后推荐的列表。Redis是一个基于键值对且具有持久化的内存存储系统，可以满足本系统中对各种数据的存储和使用。为加快运算将数据存储在Redis中。

输出部分主要将推荐的结果数据保存在Mysql数据库中。通过持久化操作，为下次计算推荐表进行参考。在实时的基础上，短时间内进行更新，那么页面上可以实时的推荐给用户需要的产品。

### 5.6.2 UI展示

在完成推荐系统构建之后，本人主要开发了一个小型的APP，通过APP将推荐结果展现出来。关于该APP的开发，是采用开源代码ECMobil，进行二次的轻量级别的APP，由于本部分不是重点，所以就不多加阐述，APP的UI展示如图5-12所示。

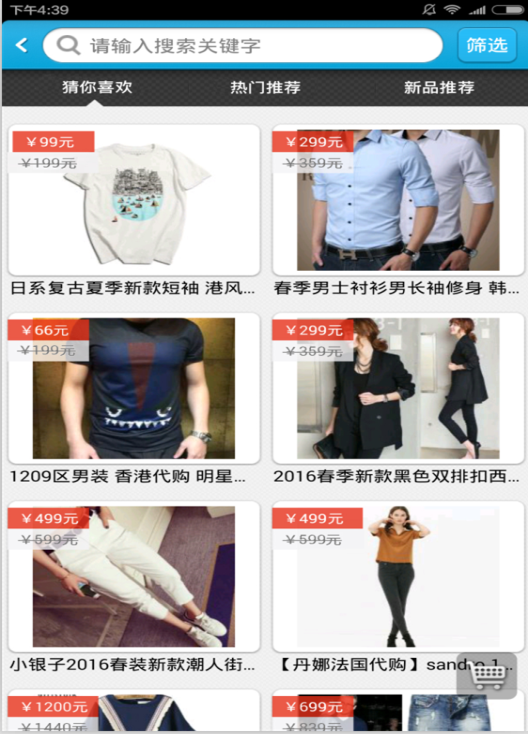
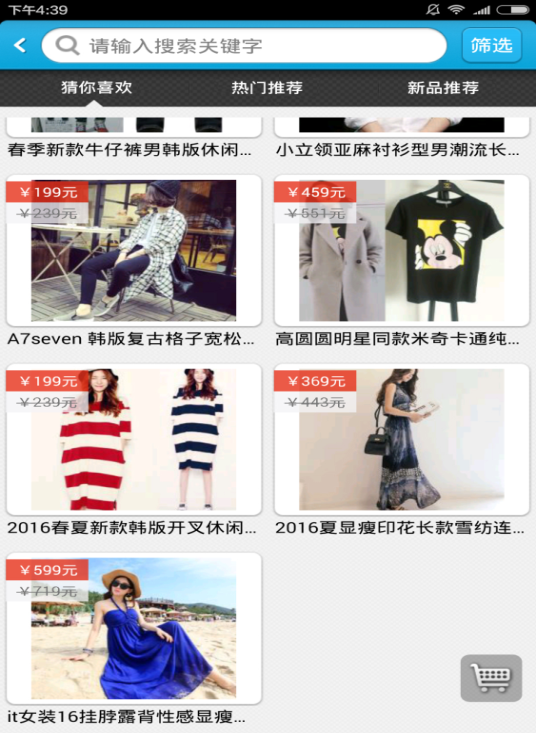
 

图5-12 小型电商APP界面展示图

图5-12中里面通过三种方式进行推荐，分别是猜你喜欢、热门推荐、新品推荐，物品的推荐主要是在“猜你喜欢”模块。当然这只是一种简单的表示。由此可知推荐系统的搭建是很复杂的，而最终只是简单的几个页面的展示。但是应用信息的展示需要考虑在哪里展示，如何展示，才能给用户一种新奇感，并且促动用户消费。

