**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4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，主要包括错误数据、相似重复数据和缺失数据等类型。这些脏数据产生的原因多种多样，例如数据来源不同，存储于不同的操作系统以及硬件平台等。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。

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，然后在前人工作成果的基础上，基于MPN算法提出了改进的IMPN算法。IMPN算法主要有以下三个创新：（1）传统的MPN算法主要依赖相关领域的专家经验生成排序关键字，不利于处理未知领域或未知类型的数据，IMPN算法使用基于统计字段区分度的方法进行排序关键字的选取，有效地修正了MPN的缺点；（2）IMPN使用动态调整滑动窗口大小的机制，与传统MPN算法所采用的固定大小的滑动窗口相比，算法操作过程更加灵活，可以更加全面地检测重复记录，同时，有效地避免了不必要的时间消耗，提高了算法运行效率；（3）对于排序关键字为空的记录，对其进行预标记，改善了MPN算法在处理排序关键字不完整的情况下的缺陷，提高算法应对缺失字段的能力，增强了鲁棒性。

虽然IMPN算法对MPN算法进行了三方面的改进，但是，随着数据量的增加，IMPN算法的查准率下降的问题随之凸显。基于这一问题，本文基于遗传算法和人工神经网络在相似重复记录检测领域的研究，提出了GANN-IMPN算法。GANN-IMPN算法将基于遗传算法改进的神经网络GANN算法应用到IMPN算法的重复记录判断中，提高了算法的查准率。传统的GANN算法首先通过带标签的数据集作为训练数据训练出多个BP神经网络，然后，将每一个训练完成的神经网络作为遗传算法的个体，所有神经网络构成遗传算法的种群，利用遗传算法使用全局信息的优势，弥补神经网络易陷入局部极小值的缺点。但是，GANN在测试阶段，需要对测试数据集中的所有记录进行全排列，然后使用训练好的神经网络判断重复记录，时间复杂度较高。本文改进了GANN的不足，综合GANN算法与IMPN算法的优势，提出GANN-IMPN算法，仅使用训练好的网络对滑动窗口内的记录进行判断，避免了不必要的判断，改进了GANN的运行效率，同时提高了IMPN的查准率。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。经对比实验证明，IMPN算法的查准率取得了显著提升。

最后本文将相似重复记录检测技术应用到航天情报信息管理系统中，用以解决多源数据合并造成的重复记录问题，提高了数据质量。

**关键词**：数据清洗，相似重复记录检测，BP神经网络，机器学习，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, missing data and so on. 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 and the amount of missing duplicate records by dynamically adjusting the size of sliding window, and promoting rubust by marking those records whose sort key is null.

The precision ratio’s decreasing becomes a fierce problem as the data grows too large. This paper provides a solution that the neural network improved by genetic algorithm is applied to the IMPN to judge whether two records are duplicate. 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 whether 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.

At last, the duplicate record detection method is applied in the spaceflight information management system in order to eliminate those duplicate records which is caused by the merge of multi-source data, that improves the data quality.

**Keywords:** Data Cleansing, Duplicate Record Detection, BP Neural Network, Machine Learning, Genetic Algorithm

插图索引

**未找到图形项目表。**

表格索引

**未找到图形项目表。**

**未找到图形项目表。**

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

SNM Sorted Neighborhood Method 近邻排序算法

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

ANN Artificial Neural Network 人工神经网络

GA · Genetic Algorithm 遗传算法

目录

[摘要 I](#_Toc510644166)

[ABSTRACT III](#_Toc510644167)

[插图索引 V](#_Toc510644168)

[表格索引 VII](#_Toc510644169)

[符号对照表 IX](#_Toc510644170)

[缩略语对照表 XI](#_Toc510644171)

[第一章 绪论 1](#_Toc510644172)

[1.1 研究的背景和意义 1](#_Toc510644173)

[1.2 国内外研究现状 2](#_Toc510644174)

[1.3 论文研究的主要内容 3](#_Toc510644175)

[1.4 论文结构 3](#_Toc510644176)

[第二章 重复记录检测相关算法概述 5](#_Toc510644177)

[2.1 相似重复记录概述 5](#_Toc510644178)

[2.2 相似度匹配算法 6](#_Toc510644179)

[2.2.1 基于单字段相似度匹配 6](#_Toc510644180)

[2.2.2 基于多字段相似度匹配 7](#_Toc510644181)

[2.3 相似重复记录检测算法 8](#_Toc510644182)

[2.3.1 近邻排序算法 8](#_Toc510644183)

[2.3.2 多趟近邻排序算法 9](#_Toc510644184)

[2.3.3 其它算法 10](#_Toc510644185)

[2.4 BP神经网络理论基础 11](#_Toc510644186)

[2.4.1 神经元模型 11](#_Toc510644187)

[2.4.2 梯度下降法 12](#_Toc510644188)

[2.4.3 BP网络前向传播和反向传播 13](#_Toc510644189)

[2.6 算法的衡量标准 29](#_Toc510644190)

[2.7 本章小结 15](#_Toc510644191)

[第三章 改进的IMPN算法 17](#_Toc510644192)

[3.1 基于字段区分度提取关键字的方法 17](#_Toc510644193)

[3.1.1 传统的提取关键字的方法 17](#_Toc510644194)

[3.1.2 改进的字段区分度方法 20](#_Toc510644195)

[3.2 自适应大小的滑动窗口检测方法 21](#_Toc510644196)

[3.2.1 传统的滑动窗口检测方法 21](#_Toc510644197)

[3.2.2 改进的自适应大小的滑动窗口 22](#_Toc510644198)

[3.3 基于标记的处理不完整排序关键字方法 24](#_Toc510644199)

[3.3.1 MPN的排序方法存在的问题 24](#_Toc510644200)

[3.3.2 改进的基于标记的方法 24](#_Toc510644201)

[3.4 IMPN算法设计 25](#_Toc510644202)

[3.4.1 算法流程设计 25](#_Toc510644203)

[3.4.2 时间复杂度分析 27](#_Toc510644204)

[3.5 实验设计与结果分析 27](#_Toc510644205)

[3.5.1 实验数据介绍 27](#_Toc510644206)

[3.5.2 SNM、MPN、IMPN综合对比实验 29](#_Toc510644207)

[3.6 本章小结 32](#_Toc510644208)

[第四章 基于遗传神经网络改进的IMPN算法 35](#_Toc510644209)

[4.1 遗传神经网络用于相似重复记录检测 35](#_Toc510644210)

[4.1.1 BP神经网络的设计 35](#_Toc510644211)

[4.1.2 遗传神经网络与重复记录检测 38](#_Toc510644212)

[4.2 改进的遗传神经网络 41](#_Toc510644213)

[4.2.1 神经网络过拟合现象 41](#_Toc510644214)

[4.2.2 “Dropout”防止过拟合 42](#_Toc510644215)

[4.3 对比实验 44](#_Toc510644216)

[4.4 本章小结 44](#_Toc510644217)

[第五章 航天情报系统中的相似重复记录检测 45](#_Toc510644218)

[5.1 系统需求分析 45](#_Toc510644219)

[5.1.1 系统建设背景与目标 45](#_Toc510644220)

[5.1.2 需求分析 45](#_Toc510644221)

[5.2 系统设计与实现 46](#_Toc510644222)

[5.2.1 系统概要设计 46](#_Toc510644223)

[5.2.2 系统实现 50](#_Toc510644224)

[5.3 数据清洗模块 51](#_Toc510644225)

[5.3.1 “脏数据”产生原因 51](#_Toc510644226)

[5.3.2 重复记录检测算法的应用 52](#_Toc510644227)

[5.3.3 实验设计与结果分析 53](#_Toc510644228)

[5.4 本章小结 53](#_Toc510644229)

[第六章 总结与展望 55](#_Toc510644230)

[参考文献 57](#_Toc510644231)

[致谢 59](#_Toc510644232)

[作者简介 61](#_Toc510644233)

# 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，数据库的信息量也与日逐增。

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[[[1]](#endnote-2)][[[2]](#endnote-3)]。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[[[3]](#endnote-4)][[[4]](#endnote-5)][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[[[5]](#endnote-6)]、数据实体识别[[[6]](#endnote-7)]、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[[[7]](#endnote-8)]。

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[[[8]](#endnote-9)]。

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提出的“排序&合并（Merge/Purge）”算法[[[9]](#endnote-10)]。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等人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[[[10]](#endnote-11)]。Hernandez等提出了一种名为“多趟近邻排序”的算法[[[11]](#endnote-12)][[[12]](#endnote-13)]，即MPN（Multi-Pass Sorted Neighborhood），该算法独立地执行多次SNM（Sorted Neighborhood Method），每次采取不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C语言重写的OPS5[[[13]](#endnote-14)]规则编程。Newcombe等则采用应用相关的关键字以将重复记录聚类到相近的位置[[[14]](#endnote-15)]。Qiu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。[[[15]](#endnote-16)]Gianni Costa等人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[[[16]](#endnote-17)]。Alfredo Ferro使用了基于q-grams的相似度衡量函数[[[17]](#endnote-18)]，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。复旦大学周傲英等比较早开始数据清理的研究工作[[[18]](#endnote-19)]。邱越峰等提出了基于N-Gram的相似记录检测算法[[[19]](#endnote-20)]。算法以一条数据的N-Gram值作为排序键，该算法在对因为拼写错误造成的重复记录进行检测时表现良好。陈伟提出了基于权重进行相似重复记录检测的方法[[[20]](#endnote-21)]，具体实现按照字段的等级划分权重，并结合长度过滤的思路减少冗余的字段相似度计算。

在数据清洗市场化领域，国内外涌现了一批优秀的数据清洗软件以及框架，包括商业上和各大学以及研究机构开发的数据清洗软件[[[21]](#endnote-22)][[[22]](#endnote-23)]。

综上，在相似重复记录检测领域，国内外研究人员已取得了诸多进展，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，所以仍然有研究的价值和改进的空间。

## 1.3 论文研究的主要内容

由目前的研究现状可以看出，相似重复记录检测领域的发展已取得诸多有效成果，但仍旧存在一定问题，主要体现在：

(1)检测效率与查全率存在提升空间，尤其是较大数据量的相似重复检测问题。

(2)大多数数据清理只针对特定领域以及业务场景，各行业需要更加通用的相似记录检测方案。

(3)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，针对传统的基本近邻排序算法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大小以节约时间并减少被遗漏的重复记录。

(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。

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。本文将反向传播神经网络应用于相似重复记录检测中，将两条记录对应字段间的相似度组成的向量作为神经网络的输入，利用有监督的学习训练出多个三层BP神经网络，然后利用遗传算法将这些神经网络组成的种群进行迭代优化，选择出适应度最好的个体，这样可以克服BP神经网络容易陷入局部最小值的缺点。将遗传神经网络判断两条记录是否相似的方法应用到IMPN算法中，提高了IMPN算法在面对较大数据量时的查准率。