## 5.7 实验与测试

### 5.7.1 评测标准

针对推荐结果的评测，主要根据推荐结果的准确率和召回率进行评价。

推荐的准确率表示物品和数据集中用户喜爱的物品的取交集占物品的比例，准确率的公式如(5-3)所示：

|  |  |  |
| --- | --- | --- |
|  |  | (5-3) |

推荐的召回率表示物品和数据集中用户喜爱的物品的取交集占数据集中用户爱好的物品的总数的比例，公式如(5-4)所示：

|  |  |  |
| --- | --- | --- |
|  |  | (5-4) |

### 5.7.2 实验设计

（1）数据集

由于平台上面的数据量不够大，所以实验数据采用的是阿里巴巴大数据竞赛“天池”的淘宝穿衣搭配大赛的数据。这里面数据包含了十万条的商品信息，和一千多万条的用户购买记录信息。由于没有物品的评分，我采用自定义的评分机制，通过购买记录对商品进行一个评分，得到可以处理的信息数据。

（2）实验环境

实验中的算法使用Java语言实现，算法的开发工具是eclipse。使用[Storm 0.10.2 Released](http://storm.apache.org/2016/09/14/storm0102-released.html)版本构建分布式计算环境，Storm集群由四台虚拟机模拟Storm的集群环境，每台计算机都分配得到单核的CPU进行试验，具体的配置情况如表5-12所示：

表5-12集群环境配置表

|  |  |  |
| --- | --- | --- |
| 主机名 | 配置参数 | 系统 |
| Master | GPU：i7 2.3GHz 内存 8G | Ubuntu 15.04 |
| Slave1 | GPU：i7 2.3GHz 内存 8G | Ubuntu 15.04 |
| Slave2 | GPU：i7 2.3GHz 内存 8G | Ubuntu 15.04 |
| Slave3 | GPU：i7 2.3GHz 内存 8G | Ubuntu 15.04 |

（3）实验内容和结果

实验内容包括两个实验，第一个实验是关于实时推荐算法的准确率和召回率的验证，第二个实验是关于该系统的实时性能的验证，第三个实验是关于不同算法计算时间性能验证。

第一个实验采用的方法如下：

第一步，由于数据中没用用户的评分，因此需要模拟评分，那么按照公式(5-5)，将用户购买物品转化为用户对物品的评分，其中系数x代表用户的购买次数，系数a为评分的上限，通过调节系数b来调节评分变化速率。

|  |  |  |
| --- | --- | --- |
|  |  | (5-5) |

第二步，由于实时的推荐物品需要根据当前的用户的行为来更新物品相似度矩阵，并得到推荐列表，因此需要模拟用户的实时行为。同时从数据集合中按照时间读取用户记录R1，然后选取和这次购买记录时间相近并且晚于R1的数据集R2（该部分就是模拟的用户实时行为），并对R1的数据在离线平台上通过基于物品协同过滤算法进行计算得到离线的物品相似度矩阵和推荐列表，最终根据R2的模拟实时数据进行更新物品相似度矩阵，再计算得到推荐列表，最后计算推荐系统中的召回率和准确率。

通过实验可以得到实验结果，下列通过柱状图来阐明实验得到的结果：



图 5-13 推荐准确率柱状图



图 5-14 推荐召回率柱状图

图5-13和图5-14，分别是根据公式(5-3)和(5-4)得出离线推荐和在线推荐的准确率和召回率的数据。其中离线推荐利用了该系统的离线平台进行计算得出结果，而在线推荐则是在离线平台基础上，通过最近的行为信息（即：上述中模拟的R2的数据集）来对物品进行更新推荐，然后通过公式计算得出结果。从两个图中可以看出随着推荐个数的增加，这两种推荐方法的准确率和召回率就趋于接近。说明实时产生的行为对推荐的效果还是有限的，推荐个数越少则实时的行为对推荐的效果越好。

第二个实验采用的方法如下：

该实验主要测试关于实时推荐系统的实时性能，和第一个实验一样，该实验也需要对数据的评分进行处理。首先从数据集合中按照时间读取用户购买记录R1，这部分作为离线的数据进行处理，从而计算的物品相似度矩阵和推荐列表。选取和这次购买记录时间相近并且晚于R1的数据R（该部分就是模拟的用户实时行为），然后根据通过数据输入R的数据进行计算得到推荐结果。选取不同的R1数据集合的数量进行测试。得出离线计算时间和实时计算时间表结果如下所示：

表 5-13 推荐计算时间表

|  |  |  |
| --- | --- | --- |
| R1（条） | 离线计算时间（s） | 实时计算时间（s） |
| 1000 | 120 | 1.9 |
| 10000 | 445 | 2.3 |
| 100000 | 1239 | 2.5 |

通过表5-13可知，离线计算部分会因为数据的增加，而导致计算时间的增加。而实时推荐因为是在离线的基础上做增量的更新，所以可以在短时间内得出因为用户新的操作行为而得到新的推荐结果。这也很好的证明了，本系统采用离线平台来进行重操作。而在线推荐在此基础上，进行计算量小的更新。由上述的实验结果，可以得出实时推荐结果响应时间都在用户可接受的范围之内。

第三个实验采用的方法如下：

本实验只是简单的利用同样的方式进行离线计算和在线的更新计算，采用不同模型和方法进行一个简易的比较，同时都取得数据记录集10000条进行操作。

表 5-14 不同算法计算时间表

|  |  |  |
| --- | --- | --- |
| 采用的方法 | 离线计算时间（s） | 实时计算时间（s） |
| 系统采用的方法 | 445 | 2.3 |
| User-CF | 610 | 3.4 |
| Logistic模型 | 569 | -- |
| LFM模型 | 538 | -- |
| FM模型 | 579 | -- |

通过表5-14可知，通过简单的比较，不同的算法之间的离线计算时间和实时计算时间都不同，由于比较复杂的算法，本人没有具体的研究这些算法对应的在线更新算法，所以没办法得到实时的计算时间，最后得到的结果是本文采用的方法的计算时间和其它几种算法计算时间的对比。

通过上述的三个实验进行，可以发现在经过实时的计算处理之后，物品推荐的准确率和召回率和基于物品协同过滤算法的离线计算有了略微的提升，当然采用该算法里面最简易的相似度和评分方法，和其他比较复杂的公式方法对比，确实是牺牲掉了一些准确率，目的是为了保证在实时性能上的快速更新，提供更好的用户体验。而且第二个实验和第三个实验结果，说明该算法在实时性能方面能更好的满足用户的体验，从而表明本系统的设计思想有一定的价值意义。

## 5.7 本章小结

本章是整个论文的核心部分，具体的阐述了系统的整体框架，同时通过详细的流程图，来具体的说明整个系统的设计。该部分将系统通过五个模块进行表述以及最后关于算法的进行的实验测试。

第一个模块，日志采集模块，采用Flume框架进行搭建，同时介绍了系统应用端涉及到的一些日志信息的数据格式，然后概述了数据采集框架的流程，最后对Flume进行配置，并将各种配置方案以表格的形式进行展现，通过时序图来说明采集数据流的过程。

第二个模块，数据接入模块，该模块采用Kafka框架，简要说明了利用该框架对整个系统的优势。介绍了该模块和采集系统连接的配置和模块之间的连接器。

第三个模块，离线计算平台模块，该模块是基于Hadoop上进行批处理的计算平台。该小节也具体通过流程图和实验例子来阐述整个算法的计算流程。

第四个模块，流式计算平台模块，该模块是本文的重点，前面的框架搭建和设计都是为这块服务的。流式计算平台采用Storm框架，从接入数据模块中获得实时的流数据，并结合离线提供的物品相似度矩阵和推荐列表，同时对物品相似度和推荐列表进行更新，最后实现实时的推荐。系统中的创新点是如何利用时间因子，来优化评分。同时针对用户冷启动问题提出解决方案。

第五个模块，数据存储和输出模块。该模块只是很简单得介绍了在计算过程需要将这些计算数据缓存在Redis中，为了方便实时调用，增加响应速度，将最终的推荐结果保存在Mysql中，进行持久化存储。

最后介绍了评测标准：准确率和召回率，然后设计实验，实验包括: 数据集的选择（阿里大数据竞赛的数据集）、实验环境搭建、实验内容，最终通过柱状图的方式展现推荐的准确率和召回率，并通过表格阐述了实时推荐消耗的时间和离线消耗时间的比较，通过实验可以表明本系统的实时推荐确实起到一定的作用。