## 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，分析了它们各自的优缺点以及适用条件。然后在第二部分对最基本的近邻排序算法和多趟近邻排序算法，介绍了算法的基本原理、设计思路以及算法步骤和算法的优缺点。除此之外还介绍了其它常用算法包括优先级队列算法、N-Gram算法等。然后对本文用到的BP神经网络理论基础进行说明。最后介绍了相似重复记录检测领域衡量算法的几个常用标准及其计算方法。

第三章，首先介绍了改进的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IMPN算法的查全率较高的优势，并分析了算法的缺点，即随着数据量的增大查准率不够理想。

第四章，首先介绍了神经网络在相似重复记录检测中的应用，以及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。接下来重点介绍了遗传BP神经网络应用于IMPN算法中，并通过实验验证了遗传BP神经网络对IMPN算法查准率的提升。

第五章，主要介绍了航天情报信息管理系统中的数据清理模块，该模块是相似重复记录检测算法在该系统中的应用，主要内容包括数据清理模块的设计、重复记录产生的原因、IMPN算法在系统中的应用以及该算法对数据质量的提高等。

第六章，总结了本文的内容以及相似重复记录检测算法研究过程中遇到的问题，并对未来的研究方向进行了展望。

# 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

相似重复记录是指，数据库中存在这样的两条记录、，它们的内容相同或者相似，且都对应着同一个现实实体，则记录对互为相似重复记录。实际数据库中可能拥有多对互为相似重复的记录，它们的存在降低了数据的质量，可能会妨碍系统的正常运行，甚至会影响企业信息管理系统的决策正确性。

表2.1给出了学生信息表中的相似重复记录示例：

表2.1 学生信息表中的重复记录

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stu\_ID | Name | Gender | Brithday Date | School |
| 1801001 | Sam Water | M | 1993/01/02 | School of Computer Science, Xi’an University of Electronic Science and Technology |
| 1802002 | Jack Panda | Female | 1990/07/20 | School of Artificial Intelligence, Xi’an University of Electronic Science and Technology |
| 1801003 | S. Water | Male | 1993/1/2 | Schol of Computer Science, Xi’an University of Electronic Science and Technology |
| 1802004 | Jack Panda | Female | 1990/07/20 | School of Artificial Intelligence, Xi’an University of Electronic Science and Technology |
| 1801005 | Mr.Sam W | Male | 1993-01-02 | College of Computer Science, Xi’an University of Electronic Science and Technology |

表2.1展示了5条学生记录，其中Stu\_ID为1802002和1802004的两条记录的所有字段内容完全一致，说明这两条记录对应现实世界中的同一个学生的信息，所以它们互为相似重复记录。表中Stu\_ID为1802001、1802003、1802005的三条记录表面上内容是不一样的，它们的区别在于：Name字段值分别为“Sam Water”、“S. Water”、“Mr.Sam W”，明显是对“Sam Water”采用了不同的书写方式而造成的；Gender字段值“Male”、“M”则是全称和缩写的区别，均代指男性；Brithday Date字段则是使用的不同的时间格式，但它们都是代表相同的一天“1993/01/02”；所属学院字段中，出现了“Schol”这样的拼写错误；经过以上观察分析可以发现，这三条内容相似的记录同样对应同一个学生。

相似重复记录产生的原因多种多样，包括人工操作过程中的录入错误或者管理错误造成的重复、不同来源的数据集进行合并时产生的重复、信息系统重构时新旧版本的数据库合并造成的重复等。

相似重复记录检测目前应用最广泛的手段是基于“排序/合并”的方法：首先对包含重复记录的数据集进行排序，排序使用的关键字按照某种固定的方式（如某字段的前三个辅音字母等）从记录的相应字段中提取，排序之后相似重复的记录汇聚在相邻的位置，然后通过对相邻位置的记录进行对比判等，可以检测出相似重复记录。

## 2.2 相似度匹配算法

相似重复记录检测过程中需要对不同的记录对进行整体相似性判断，这就需要用到字段相似度匹配算法。目前该领域的算法主要有两大类：基于单个字段的匹配算法和基于多个字段的匹配算法。

### 2.2.1 基于单字段相似度匹配

基于单字段的相似度匹配算法在相似重复记录检测过程中的应用思想在于，通过计算两条记录相同字段对应内容的相似度来衡量记录整体的相似与否，这是一个从部分到整体的过程。常用的算法包括：编辑距离算法、Smith-Waterman算法、Jaro算法等。

编辑距离算法是Levenshtein于1965年提出的一种基于字符的相似度匹配算法，又名L-距离算法[[[23]](#endnote-24)]。定义两个字符串和的编辑距离：变成需要对其单个字符进行插入、替换、删除操作的次数。编辑距离越小代表和越相似。



图2.2 编辑距离示意图

如图2.2所示，字符串“change”经过3次插入操作和一次删除操作可以变成字符串“challenge”，所以这两个字段的编辑距离为4。计算两个字符串间的编辑距离的经典解法是使用动态规划方法。L-距离算法在某些场景下（字母书写错误、存在缩写等）效果较好。

Smith-Waterman算法[[[24]](#endnote-25)]最早是在生物学序列比对领域被提出的，用于匹配遗传序列。S-W算法也是一种动态规划算法，它是Needleman-Wunsch算法的一个变种，主要思路是通过罚分和空位计算不同字段内容的相似度。S-W算法可以有效应对包含不正确值的相似重复记录，但处理字符串缩写、字母颠倒情况的能力较差。

Jaro算法[[[25]](#endnote-26)]由Jaro在1976年提出的基于字符串公共子集的相似度匹配算法。Jaro距离用来衡量两个字符串的相似度，对于给定的字符串和，两者的Jaro距离如下公式所示：

 公式2-1

其中，代表匹配的字符个数，代表换位的数目。Winkler提出的Jaro-Winkler相似度匹配算法在Jaro算法的基础上，在开始时赋予相同的字符串更高的分数，减小了原算法对于字符距离限制的影响，提高了算法在面对较分散的长字符串时的检测准度。Jaro-Winkler距离的计算公式如下：

 公式2-2

其中，是Jaro距离，是前缀的匹配长度，是一个常数，作用是可以调整前缀匹配的权值，0.25。

### 2.2.2 基于多字段相似度匹配

基于多字段的相似度匹配算法的思想是将一条记录视为一个整体，通过计算两条记录整体上的相似度判断是否互为相似重复记录。常用的算法包括余弦相似度匹配算法、基于监督训练的机器学习方法等。

余弦相似度[[[26]](#endnote-27)]是一种基于TF-IDF加权算法的多字段相似度匹配方法。算法的步骤如下：

表2.2 余弦相似度匹配算法步骤

|  |
| --- |
| **余弦相似度匹配算法** |
| 1.将需要匹配的字段内容进行分词，得到互相独立的单词；  2.对每个单词分配权重，，其中单词出现的次数（词频）用表示，表示记录总数除以包含的记录个数（逆文档频率）；  3.将待匹配的字段转化成向量和；  4.计算向量的余弦相似度：；  5.余弦结果越接近1证明记录间相似度越高，将结果与阈值进行比较判断记录是否相似。 |

除此之外，机器学习领域中的分类技术可以用来检测判断重复记录。依赖于多个字段进行相似判断时，不同的字段拥有不同的权重，所以对记录相似与否的影响程度也不同，多字段之间相似度到记录整体的相似度关系是非线性的。通过已知数据集（可以明确不同记录之间相似与否）对神经网络进行训练，然后采用训练好的网络对由记录对生成的输入向量进行计算，得到的结果若大于阈值则认定两条记录是重复记录。

## 2.3 相似重复记录检测算法

相似重复记录最直接简单的方法是对数据集中的数据进行一对一地比对，这种做法简单，查重效果好，但是时间复杂度为，处理较大数据量时间消耗过多。“排序/归并”是目前相似重复记录检测算法的主要方法，即通过对记录排序，将相似记录汇聚到邻近位置然后进行邻域查重，常见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）、N-Gram算法等。

### 2.3.1 近邻排序算法

SNM算法的设计思路是：首先指定数据集排序采用关键字的生成方式，然后遍历数据集对每一条记录生成排序关键字并附加到记录后，然后对数据集按照对应关键字字段进行排序，根据相似记录的关键字也是相似的原理，不同的重复记录在排序完成后理论上会处于邻近的位置，最后采用滑动窗口的方式对数据集进行重复检测。SNM算法步骤为：

表2.3 SNM算法步骤

|  |
| --- |
| **SNM算法** |
| 1.确认排序关键字的生成方案；  2.对每条记录生成排序关键字；  3.按照对数据集的记录进行排序（只考虑内部排序）；  4.确定滑动窗口的大小，每次比较时将新进入窗口的记录与窗口内剩余的条记录进行相似性判断。 |

滑动窗口的过程如图2.3所示：



图2.3 SNM算法滑动窗口处理过程

数据集大小为情况下使用SNM算法生成排序关键字过程的时间复杂度为，排序过程的时间复杂度为，滑动窗口归并过程的时间复杂度为，其中为窗口的固定大小。可以看出SNM算法的优点在于比较过程效率较高，运行速度较快。但它存在比较明显的缺点：

（1）过于依赖生成的排序关键字。选择不当的关键字生成方案可能导致相似重复记录相距较远，不相似的记录处于邻近位置，这就导致算法的检测效果大打折扣。

（2）滑动窗口的大小较难选择。若太大，虽然检测效果可能提高，但是会导致算法的运行时间增大；若太小，很可能导致相似重复记录无法被窗口覆盖到，导致算法查全率下降。

### 2.3.2 多趟近邻排序算法

多趟近邻排序算法（Multi-Pass Sorted Neighbothood，MPN）算法是在SNM算法的基础上提出来的一种改进算法。该算法的改进点在于：

（1）对数据集互不干扰地执行多趟近邻排序算法，每次采用不同的排序关键字生成方案，并且滑动窗口的大小相对于传统的SNM算法可以更小。

（2）对执行完多趟SNM算法的结果求传递闭包。

计算传递闭包的理论基础是相等的传递性：若记录和是相似重复记录，记录和是相似重复记录，则和也互为相似重复记录。传递闭包的计算多采用warshall算法。

MPN算法的步骤如下所示：

表2.4 MPN算法步骤

|  |
| --- |
| **MPN算法** |
| 1.确认个排序关键字的生成方案，独立重复地执行步骤2~4次；  2.对数据集生成排序关键字；  3.按照对数据集的记录进行排序（只考虑内部排序）；  4.确定滑动窗口的大小，每次比较时将新进入窗口的记录与窗口内剩余的条记录进行相似性判断得到重复记录集合；  5.对个重复记录集合求传递闭包得到最终的重复记录集合。 |

含有两趟SNM过程的MPN算法的流程图如下所示：



图2.4 MPN算法流程图

正式由于MPN算法引入了传递闭包的计算，使得一些容易被遗漏的重复记录被检测了出来，提高了算法的查全率，同时每轮SNM过程的滑动窗口也可以变得更小，缩短了滑动归并的执行时间。但是MPN算法也存在缺点：依旧没有克服SNM算法对于排序关键字的依赖性；计算传递闭包容易导致算法的误识别率上升。

### 2.3.3 其它算法

N-Gram算法是一种聚类思想的算法[[[27]](#endnote-28)]。N-Gram值由记录中每个单词出现的概率综合计算得出，相似重复记录的N-Gram值在数值上也相似，算法的优势在于对常见拼写错误表现较好。邱越峰等人用基于域的重复矩阵代替全局重复矩阵，将相似重复记录聚类到同一个簇中，并进行“pair-wise”比较，提高了检测精度。韩京宇等人提出了一种基于N-Gram层次空间的聚类算法DGHS[[[28]](#endnote-29)]，记录、的N-Gram相似性计算公式如下：

 公式2-4

其中表示记录的所有字段组成的集合。将记录整体作为字符串，由逐步变长的N-Gram映射相应的子空间，然后归并的时候采用层次聚类，最终实现对相似重复记录的检测。

优先级队列算法进行相似重复记录检测的思想是：用包含有不同重复记录簇的优先级队列来替换传统SNM算法中的滑动窗口，算法在扫描过程中，遇到队列中不含有的记录则赋予其最高的优先级然后加入到队列中，如果含有该记录，则将相应的重复记录簇的优先级设为最大。使用多个关键字进行多趟优先级队列算法的过程示意图如下所示：



图2.5多趟优先队列扫描算法过程示意图

## 2.4 BP神经网络理论基础

### 2.4.1 神经元模型

人工神经网络简单模拟大脑处理信息的机制，它是由许多互相连接并传递信息的神经元组成的非线性处理系统[[[29]](#endnote-30)]，每个组成单元的结构功能并不复杂，整体却能以任意精度逼近线性或者非线性函数，从而可以表征真实社会中更加复杂的问题。神经网络中的一个神经元所起到的作用是接收来自其他神经元的加权输入，然后结合自身的阈值(偏置)，最后经过非线性函数的处理，得到输出结果[[[30]](#endnote-31)]。典型的神经元模型如图2.6所示：



图2.6 神经元模型

其中，~是输入向量的不同分量，~是神经元各个突触的权值，表示偏置，是加权求和操作，则是激活函数。则对于输入向量经过此神经元时，经过加权求以及激活函数得到的输出为：。许多的类似于这样的神经元则组成了人工神经网络。

### 2.4.2 梯度下降法

梯度下降法[[[31]](#endnote-32)]是一种经典的最优化方法，其主要思想是不断沿着负梯度方向进行搜索。给定目标函数如公式2-5所示，是要学习的参数，是第个输入特征向量的第个分量，，表示偏置，共有维特征。

 公式2-5

采用均方误差损失函数，如公式2-6所示，共有个训练样本，表示第个训练样本的真实类标向量，表示第个训练样本的预测类标向量，当损失函数的值最小时，说明所训练出的模型参数最能拟合训练样本，因此求解参数的过程就是最小化损失函数。

 公式2-6

首先求出损失函数对参数的导数，如公式2-7所示。然后，根据损失函数对参数的负梯度方向更新参数，如公式2-7所示。是梯度下降法的学习速率，一般情况下，随着学习次数的增加，参数逐渐减小，即参数在学习过程中的变化越来越小。

 公式2-7

 公式2-8

由公式2-8可以看出，对每一个参数，都需要使用全部样本来学习该参数的变化量，将这种梯度下降法的实现方式称作“批梯度下降法”。在实际操作中，由于样本个数较大，所以这种参数更新方法会使导致训练过程缓慢，难以应用于实际问题。

为了克服批梯度下降法的缺点而出现了随机梯度下降法和小批量梯度下降法，这两种方法使用样本全集中的一个或部分样本来更新参数，这样操作使得每一次并不是按照严格意义上的最优方向来更新参数，但是从整体来看，依旧是朝着负梯度的方向更新参数，这两种梯度更新方式使参数学习的速度大大提高，适用于大规模训练样本的情况。

### 2.4.3 BP网络前向传播和反向传播



图2.7 典型三层BP神经网络

一个典型的三层BP网络如图2.7所示，第一层是输入层，共有个输入神经元；第二层是隐层，共有个隐层神经元，第三层是输出层，共有个输出神经元。表示第个输入神经元与第个隐层神经元的连接权值，表示第个隐层神经元的偏置，表示第个隐层神经元与第个输出神经元的连接权值，表示第个输出神经元的偏置，表示第个输入神经元的输入特征值，表示第个隐层神经元的输出，表示第个输出神经元的输出，表示第个隐层神经元的输入，表示第个输出神经元的输入。

BP网络的前向传播是指将上一层的输出与层间的对应权值相乘并求和，最终经非线性函数的处理，得到下一层对应神经元的输出值。常用的非线性函数有sigmoid函数(如公式2-9所示)、tant函数(如公式2-10所示)、ReLU函数(如公式2-11所示)等。

 公式2-9

 公式2-10

 公式2-11

如图2.7所示，前向传播过程如公式2-12和公式2-13所示，设使用sigmoid函数。

 公式2-12

 公式2-13

对于第个训练样本，，，设神经网络对于该样本的输出为，使用均方误差损失函数，则对样本的损失函数如公式2-14所示，对整个样本的损失函数如公式2-15所示，其中第一项表示在所有样本上的平均损失，第二项表示正则项，表示所有参数组成的向量，正则项可以避免过拟合。Sigmoid函数对参数求导结果如公式2-16所示。BP网络反向传播的目的是求得使损失函数最小时的参数，使用梯度下降法进行求解。

 公式2-14

 公式2-15

 公式2-16

根据链式求导法则，损失函数对参数和的导数分别如公式2-17和公式2-18所示。

 公式2-17

 公式2-18

在样本上，参数的更新如公式2-19所示。

 公式2-19

BP误差反向传播算法的基本过程为：

**输入：**训练集，学习率.

**输出：**连接权值与偏置值确定的多层前馈神经网络.

**过程：**

产生(0,1)范围内的随机数，初始化所有连接权值和偏置值；

Repeat

For all ：

根据当前参数执行前馈传播，计算当前样本的输出值；

根据公式(1.15)进行反向传播，更新当前参数；

End for

Until 达到终止条件

## 2.5 本章小结

本章首先介绍了相似重复记录的概念以及产生的原因，然后介绍了常用的基于单字段和多字段的相似度匹配算法。

在第三小节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算法等进行了介绍说明。

最后介绍了神经网络中的神经元模型、BP神经网络更新所采用的梯度下降法、BP神经网络的正向传播和反向传播的过程等。

# 第三章 改进的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，并且通过计算传递闭包可以检测到一些人工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。所以本文在MPN算法基础上提出了一种改进算法IMPN，它的改进点在于：基于统计规律的提取关键字的IMPN算法，通过统计不同字段对于数据集记录的区分度，可以一定程度上克服MPN对于专家知识的依赖；对排序关键字为空或者不完整的记录进行标记，初步检测完成之后再对这些标记记录进行归并处理，提高了算法的鲁棒性。MPN算法在归并过程中采用的滑动窗口方法是继承自传统的SNM算法的，即固定大小的滑动窗口。这种方式不够灵活，如果当前窗口内含有较少相似重复记录时则会出现较多的无效比较，IMPN算法对于MPN算法的第三个改进点就在于采用了可伸缩大小的自适应滑动窗口，这种方式能够减少一些不必要的判等比较，从而一定程度减少了时间消耗。

## 3.1 基于字段区分度提取关键字的方法

### 3.1.1 传统的提取关键字的方法

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关键字，然后再对记录按照关键字的顺序进行排序。完全理想状态下相似重复记录的关键字也相等，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。

关键字的定义是，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。关键字的选择十分关键，因为只有选择恰当的排序关键字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，合适的排序关键字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，排序关键字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。

接下来介绍三种不同的生成排序关键字的方案：【方案一】、【方案二】以及【方案三】，并说明每种方案下生成的排序关键字对整个算法可能造成的影响。

表3.1 四条相似重复记录的例子

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

举例来说，表3.1是四条相似重复记录，，，。它们由四个字段组成，分别是First Name、Last Name、Address和ID。可以看出和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；和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；然而和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。

假设【方案一】的关键字设计者选择的排序关键字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关键字分别如表3.2所示：

表3.2 相似重复记录及其生成的关键字

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

则可以看出该方案下，、，对应的关键字相同均为“STOLFO123FRT123”，和对应的关键字不同，经过排序后和聚集在邻近位置，和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。和原本是相同的记录，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，很可能会导致这对重复记录无法被正确地检测出来。

假设【方案二】的关键字设计者选择的排序关键字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关键字分别如表3.3所示：

表3.3 相似重复记录及其生成的关键字

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

则可以看出该方案下这四条记录的排序关键字均为“STLJAC123FRT213”。然而从表格中可以看出：和的排序关键字的相同是合理的；和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克服发音上的错误，和的排序关键字的相同也是合理的；然而和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】会导致和与其本身并不相同三条记录、、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与、、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】的关键字设计者选择的排序关键字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关键字分别如表3.4所示：

表3.4 相似重复记录及其生成的关键字

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

则可以看出该方案下，、、的排序关键字相同，它们对应的现实实体也是一致的。和则对应不同的排序关键字。这种情况下可以看出【方案三】是最理想的一种关键字选择。

综合以上的三种情况不难发现：对于特定的数据集，生成排序关键字的方式优劣取决于该数据所在领域的专家经验。比较合适的排序关键字才能达到较为理想的检测效果和检测效率，这就需要较多的人为参与，工作量大且影响了算法的普适性。本章3.1.2提出了一种较少依赖专家知识的方法。

### 3.1.2 改进的字段区分度方法

衡量排序关键字优劣的一个原则即为不同的记录对应的关键字不同，相同的记录对应的关键字理应相似或一致。对于相似重复的记录，无论排序关键字以何种方式选取，理论上生成的key值是相同或者相似的，所以衡量排序关键字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关键字方案下生成的key值应当不一致。所以在选择排序关键字时，应当选择有足够区分度的字段来提取关键字。