# 第六章 总结与展望

随着“大数据”时代的来临，互联网的信息资源越来越丰富，从而造成信息的过载，导致人们在网上很难容易得到需要的信息。于是，推荐系统的发展对于互联网越发的重要。

本文主要是实现关于电商的实时推荐系统，利用目前比较流行的实时计算框架Storm来实现。前期也尝试过通过流式计算直接实现推荐算法，但是在数据量比较大的情况下，实现的效果并不太理想，所以在后面加入离线平台，把复杂的算法计算放在线下处理，来解决由于计算过于复杂无法及时得到推荐列表的问题，这算是该系统的一个亮点。

该系统包括了系统采集模块、系统接入模块、系统的离线计算、系统流计算模块、系统存储和输出模块。其中，也用简单例子说明算法的计算的过程。由于为了体现实时性，所以该系统上使用比较基础的基于物品的协同过滤算法，目的是为了让线上更新推荐计算不会过于复杂，导致推荐结果输出太慢而影响用户的体验。在文中已经分析了如何通过增量算法更新物品相似度，并利用时间窗口来获取近期的用户行为来进行推荐结果的计算。当然本人也还在尝试利用复合的算法来实现的推荐系统，但是目前效果还是不理想。

由于各种因素的存在，包括时间、环境和本身的能力。本系统还有很大的改进空间。本系统还可以向这几个方向进行探讨：

（1）关于算法方面，可以研究和探讨更适合的算法，不管是复合算法还是新型的算法；本系统只是采用基于物品的协同过滤的算法来实现，结合了在线的更新算法满足推荐效果。但是不同的推荐算法有不同的优势，所以还需要继续试验和探讨。

（2）需要更多的真实商业环境下进行测试工作，目前，这边的数据集，只是从

网上获取的数据用来测试，还是具有片面性。

（3）在离线方面可以利用深度学习，来获取更多有用的信息，以便于在线的推荐。

参考文献

1. 蔡自兴. 人工智能及其应用:研究生用书[M]. 清华大学出版社, 2004.
2. 李勇, 徐振宁, 张维明. Internet个性化信息服务研究综述[J]. 计算机工程与应用, 2002, 38(19):183-188.
3. 项亮. 推荐系统实践[M]. 人民邮电出版社, 2012.
4. Liu F, Tang B, Yuan X, et al. Recommender System in E-commerce[C]// International Conference on E-Business and E-Government. IEEE Computer Society, 2012:700-703.
5. D’Souza S. Apache flume[J]. 2003.
6. Garg N. Apache kafka[M]. Packt Publishing, 2013.
7. Components S, Implementation S. Easy, Real-Time Big Data Analysis Using Storm[J]. Dr Dobbs Journal.
8. Somasundaram N. "Apache Samza - A Stream Processing Framework"[J]. 2014.
9. Akidau T, Balikov A, Bekiro, et al. MillWheel: fault-tolerant stream processing at internet scale[J]. Proceedings of the Vldb Endowment, 2013, 6(11):1033-1044.
10. Neumeyer L, Robbins B, Nair A, et al. S4: Distributed Stream Computing Platform[C]// IEEE International Conference on Data Mining Workshops. IEEE, 2010:170-177.
11. Vavilapalli V K, Murthy A C, Douglas C, et al. Apache Hadoop YARN: yet another resource negotiator[C]// Symposium on Cloud Computing. 2013:5.
12. Shvachko K V. Apache hadoop[J]. Usenix Org.
13. Ghemawat S, Gobioff H, Leung S T. The Google file system[C]// Nineteenth ACM Symposium on Operating Systems Principles. ACM, 2003:29-43.
14. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[C]// Conference on Symposium on Opearting Systems Design & Implementation. DBLP, 2004:137-150.
15. Miner D, Shook A. MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems[M]. O'Reilly Media, Inc. 2012.
16. Su X, Khoshgoftaar T M. A survey of collaborative filtering techniques[J]. Advances in Artificial Intelligence, 2009, 2009(12):4.
17. Schafer J B, Konstan J A, Riedl J. E-Commerce Recommendation Applications[J]. Data Mining and Knowledge Discovery, 2001, 5(1):115-153.
18. 张振亚, 王进, 程红梅,等. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学, 2005, 32(9):160-163.
19. Adler J, Parmryd I. Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient.[J]. Cytometry Part A, 2010, 77(8):733.
20. 孟海东, 张玉英, 宋飞燕. 一种基于加权欧氏距离聚类方法的研究[J]. 计算机应用, 2006, 26(s2):179-180.
21. Fligner M A, Verducci J S, Blower P E. A Modification of the Jaccard–Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings[J]. Technometrics, 2002, 44(2):110-119.
22. 刘庆鹏, 陈明锐. 优化稀疏数据集提高协同过滤推荐系统质量的方法[J]. 计算机应用, 2012, 32(4):1082-1085.
23. Canny J. Collaborative filtering with privacy via factor analysis[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2002:238-245.
24. O'Donovan J, Smyth B. Trust in recommender systems[C]// International Conference on Intelligent User Interfaces, January 10-13, 2005, San Diego, California, Usa. DBLP, 2005:167-174.
25. Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[J]. New Page, 2013, 7(7):43--52.
26. Cannon R K. Scribe: US, US3100346[P]. 1963.
27. 张川, 邓珍荣, 邓星,等. 基于Chukwa的大规模日志智能监测收集方法[J]. 计算机工程与设计, 2014, 35(9):3263-3269.
28. Pazzani M J, Billsus D. Content-based recommendation systems[M]// The adaptive web. Springer-Verlag, 2007:325-341.
29. Cantador I, Bellog, Alejandro N, et al. Content-based recommendation in social tagging systems[C]// ACM Conference on Recommender Systems. ACM, 2010:237-240.
30. Resnick P, Varian H R. Recommender systems. Commun ACM[J]. 1997, 40(3):56–58.
31. 李忠俊, 周启海, 帅青红. 一种基于内容和协同过滤同构化整合的推荐系统模型[J]. 计算机科学, 2009, 36(12):142-145.