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。若数据库中一共有条记录，每条记录都个字段组成，即，对于第个字段，它的区分度如下公式3-1：

 公式3-1

其中，代表数据集中在字段一共有种取值，也就是将数据集按照进行聚类一共有簇。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。提取排序关键字之前首先计算所有字段的“字段区分度”值然后对其排序，排序关键字每次选择值最大的一个字段的部分或者整体来作为关键字的一部分。

这里的“字段区分度”有效的前提在于，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，这也是较少部分的情况。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。但是若数据集的质量过差，即错误记录以及缺失记录所占的比例过高时，“字段区分度”的效果将接近于随机选择方法。

在没有人工参与或者较少参与选取排序关键字的情况下，利用“字段区分度”去选择排序关键字的生成方式，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 自适应大小的滑动窗口检测方法

### 3.2.1 传统的滑动窗口检测方法

MPN算法的滑动窗口方法是继承自传统的SNM算法的。首先给定一个固定大小为的滑动窗口如下图3.1所示：



图3.1 滑动窗口扫描过程

在排序之前首先要将数据源合并成一个线性序列，然后将这个滑动窗口从第一条记录开始，每次向下滑动一条记录同时原来窗口的第一个记录被移出，直到最后一条记录进入窗口内。滑动过程中重复执行如下操作：刚进入的一条记录分别于之前的条记录进行比较，判断是否是相似重复记录。需要注意的是，初始化大小的滑动窗口之后，需要将最开始的这条记录互相之间重复检测，因为后面的滑动过程中忽略掉了这前条记录。

这种传统的固定滑动窗口存在一些问题，可以看出来滑动归并过程的时间复杂度为，所以如果窗口的长度过大则会增加算法的的复杂度，长度如果太小又会降低检测的精度。

如下图3.2所示：



图3.2 不同滑动窗口大小时的归并过程

图中的record0\_origin是原始记录，record0\_dup0、record0\_dup1、record0\_dup2是其相似重复记录，而record1\_origin和record2\_origin是另外的两个不重复记录。若滑动窗口的大小为2（这里为了举例说明而采用了一个比较极端的例子），按照传统的滑动窗口归并方法，无法检测到(4)与(1)、(2)是重复记录；若滑动窗口的大小为，虽然能够实现最好的检测效果，但是也最耗时间。可以看出(1)、(2)这段数据中重复率最高所以这时的滑动窗口应当较大，而当窗口移动到(3)、(4)、(5)的时候，记录互相不重复，所以这里窗口应当较小，以节约时间。所以总的来看MPN采用的这种滑动窗口方式还有很大的优化空间，一个较为合理的方案应当是根据数据内容动态变化滑动窗口的大小。

### 3.2.2 改进的自适应大小的滑动窗口

滑动窗口的大小可以根据当前滑动窗口的数据重复情况而做出动态地调整，当窗口内的数据重复度比较高时，证明当前窗口整处于重复记录比较集中的位置附近，而且重复记录的数量可能更多，所以为了实现更精准的检测，窗口应当增大尺寸以包含位置比较远的相似重复；反之，若滑动窗口内的数据集重复度比较低，即相似重复数据较少时，说明当前位置附近的数据之间可能互相不重复，所以如果窗口较大的话就会造成这些互不重复的数据需要数次与新进入的数据进行判重检测之后才能退出滑动窗口，所以这时应当减小滑动窗口的大小。

现有的自适应窗口算法[[[32]](#endnote-33)]大多是统计出当前滑动窗口内的所有的重复记录数，依据重复数据的个数占滑动窗口的大小来决定当前窗口的尺寸变化，此类方式存在很大的弊端：原本MPN的归并过程的时间复杂度只有，其中是滑动窗口的大小，是待检测的相似重复记录集合中记录的条数，但是这种方法每次要对窗口内的记录进行一对一的重复检测，使得时间复杂度升高到。

考虑到MPN原本在滑动窗口内的比较是将新进入的一条记录与窗口内仍保留的-1条记录进行比较，当重复度较高的时候，下一次比较应当保证的是新进入的记录应该能够尽可能的照顾到即将要离开的记录，才能保证检测效果。换个角度说，在当前滑动窗口内，即将离开滑动窗口的记录若和新进入的元素是相似重复记录，这时应当扩大窗口尺寸。设窗口内的重复记录在窗口内的位置依次为0、1、2...-1，其中-1是刚滑入的数据，0号位置的记录是即将被滑出的记录，则越是靠近0号位置的数据对于滑动窗口的尺寸影响越大。根据以上条件本文给出公式如下3-2：

 公式3-2

其中，滑动窗口的最大值为，最小值为，即将滑出的记录在数据集中的索引位置为，当前滑动窗口的大小为，代表数据集中索引为的记录是否与位置的记录互为重复，若是则=1，否则=0。可以看出，若内的记录都是重复记录，则下一个窗口的大小去最大值，相反若互不相似，则取最小值；并且距离越远位置的记录对下一个滑动窗口大小的影响越大（当其与位置的记录互为重复记录时）。

如下图3.3所示：

图3.3 自适应大小的滑动窗口

其中R0、R0\_0、R0\_1、R0\_2、R0\_3、R0\_4互为相似重复记录，剩余的记录R1、R2、R3互不重复。归并过程从左向右进行。最大窗口为5，最小窗口为3，第①步窗口大小为4，到了第②步，由公式计算得到窗口变为5，因为窗口内的记录和R0\_2全部重复，第③步窗口大小又变成最小3，因为窗口内的数据和R1互不重复，④⑤⑥的变化同理。图上的例子简明生动地表现了自适应滑动窗口的变化过程，展现了IMPN算法相对于传统MPN算法更加灵活高效的处理滑动窗口的方式。

## 3.3 基于标记的处理不完整排序关键字方法

### 3.3.1 MPN的排序方法存在的问题

由于待清洗的数据集本身的数据质量并不高，所以记录中可能存在字段为空或者字段不完整的情况，如表格3.5所示：

表3.5 缺失数据及不完整数据示例

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Record | First Name | Last Name | Address | ID |
|  | Jack | tolfo | 123 First Street | 12345678 |
|  | Jack |  | 123 First Street | 12345673 |
|  | Jack | Stolpho | 123 First Street | 12345678 |
|  | Jacon | Stiles | 123 Forest Street | 12345432 |

记录R1的Last Name字段原本应该是“Stolfo”，现在缺失了一部分成为“Sto”，在生成排序关键字时若采用3.1.1中的方案三，即提取Last Name的前三个辅音字母是则只能得到“s”和“t”两个字母，则排序关键字位数少了一位，所以生成的关键字为“TLFJCK123FRT213”。记录R2的Last Name字段则直接完全缺失，所以提取的关键字为“JCK123FRT213”，在排序的过程中，本来属于相似重复记录的R1和R2则由于开头字母的差异而被分到距离比较远的位置，从而可能无法被检测到。这个例子展示了不完整数据和缺失数据对MPN算法检测效果的影响，本文3.2.2提出了针对不完整排序关键字的改进方法。

### 3.3.2 改进的基于标记的方法

缺失数据的处理是数据清洗的另一个分支研究领域，面对缺失值常见的做法主要有三种：

（1）以同一指标的计算结果（包括均值、中位数、众数等）填充缺失值

（2）以业务知识或者经验填充缺失值

（3）从本数据集或者其他来源的数据集推测出来

其中第（1）种做法填充结果不够精细甚至过于粗糙，对检测结果可能造成负面影响；第二种做法填充结果可能较为准确但是需要人工干预，工作量较大；第三种做法对数据集的数据质量要求较高并且能达到的效果下限很低。考虑到出现排序关键字为空的记录主要是影响到排序后记录的位置，所以本文提出了基于标记的处理方法：先将遇到的带有不完整关键字的记录的ID进行标注，然后在其余完整关键字的记录归并结束之后，对这些标注过的记录分别进行检测，将其一一聚类到已知的重复数据簇中。

本文这种基于标记的做法能弥补MPN算法面对缺失排序关键字情况检测效果较差的不足，但是对于对含有缺失字段的记录占数据集比例较低的数据集合进行操作时不会耗费太多时间，但是当数据集中的大部分记录都含有关键字段的缺失时，该方法会导致时间复杂度升高，这是一个不足之处。

## 3.4 IMPN算法设计

### 3.4.1 算法流程设计

结合3.1~3.3的内容可以看出，IMPN算法的改进思想在于以下三点：

（1）引入“字段区分度”概念减弱算法对专家知识的依赖。

（2）采用可伸缩大小的滑动窗口提高算法的查全率，减少不必要的比较次数。

（3）通过标记含有不完整排序关键字的记录，提高了算法的查全率。

设有趟SNM过程的IMPN算法，则算法的步骤如表3.5所示：

表3.5 含有N趟SNM过程的IMPN算法步骤表

|  |
| --- |
| **IMPN算法** |
| 1.读取数据集，得到待检测的数据；  2.计算数据集字段的字段区分度并排序；  3.优先选取区分度较大的字段去生成组排序关键字；  4.独立地执行步骤5~8次；  5.按照排序关键字的产生方式对每条记录提取其排序关键字；  6.对数据集按照关键字排序，如果某条记录的不完整或者为空则将该记录的ID加入到缺失关键字记录集合中，完整则正常排序；  7.进行可伸缩大小的滑动窗口重复检测得到重复集合；  8.将与进行重复归并，然后计算此集合的传递闭包；  9.将次SNM重复检测得到的集合进行归并，然后计算传递闭包得到最终的重复记录集合。 |

含有两趟SNM过程的IMPN算法流程如下图3.4所示：

图3.4 含有2趟SNM过程的IMPN算法流程图

### 3.4.2 时间复杂度分析

接下来分析说明MPN算法和IMPN算法各自的时间复杂度。在相似重复记录检测算法的处理过程中，理想状态下所有的数据都可以在内存中处理，不考虑磁盘I/O的情况。

MPN算法首先创造排序关键字需要对数据集进行整体遍历，所以该阶段的时间复杂度为；排序过程采用快速排序，算法时间复杂度为；滑动窗口的归并检测过程相当于比较了次，所以该阶段的时间复杂度为；其中代表待检测数据集中的记录总数，代表滑动窗口的大小；在传递闭包的计算过程中，假设重复记录数据集的大小为，则该阶段的时间复杂度为。所以对于MPN算法来说，总的时间复杂度为：

 公式3-3

由公式3-3可以看出，IMPN算法首先需要对数据集中的所有字段进行区分度统计，假设每条记录的字段总数为，则区分度统计阶段的时间复杂度为；生成排序关键字过程、排序过程、以及滑动窗口归并过程与MPN算法的时间复杂度相同，分别为、、；设因为排序关键字为空而被标注的数据集包含记录数为，则与已检测识别出的数据集进行重复记录检测过程的时间复杂度为；传递闭包过程中的时间复杂度同理为，为检测结束后重复记录集合的大小。所以IMPN算法的时间复杂度为（一般情况下带有不完整排序关键字的记录数满足>k并且>）：

 公式3-4

观察公式3-3和公式3-4组成部分可以发现，两者结构相似且变量数量级一致，所以两者的时间复杂度在特定的数据集上是一致的。若用代表常数，两个公式都可以简化成下面的式子：

 公式3-5

当重复记录较多时，即的值较大，此时两个算法的时间复杂度均为，这体现出了MPN算法和IMPN算法的时间消耗受重复数据比例的影响均较大。

## 3.5 SNM、MPN、IMPN综合对比实验

### 3.5.1 实验数据介绍

为了方便研究使用，本文实验采用的数据集是由第三方的数据生成器“febrl”[[[33]](#endnote-34)]（开源地址：[https://sourceforge.net/projects/febrl/）生成的。“febrl](https://sourceforge.net/projects/febrl/）生成的。)”的数据源是澳大利亚某卫生部门的数据库。生成数据集中记录包含的字段及字段含义如下表所示：

表3.6 记录字段说明

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名称 | 字段描述 | 举例1 | 举例2 |
| rec\_id | 记录ID | rec-454-org | rec-454-dup-0 |
| culture | 文化 | pak | pak |
| sex | 性别 | f | f |
| age | 年龄 | 30 | 30 |
| date\_of\_birth | 出生日期 | 19870221 | 19870221 |
| title | 头衔 | hon | hon |
| given\_name | 名字 | sophie | sophie |
| surname | 姓氏 | bozdar | bozdaa |
| state | 州 |  |  |
| suburb | 郊区 | holsworthy | holsworthy |
| postcode | 邮编 |  |  |
| street\_number | 街道号码 | 46 | 46 |
| address\_1 | 地址1 | thurgood court | thurgood court |
| address\_2 | 地址2 |  |  |
| phone\_number | 电话号码 | 08 42167414 | 08 42167414 |
| soc\_sec\_id | 社保ID | 3920942 |  |

采用数据生成器的好处在于：生成器公开的接口中提供了多个参数，这些参数能够方便用户自定义数据集的大小与特征等；相较于真实数据，生成的数据记录拥有唯一的标识符，更方便后期对算法的查准率、查全率等进行计算和评估；由该生成器得到的数据集非常接近现实数据。

“febrl”的公开参数列表及其说明如下表所示：

表3.7 febrl公开接口的参数说明

|  |  |
| --- | --- |
| 参数名称 | 参数说明 |
| outputFileName | 输出文件名（.CSV格式） |
| numberOfOriginalRecords | 原始数据集大小 |
| numberOfDupRecords | 由原始数据集生成的重复数据集大小 |
| maxNumOfDupPerRec | 一条原始记录能够最多生成的重复记录个数 |
| maxNumOfModPerRec | 一个字段最多可修改数目 |
| maxNumOfModPerRec | 一条记录最多可修改字段数目 |
| probabilityOfDup | 重复记录在数据集中的概率分布  （均匀分布、泊松分布或齐夫分布） |
| typeOfModification | 字段可能发生的错误类型（typo：印刷错误、ocr：  扫描错误、phonetic：发音错误或者以上所有） |

### 3.5.2 算法的衡量标准

衡量相似重复记录检测算法效果的常用标准主要有查全率（Recall），查准率（Precision）以及F-measure。

将算法的重复记录检测结果与实际数据集重复记录进行比较时，会出现以下四种可能的情况：

（1）True Positive（TP）：算法判定为重复记录，实际上也是重复记录；

（2）False Positive（FP）：算法判定为重复记录，但实际不是重复记录；

（3）True Negative（TN）：算法判定为非重复记录，实际也不是重复记录；

（4）False Negative（FN）：算法判定为非重复记录，但实际上却是重复记录。

查全率（Recall）代表了算法检测重复记录是否完备的能力，它的计算方法是算法检测出的正确重复数除以实际数据集中的重复记录总数，如下公式：

 公式3-6

查准率（Precision）代表了算法正确识别重复记录的能力，它的计算方法是算法检测出的正确重复数除以被算法识别为重复记录的总数：

 公式3-7

F-measure是查全率和查准率的加权调和平均，F-measure的计算公式如下：

 公式3-8

其中参数取值为1时，F-measure即为最常见的F1-measure：

 公式3-9

假设有8条记录、、、、、、、，其中{，，}和{，，}互为相似重复记录，若通过算法检测出的重复结果为{，，，}和{，，，}互为相似重复记录，则TP=4，FP=2，TN=2，FN=0，所以算法的查全率为4/(4+0)=100.00%，查准率为4/(4+2)=66.67%，F1-measure=2\*1\*0.6667/(1+0.6667)=80.00%。

### 3.5.3 实验设计与结果分析

本节实验主要是为了综合对比IMPN、MPN和SNM算法的查全率和查准率。实验之前首先确保三个算法中的常数要一致。令SNM算法和MPN算法中滑动窗口的大小=5，IMPN算法的最大滑动窗口大小=20，最小滑动窗口大小=3，IMPN和MPN算法过程中进行单趟SNM的次数=3，在字段相似度检测过程中用到的两个常数VERY\_CLOSE\_CONSTANT（字符串非常接近）、CLOSE\_CONSTANT（字符串比较接近），对它们分别赋值：VERY\_CLOSE\_CONSTANT=0.8，CLOSE\_CONSTANT=0.6。

数据集的大小与参数配置见下表：

表3.8 测试数据集说明

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 |  |  |  |  |  |  |  |
| dataset1 | 5000 | 1000 | 3 | 1 | 1 | uniform | phonetic |
| dataset2 | 10000 | 2000 | 3 | 1 | 1 | uniform | phonetic |
| dataset3 | 20000 | 4000 | 3 | 1 | 1 | uniform | phonetic |
| dataset4 | 50000 | 10000 | 3 | 1 | 1 | uniform | phonetic |
| dataset5 | 80000 | 16000 | 3 | 1 | 1 | uniform | phonetic |
| dataset6 | 100000 | 20000 | 3 | 1 | 1 | uniform | phonetic |
| dataset7 | 200000 | 40000 | 3 | 1 | 1 | uniform | phonetic |
| dataset8 | 500000 | 100000 | 3 | 1 | 1 | uniform | phonetic |

上表中的参数~代表的含义分别为：-原始记录数、-重复记录数、-单个记录最多重复数、-单个字段最多修改数、-单个记录最多修改字段数、**-**重复记录的概率分布、错误类型。

本实验的实验环境配置如下：

表3.9 实验环境配置

|  |  |
| --- | --- |
| 操作系统 | Windows 7旗舰版 |
| 处理器 | Core i7（8核心） |
| 内存大小 | 8G |
| JDK版本 | JDK1.8.0 |
| JVM配置 | -Xmx4096m（堆内存最大4G） |

由上述8个数据集（dataset1~dataset8）分别运行SNM算法、MPN算法以及IMPN算法，得到的算法查全率对比表格如下：

表3.10 SNM、MPN、IMPN算法查全率对比表

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集大小（千条） | SNM算法（%） | MPN算法（%） | IMPN算法（%） |
| 5 | 80.60 | 94.80 | 98.80 |
| 10 | 78.15 | 91.80 | 98.40 |
| 20 | 78.03 | 89.28 | 98.75 |
| 50 | 78.01 | 87.39 | 97.90 |
| 80 | 76.38 | 82.08 | 98.58 |
| 100 | 75.85 | 80.76 | 98.73 |
| 200 | 73.83 | 79.72 | 99.10 |
| 500 | 72.90 | 78.91 | 96.91 |

三种算法的查准率对比表格如下：

表3.11 SNM、MPN、IMPN算法查准率对比表

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集大小（千条） | SNM算法（%） | MPN算法（%） | IMPN算法（%） |
| 5 | 100.00 | 99.89 | 99.50 |
| 10 | 99.94 | 100.00 | 99.54 |
| 20 | 99.97 | 100.00 | 98.78 |
| 50 | 99.87 | 99.98 | 96.93 |
| 80 | 99.81 | 99.92 | 95.69 |
| 100 | 99.92 | 99.87 | 94.77 |
| 200 | 99.73 | 99.89 | 89.70 |
| 500 | 99.53 | 99.70 | 88.17 |

三种算法的运行时间对比表格如下：

表3.12 SNM、MPN、IMPN算法运行时间对比表

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集大小（千条） | SNM算法（s） | MPN算法（s） | IMPN算法（s） |
| 5 | 0.503 | 0.938 | 1.274 |
| 10 | 1.196 | 2.320 | 2.961 |
| 20 | 3.540 | 6.985 | 10.327 |
| 50 | 17.169 | 26.347 | 59.420 |
| 80 | 47.255 | 239.434 | 197.439 |
| 100 | 67.173 | 372.245 | 290.903 |
| 200 | 237.805 | 1668.995 | 1391.415 |
| 500 | 1493.305 | 10464.894 | 11296.066 |

由查全率对比表可以画出三种算法的查全率折线图如下所示：



图3.5 SNM、MPN、IMPN算法查全率折线图

由上图可以看出：改进的IMPN算法面对不同大小的数据集（其他条件如滑动窗口大小、数据集特征参数等均一致），查全率均高于SNM算法和MPN算法，并且IMPN算法查全率的数值稳定在96%以上。所以证明了改进的IMPN算法在查全率上相较于传统的MPN算法拥有较大的提升；随着数据集的增大，SNM算法和MPN算法的查全率均呈下降趋势，而IMPN算法的查全率较为稳定。

由查准率对比表可以画出三种算法的查全率折线图如下所示：



图3.6 SNM、MPN、IMPN算法查准率折线图

由上图可以看出：面对不同的数据集，SNM和MPN算法表现较好，这是因为算法中采用的相似记录判等方法是基于当前数据集特征的判定规则，所以需要领域专家的经验和知识；IMPN算法在检测大小从5000到200000的数据集时，查准率均在90%以上，属于尚可接受的范围，但是当数据集大小达到500000时，查准率低于90%，所以IMPN算法的查准率还有待改进的空间。本文的第四章针对这一问题提出了新的改进思路，并取得了良好的效果。

由运行时间对比表可以画出三种算法的运行时间折线图如下所示：



图3.7 SNM、MPN、IMPN算法运行时间折线图

由上图可以看出：SNM算法时间消耗最低，因为MPN和IMPN算法均包含多趟独立的SNM过程，并且含有传递闭包的计算；与MPN相比较，IMPN算法采用了自适应大小的滑动窗口，这一改进是可以减少时间消耗的，但是IMPN算法还包括对含有空（或者不完整）排序键值的记录进行归并的过程，所以总体的时间和MPN算法差距不大。