致谢

时光荏苒，我的硕士生涯已接进尾声。这几年的时光既漫长又短暂，其中充满了酸甜苦辣，更有收获和成长。几年来，感谢陪我一起度过美好时光的每位尊敬的老师和亲爱的同学，正是你们的帮助，我才能克服困难，正是你们的指导，我才能解决疑惑，直到学业的顺利完成。

本人的学位论文是在我的恩师刘志镜教授和姚勇副教授的殷切关怀和耐心指导下进行并完成的，衷心感谢我的恩师对我的淳淳教诲和悉心关怀。从课题的选择、项目的实施，直至论文的最终完成，他们都始终给予我耐心的指导和支持，我取得的每一点成绩都凝聚着恩师的汗水和心血。他们开阔的视野、严谨的治学态度、精益求精的工作作风，深深地感染和激励着我，在此谨向他们致以衷心的感谢和崇高的敬意。

感谢西安电子科技大学大学 14 级硕士班的全体同学陪我一起走过这段人生难忘的历程!

作者简介

##### 基本情况

朱群，男，福建人，1991年11月出生，西安电子科技大学计算机学院计算机技术专业2014级硕士研究生。

##### 教育背景

2010.09～2014.07 福建师范大学，本科，专业：软件工程

2014.09～2017.07 西安电子科技大学，硕士研究生，专业：计算机技术

##### 攻读硕士学位期间的研究成果

###### 3.1发表学术论文

1. Hong ZHOU, Zhi-jing LIU, Qun Zhu, Bing-hua WANG, Jing-jie LI. The Problem-Solving Agricultural Knowledge Service System for Farm Households, ICAMM, Bangkok Thailand, Octobe 2016, 2016 International Conference on Applied Mathematics and Mechanics, DEStech Publications,Inc. 2016, 472-478.

###### 3.2参与的科研项目

1. “陕勤网”项目，2015-5~2016-07，负责该项目手机Web的开发和维护。
2. 陕西省慧农科技“农掌门”项目，2015-10~2016-04，负责该项目Android APP农民端和专家端的开发和维护。