由以上表格中的查全率和查准率可以计算出三种算法各自的F1-measure值如下表所示：

表3.13 SNM、MPN、IMPN算法F1-measure值对比表

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集大小（千条） | SNM算法（%） | MPN算法（%） | IMPN算法（%） |
| 5 | 89.258 | 97.278 | 99.149 |
| 10 | 87.712 | 95.725 | 98.967 |
| 20 | 87.648 | 94.336 | 98.765 |
| 50 | 87.597 | 93.262 | 97.413 |
| 80 | 86.537 | 90.126 | 97.114 |
| 100 | 86.237 | 89.304 | 96.709 |
| 200 | 84.847 | 88.672 | 94.166 |
| 500 | 84.159 | 88.095 | 92.334 |

由上表可以画出三种算法的F1-measure数值折线图如下：



图3.8 SNM、MPN、IMPN算法F1-measure折线图

综合以上结果可以看出，IMPN算法相对于MPN算法较为明显的改进在于：在和MPN消耗时间差距非常小的同时实现更高的查全率；算法的缺点也很明显，即随着数据量的增大，相似重复记录判等方法表现较差，导致查准率的下降。但是由F1-measure折线图可以看出，IMPN算法的综合表现要好于MPN。

## 3.6 本章小结

本章首先介绍了MPN算法的缺点，然后在此基础上提出了一种改进的算法IMPN，并从三个方面介绍了传统MPN的处理方式的不足以及IMPN算法的处理方法。之后介绍了IMPN算法的设计与流程步骤，最后通过采用数据生成器febrl生成的CSV数据集进行对比实验验证了IMPN算法的查全率较高的优势，但是也反映了IMPN算法查准率不够理想的问题。

# 第四章 基于遗传神经网络改进的IMPN算法

从第三章的实验结果可以看出，IMPN算法查准率不够理想，其最根本原因在于算法依旧是“排序/归并”的本质思想，在归并过程中对滑动窗口内的记录进行判等时采用的算法准确率不够，被误识别为重复记录的数据较多。

第二章介绍了一些常用的字段匹配算法，包括编辑距离法、S-W算法、N-Gram算法等，目前常见的做法是利用这些算法计算得到两条记录的相似度值，如果大于相似度阈值，则断定两条记录相似。这种方式的缺点主要有两个：首先是这些字段的相似度算法往往只对某种特定情况的字段特别有效[[[34]](#endnote-35)]，普适性较差；其次是整条记录的相似度和某字段的相似度并不是简单的线性关系，所以即便是对不同的字段赋予不同的权重，也可能会出现判断失误的情况，况且如何合理地对字段分配权重也是一个复杂的问题，这也会导致时间消耗增加。所以理论上最好的方法是通过学习给定的数据集合的特征，利用数据集记录和字段的内在关系来判断记录之间是否重复。

ANN（Artificial Neural Network，人工神经网络）是一种由大量处理单元组成的自适应系统，它的数学计算模型模仿生物神经网络的结构和功能，可以任意精度逼近线性或者非线性函数，从而可以表征真实社会中更加复杂的问题。人工神经网络在相似重复记录检测领域也有诸多应用，本章4.1节首先介绍了一种遗传神经网络算法[[[35]](#endnote-36)]，采用BP（Back Propagation，反向传播）神经网络，通过基于监督的学习实现对训练数据集的相似性关系的非线性模拟，训练完成后即可对测试数据集的相似性进行预测，并通过引入遗传算法克服BP神经网络易陷入局部最优值的缺点。

本章将遗传神经网络与IMPN算法结合，将该算法应用到IMPN的滑动窗口归并过程的相似性判断中，提高了IMPN算法的查准率。

## 4.1 遗传神经网络用于相似重复记录检测

### 4.1.1 BP神经网络的设计

误差反向传播算法是一种前馈神经网络，它具有学习方式较为简单的优点，应用到结构简单层数少的网络模型中时收敛较快等优点，是一种目前应用最为广泛的有监督学习算法。接下来分析介绍BP神经网络结构的设计思路。

一般情况下的BP神经网络层数多采用3层或者以上，包括输入层、输出层、隐含层，定理[[[36]](#endnote-37)]指出：三层的反向传播神经网络能够以任意的精度逼近非线性函数。所以设计BP神经网络应当优先考虑增加隐含层结点的个数而不是增加网络层数，网络层数的增加在降低误差的同时也带来了网络复杂度的增加，这会使得训练过程花费的时间大大增加，所以这里的设计采用三层的BP神经网络对相似重复记录进行处理。

输入层的结点数由数据集中记录的字段数目确定，举例来说，第三章中实验部分生成的数据集（表3.7）所示，每条记录都有16个字段，但是第一个字段“rec\_id”是为了方便计算算法的查重效果而设计的，所以输入层应当有15个结点（不含一个偏置结点）；输出层结点为1，输出的结果介于0到1之间，表示输入向量对应的两条相似重复记录的相似程度；隐含层结点数目的计算由公式4-1[[[37]](#endnote-38)]计算得到。

 公式 4-1

其中代表隐含层结点的个数，代表输入层结点的个数，代表输出层结点的个数，计算过程的结果需要四舍五入进行修整。在本例中，=1，=15，=7，=24，所以计算得到隐含层结点数=24。

激活函数的选择有较多种这里采用常用的Sigmoid函数，即。网络的初始权值的选择较为关键，它影响着网络的学习是否达到局部最小、最终是否能够收敛，甚至包括训练时间的长短。一般最理想的情况是初始化赋值之后，每个神经元的输出接近于0，这样可以在S型激活函数导数最大的地方进行调节与变化，一般取初始值在之间的随机数。

学习速率决定了每次训练之后权值更新的幅度，过大的学习速率会导致系统震荡，过小的学习速率又会导致训练学习时间变长，收敛速度也变得很慢。一般倾向于选择较小的学习速率，因为这样能保证系统稳定性而不会导致修改幅度过大，取值范围一般在0.01~1之间。期望误差代表着当训练后的误差结果在可接受范围之内则人为进行收敛，停止训练。

表4.1对应例子的3层神经网络的拓扑结构如图4.1：



图4.1 神经网络的拓扑结构

当网络结构确定以后需要提取训练数据集以及网络理想输出对BP神经网络进行训练，训练的目的是为了找到最优情况下的权值以及阈值，通过理想输出和实际输出的误差，然后沿梯度方向修改更新网络权值。使用BP神经网络进行相似重复记录检测的算法流程图如图4.2所示：



图4.2 BP网络检测相似重复记录流程图

其中，停止训练的条件除了上图中的条件（实际输出和目标输出的偏差值在可接受范围之内）外，还可以是达到了人为设定的迭代次数或者权重的更新低于某个阈值。

### 4.1.2 基于遗传算法改进的神经网络

BP神经网络通过训练能够有效地对测试数据集中的相似重复记录进行检测判断，解决了大数据量情况下传统的基于“排序/归并”思想的检测算法检测效果较差的问题，并且拥有较好的适应性。但它也有一个很明显的缺点是容易陷入局部最优，并且收敛速度慢。GA（Genetic Algorithm遗传算法）是一种模拟生物进化过程的进化算法，它利用“物竞天择，适者生存”的进化思想，通过选择、交叉、变异等操作不断演化，并在演化过程中淘汰掉适应度较差的个体，经过数代进化之后产生该环境条件下适应度较高的个体，也就是找到全局的最优解。遗传算法拥有较强的全局搜索能力，能够很好地跳出局部最优解，所以可以将其应用到BP神经网络中，以解决反向传播算法容易陷入局部最优的问题[[[38]](#endnote-39)][[[39]](#endnote-40)]。

遗传算法的编码方式有两种：实数编码、二进制编码、矩阵编码、树形编码等[[[40]](#endnote-41)]。这里的染色体个体代表的是一个训练好的BP神经网络的参数和阈值组合，所以采用实数编码更容易找到问题的最优解，并且在遗传变异过程中无需解码操作，可以提高运行速度。如表4.1中的例子，经过训练收敛之后的一个BP神经网络，将其转化为染色体，由于输入层一共4个结点，隐含层一共6个结点，加上输入层一个偏置，所以输入层和隐含层之间的权值一共(4+1)\*6=30个，输出层和隐含层之间的权值一共7个，所以网络的总权值为37个，加上1个阈值可以得出染色体的总长度为38。

适应度函数衡量染色体对于环境的适应程度，在重复检测应用中，适应度函数应当能反映出相似重复检测结果和真实结果的差距，差距越小适应度越大，差距越大则适应度越小。这里的适应度函数定义为，其中代表全局误差，的计算公式如下：

 公式4-2

其中，代表训练数据集的记录总数，代表期望的输出结果，代表实际的输出结果。

遗传算法的三大基本算子包括选择、交叉、变异。选择是为了产生群体的下一代，采用轮盘赌算法能够保证种群中适应度函数值越大的个体有更大的概率进入下一代。轮盘赌算法的计算公式如下：

 公式4-3

其中代表种群中个体的数量，代表第个染色体的适应度函数的值。轮盘赌操作得到新的一代种群，但目前的种群相对于上一代只是增大了上一代中的适应度值高的个体所占的比例。

接下来需要进行交叉和变异以产生新的个体，也可能会出现适应度更好的个体，这两步操作即是遗传算法可以打破局部最优的根本原因。由于采用的是实数编码，所以在交叉操作时交换的是两条染色体上相同位置的两个实数。变异算子则是模仿的基因突变的过程：染色体某个位置上的基因突变成为其等位基因，从而可能引发性状表现上的变异。为了使群体能延续表现更好的个体的性状，在交叉和变异时，适应度值高的个体交叉和变异概率较低。对于适应度为的个体，在进行交叉或者变异时的概率取值为：

 公式4-4

其中、代表变异率（或者选择率）的上、下限，、代表当前代的种群中适应度的最大、最小值。若用表示一个个体，，，则个体中的每个位置上的基因的变异过程为：

 公式4-5

其中表示变异后的值，是0到1之间随机数，表示当前是迭代过程中的第代，的取值如下所示：

 公式4-6

其中是取值在0到1之间的随机数，是最大迭代次数。取值的特点是，随着的增加而越发接近于0，这就可以使得变异在刚开始较大而后逐渐拘于局部，有利于种群的稳定性。

使用遗传算法改进的BP神经网络进行相似重复记录检测的流程图如下：



图4.3遗传神经网络检测相似重复记录流程图

## 4.2 遗传神经网络对IMPN算法的改进

### 4.2.1 IMPN算法查准率问题

“排序/归并”算法在滑动窗口归并过程中，需要对邻近位置的记录进行相似判断。传统的MPN算法采用的是基于专家经验知识的规则产生式系统（OPS5），IMPN算法则是利用基于编辑距离的字段加权方式，区分度高的字段拥有更大的权值。

对两条记录、（均包含个字段）进行相似性判断的步骤如下：

表4.1相似记录判断步骤表

|  |
| --- |
| **相似记录判断算法** |
| 1.计算记录和所有字段的编辑距离；  2.由字段区分度从小到大的顺序对字段编辑距离进行排序；  3.排序完成，将所有的编辑距离值组合得到相似度向量；  4.对每个字段分配权值，保证字段区分度大的字段，其权值也高；  5.计算两条记录的相似度：；  6.将两条记录的加权相似度值和阈值进行比较，如果，则两条记录为相似重复记录，否则断定两条记录为不重复记录。 |

从第三章图3.6可以看出，随着数据量的增大，IMPN算法的查准率逐渐下降，这一问题严重影响了IMPN算法的检测效果，这也是导致图3.8中，IMPN算法的F1-measure值逐渐下降的原因。而图3.6中的SNM算法和MPN算法的查准率一直保持较高水准，这是因为算法采用的重复记录判等方法是基于数据集所在的领域知识和专家经验的，并且适用于数据生成器“febrl”生成的数据集。但是这就限制了算法的扩展性和通用性，本章4.2.2节将遗传神经网络检测重复记录的方式与IMPN算法结合，提高了算法的查准率。

### 4.2.2 遗传神经网络用于IMPN算法

在本章4.1节介绍了现有的遗传神经网络算法进行相似重复记录检测的方式：首先进行有监督的训练学习得到组BP网络，然后初始化含有条染色体的种群，经过选择、交叉、变异等操作获取全局最优的BP网络，最后对待检测数据集合进行时间复杂度为的重复检测。因为训练过程和检测过程的时间复杂度较高，虽然检测查准率和查全率都较为理想，但整个过程的时间消耗让用户难以忍受。

考虑到IMPN算法（除去重复记录判断过程）可以使得相似重复记录聚类在邻近位置，所以可以使用训练好的BP神经网络对滑动窗口内的记录进行重复判断，从而避免了复杂度为的重复检测过程。所以使用遗传神经网络改进IMPN算法的本质是使用IMPN算法对数据集进行预处理，然后再使用训练完成的BP神经网络进行重复检测。

改进的IMPN算法步骤如下：

表4.2基于遗传神经网络改进的IMPN算法步骤表

|  |
| --- |
| **基于遗传神经网络改进的IMPN算法** |
| 1.读取数据集，得到待检测的数据以及训练数据集；  2.计算数据集字段的字段区分度并排序；  3.优先选取区分度较大的字段去生成组排序关键字；  4.随机初始化出组BP神经网络；  5.对组BP网络进行迭代训练；  6.将训练完成的组BP网络初始化种群；  7.使用遗传算法的选择、交叉、变异等操作求种群的最优化解；  8.独立地执行步骤9~12次；  9.按照排序关键字的产生方式对每条记录提取其排序关键字；  10.对数据集按照关键字排序，如果某条记录的不完整或者为空则将该记录的ID加入到缺失关键字记录集合中，完整则正常排序；  11.采用进行可伸缩大小的滑动窗口重复检测得到重复集合；  12.将与进行重复归并，然后计算此集合的传递闭包；  13.将次SNM重复检测得到的集合进行归并，然后计算传递闭包得到最终的重复记录集合。 |

算法的流程图如下：



图4.4基于遗传神经网络改进的IMPN算法流程图

使用遗传神经网络进行训练迭代得到的最优BP神经网络的权值与阈值可以保存起来供下次使用，避免了相同性质的数据集进行重复记录检测时每次都要进行训练的过程，可以减少时间消耗。

## 4.3 对比实验

本节将对IMPN算法和采用遗传神经网络改进的IMPN算法（简称IMPN改进算法）进行综合对比，实验设计和数据集设定同第三章3.5节相同，均采用8个不同大小但是相同生成方式（生成器“febrl”的参数一致）的数据集做测试数据，训练过程采用的数据集大小为500，使用“febrl”生成，参数与测试数据集一致，详情见表3.7与表3.8。两者最大滑动窗口大小=20，最小滑动窗口大小=3，算法过程中进行单趟SNM的次数=3。

在BP神经网络的训练过程中，学习速率设置为0.05，动量系数设置为0.9[[[41]](#endnote-42)]，网络最大迭代次数设置为5000；遗传算法种群初始大小150，交叉概率=0.3，变异概率=0.05，种群最多迭代300次。

查全率的对比实验结果如下表所示：

表4.3 IMPN与改进的IMPN算法查全率对比表

|  |  |  |
| --- | --- | --- |
| 数据集大小（千条） | IMPN算法（%） | IMPN改进算法（%） |
| 5 | 98.80 | 98.50 |
| 10 | 98.40 | 98.50 |
| 20 | 98.75 | 98.90 |
| 50 | 97.90 | 98.35 |
| 80 | 98.58 | 98.14 |
| 100 | 98.73 | 97.76 |
| 200 | 99.10 | 96.85 |
| 500 | 96.91 | 94.83 |

查准率的对比实验结果如下表所示：

表4.4 IMPN与改进的IMPN算法查准率对比表

|  |  |  |
| --- | --- | --- |
| 数据集大小（千条） | IMPN算法（%） | IMPN改进算法（%） |
| 5 | 99.50 | 100.00 |
| 10 | 99.54 | 100.00 |
| 20 | 98.78 | 100.00 |
| 50 | 96.93 | 99.87 |
| 80 | 95.69 | 99.85 |
| 100 | 94.77 | 99.88 |
| 200 | 89.70 | 99.68 |
| 500 | 88.17 | 99.21 |

运行时间的对比实验结果如下表所示：

表4.5 IMPN与改进的IMPN算法运行时间对比表

|  |  |  |
| --- | --- | --- |
| 数据集大小（千条） | IMPN算法（s） | IMPN改进算法（s） |
| 5 | 1.274 | 21.110 |
| 10 | 2.961 | 83.169 |
| 20 | 10.327 | 336.404 |
| 50 | 59.420 | 2172.766 |
| 80 | 197.439 | 5510.201 |
| 100 | 290.903 | 8505.129 |
| 200 | 1391.415 | 32842.004 |
| 500 | 11296.066 | 204281.991 |

由查全率对比表可以画出IMPN算法与改进的IMPN算法的查全率折线图如下所示：

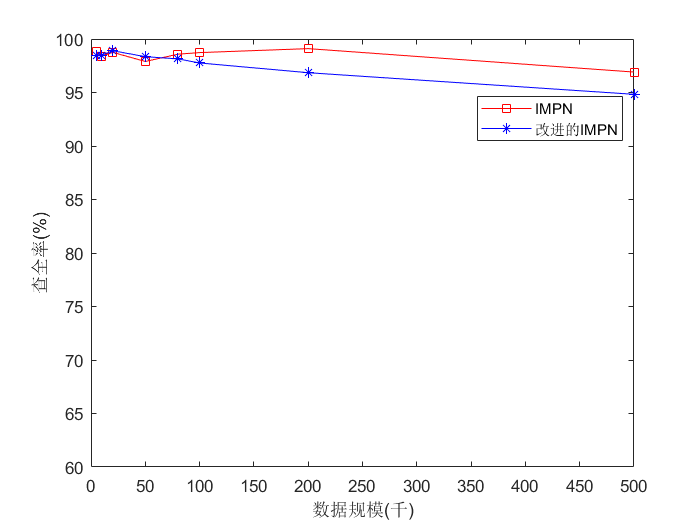


图4.5 IMPN、改进的IMPN算法查全率折线图

由上图可以看出，基于遗传神经网络改进的IMPN算法与IMPN算法相比，查全率没有太大差距，证明改进的IMPN算法可以保证IMPN算法较好的查全率。

由查准率对比表可以画出IMPN算法与改进的IMPN算法的查准率折线图如下所示：

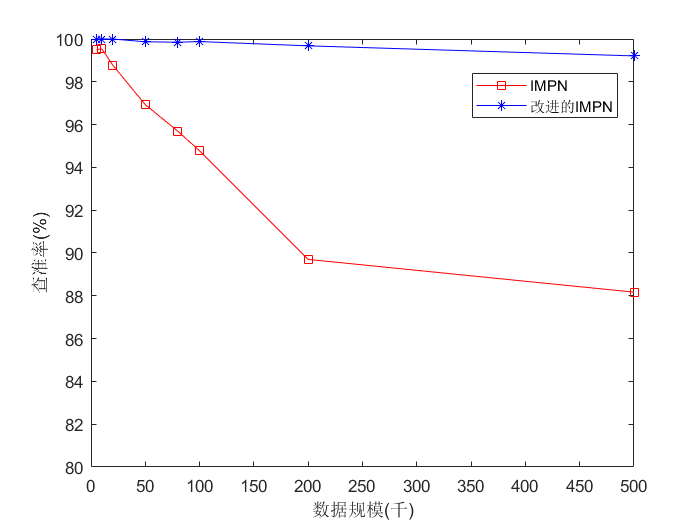


图4.6 IMPN、改进的IMPN算法查准率折线图

由上图可以看出，基于遗传神经网络改进的IMPN算法在查准率方面取得了较好的效果，相对于IMPN算法有了较大的提高。

由运行时间对比表可以画出IMPN算法与改进的IMPN算法的运行时间折线图如下所示：

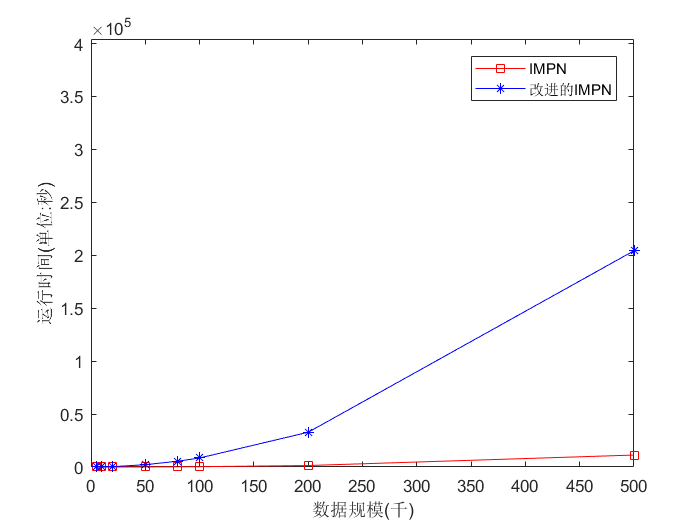


图4.7 IMPN、改进的IMPN算法运行时间折线图

由上图可以看出，改进的IMPN算法的时间消耗远高于IMPN算法。这是由于在相似记录进行重复判断时，IMPN算法的时间复杂度是常数级的，而改进的IMPN算法在判等时，若BP神经网络的输入层结点个数为，隐含层结点个数为，则时间复杂度为。

综上，基于遗传神经网络改进的算法在保证了IMPN算法较好查全率的基础上，弥补了算法查准率较低的缺陷，但是以牺牲时间为代价。取得此实验结果的理论原因在于：两者最主要的区别即是相似重复记录判等阶段，改进的算法采用遗传算法改进BP神经网络，然后通过学习训练数据集字段相似与整体记录相似的非线性关系，可以较为准确地对两条记录相似与否进行预测和判断，但同时神经网络前向传播的计算时间比IMPN算法采用基于权重的判等过程时间复杂度高，导致了算法整体时间消耗的增加。算法另外一个缺点在于，检测效果非常依赖于训练数据集的选取与训练参数的调整，令算法的应用变的更加复杂。

## 4.4 本章小结

本章首先介绍了目前较为成熟的使用遗传神经网络进行相似重复记录检测的方式：使用训练数据集对BP神经网络进行训练，然后针对其易陷入局部最小值的缺点，引入遗传算法对其进行改进。然后说明这种方式的一个很明显的缺点就是时间复杂过大，接着引出遗传神经网络应用于IMPN算法中可以互相取长补短：遗传神经网络可以提高IMPN算法的查准率，IMPN算法作为对遗传神经网络算法的预处理，降低了其时间复杂度。最后通过设计对比实验证明了改进的IMPN算法相对于IMPN算法的高查准率的优点。

# 第五章 航天情报系统中的相似重复记录检测

目前数据清洗技术在各行业的信息管理系统中取得了广泛的应用。本章首先介绍了航天情报信息管理系统的需求分析及概要设计与技术实现，然后重点介绍了作者主要负责的数据清理模块，包括数据清理模块的设计、重复记录产生的原因、IMPN算法在系统中的应用以及该算法对数据质量的提高等。

## 5.1 系统需求分析

### 5.1.1 系统建设背景与目标

北京空间科技信息研究所为了提高科技化水平，实现航天情报数据的采集、处理、分析的信息化，于2016年开展“航天情报信息管理系统”项目的研究。该研究以“知识结构化、成果产品化”为目标，立足多年的情报信息数据积累，致力于打造一款功能丰富、实用高效的情报数据信息管理系统。

### 5.1.2 需求分析

“航天情报信息管理系统”主要面向研究所内部研究人员的日常办公使用，经过项目调研与分析后将系统的总体需求概括如表5.1所示：

表5.1需求分析总结表

|  |  |  |
| --- | --- | --- |
| 运行环境需求 | Web | 操作系统为Windows XP，浏览器为Internet Explorer 8 |
| iOS | 操作系统为iOS8.0及其以上 |
| Android | 操作系统为Android4.0及其以上 |
| 功能性需求 | 1. 数据采集模块：将现有的数据采集到系统中并保持和系统中的数据格式一致，包括两种采集模式：人工在线录入以及Excel表格批量导入。 2. 数据清洗模块：对多源数据合并导致的重复数据进行检测清理。 3. 数据检索模块：方便研究人员更加快捷地检索和查询所需信息。 4. 服务支撑模块：包括用户权限管理、数据异常下载行为监视、综合营销平台建设。 5. 移动应用模块：涵盖iOS和Android操作系统在内的手机端App，是系统在手机端的简化体现，方便研究人员随时查看相关信息。 6. 数据应用模块：在已有数据集的基础上，对数据进行统计，并进行可视化展示，方便研究人员更直观地分析数据。 | |
| 非功能性需求 | 1. 性能需求：并发用户数2000，事物平均响应时间3.0s。 2. 稳定性需求：双机热备方案。 3. 安全性需求：网络/系统的安全监测与检查、反爬虫设计等。 | |

## 5.2 系统设计与实现

上一节对系统的需求进行了分析总结，本节主要介绍“航天情报信息管理系统”的设计与实现方式，主要包括系统架构、数据库设计、功能模块实现等。

### 5.2.1 系统概要设计

根据5.1节中的系统功能性需求分析，可以将本系统按照功能模块划分成6个主要的部分，如图5.1所示：



图5.1 六大功能模块示意图

“航天情报信息管理系统”的总体设计从以下三个层面展开：前端和移动端的交互设计、服务器端的逻辑功能、数据持久化的实现。

持久化层主要是负责数据的存储以及向服务器端提供增删改查的服务接口，这里存储了“航天情报信息管理系统”的核心数据信息。

服务器端是处理业务逻辑的核心层，是系统的枢纽部分。数据请求由前端发给服务器端，经用户鉴权通过之后向持久化层请求数据并进行整理发送给前端页面。

前端交互部分主要包括Web网页界面和App移动端页面，这一层是直接和用户交互的最上一层，负责接收用户的指令以及向用户呈现系统信息等。它主要包括用户的注册与登录、数据检索与查询、可视化展示、数据统计等功能页面。

系统的总体架构如下图：



图5.2 航天情报信息管理系统架构图

“航天情报信息管理系统”中的数据设计是系统设计较为核心的一个环节。数据内容主要包括航天器、轨道信息、发射场、运载火箭、航天国家与机构、航天器故障信息等。其中，航天器信息数据是系统的核心数据，按照所属类别又可以将其分成8种：通信卫星、导航卫星、遥感卫星、在轨服务与空间安全卫星、空间科学卫星、技术试验卫星、空间探测器和载人航天器。系统数据库对应的ER（实体-关系）图如图5.3所示：



图5.3实体关系结构图

由上图可知，航天器实体是系统数据库中的最关键实体，它与故障、卫星平台、航天机构、航天国家、航天发射场、运载火箭都存在直接的关系。其中，除了和航天器故障对应关系是“一对多”之外，和其他的几个实体的关系都是“多对一”。航天国家和与其相连的几个实体的关系均为“一对多”。

接下来介绍几个核心数据库表的设计：

（1）航天器表

表5.2 航天器字段设计表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 属性名称 | 属性描述 | 类型 | 可否为空 | 备注 |
| spacecraft\_id | 航天器ID | int(11) | 否 | 主键 |
| spacecraft\_name\_cn | 航天器中文名称 | varchar(60) | 否 | 无 |
| spacecraft\_name\_en | 航天器英文名称 | varchar(40) | 否 | 无 |
| spacecraft\_launch\_num | 航天器发射编号 | varchar(20) | 否 | 无 |
| spacecraft\_num | 航天器编号 | varchar(20) | 否 | 无 |
| task\_property | 任务性质 | varchar(5) | 否 | 无 |
| spacecraft\_type | 航天器类型 | varchar(50) | 否 | 无 |
| country\_id | 所属国家ID | int(11) | 否 | 外键 |
| country | 所属国家 | varchar(20) | 否 | 无 |
| institution\_id | 所属机构ID | int(11) | 否 | 外键 |
| institution\_name | 所属机构 | varchar(200) | 否 | 无 |
| operator\_id | 运营单位ID | int(11) | 否 | 外键 |
| operator\_name | 所属运营单位 | varchar(200) | 否 | 无 |
| … | … | … | … | … |
| spacecraft\_image | 外形图片 | varchar(60) | 是 | 无 |

由于篇幅限制表5.2只给出了部分关键字段的设计。从上表可以看出航天器的ID是标识一条航天器记录的唯一关键字，而该表中的外键：country\_id、institution\_id、operator\_id则是航天器与“所属国家”、“所属机构”以及“所属运营单位”三张表的关联。

（2）航天国家表

表5.3 航天国家字段设计表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 属性名称 | 属性描述 | 类型 | 可否为空 | 备注 |
| country\_id | 国家ID | int(11) | 否 | 主键 |
| country\_name\_cn | 国家中文名称 | varchar(60) | 否 | 无 |
| country\_name\_en | 国家英文名称 | varchar(60) | 否 | 无 |
| budget\_per\_year\_gov | 政府年度航天预算 | float(8,3) | 是 | 无 |
| budget\_per\_year\_civil | 民用年度航天预算 | float(8,3) | 是 | 无 |
| … | … | … | … | … |
| main\_spacecraft | 主要航天器 | text | 是 | 无 |

（3）航天器故障表

表5.4航天器故障字段设计表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 属性名称 | 属性描述 | 类型 | 可否为空 | 备注 |
| malfunction\_id | 故障ID | int(11) | 否 | 主键 |
| malfunction\_spacecraft\_id | 故障航天器ID | int(11) | 否 | 外键 |
| malfunction\_level | 故障等级 | varchar(10) | 否 | 无 |
| malfunction\_date | 故障发生时间 | date | 否 | 无 |
| malfunction\_in\_designlife | 是否发生  在寿命期 | tinyint(1) | 否 | 无 |
| … | … | … | … | … |
| malfunction\_consequence | 故障后果 | text | 是 | 无 |

（4）卫星平台表

表5.4 卫星平台字段设计表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 属性名称 | 属性描述 | 类型 | 可否为空 | 备注 |
| satellite\_platform\_id | 平台ID | int(11) | 否 | 主键 |
| platform\_dev\_org\_id | 平台研制  单位ID | int(11) | 否 | 外键 |
| platform\_dev\_data | 研制时间 | date | 否 | 无 |
| platform\_descrip | 平台描述 | text | 否 | 无 |
| … | … | … | … | … |
| platform\_image | 平台图片 | varchar(60) | 是 | 无 |

### 5.2.2 系统实现

（1）系统开发与运行环境

系统的开发与运行环境如表5.5所示：

表5.5 航天情报信息系统开发与运行环境

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 服务器端 | | Web端 | App | |
| iOS | Android |
| 开发平台 | Windows 7 | | Windows 7 | macOS Yosemite | Windows 7 |
| 开发工具 | Intellij idea  JDK 1.8.0 | | Sublime | Xcode 8 | Android Studio  JDK 1.8.0 |
| 数据库 | MySQL | |  | SQLite | SQLite |
| 运行环境 | 硬件 | 操作系统：  Windows server  2008以上  内存大小：4G  磁盘空间：300G | Internet Explorer8及以上 | iOS 8.0及以上 | Android 4.0  及以上 | |
| 软件 | JDK 7.0.71  Tomcat 7.0.54  MySQL 5.6  Navicat 11 |

（2）技术路线

本系统采用B/S加移动端C/S的综合技术方案进行实现。其中服务端采用较为成熟的SSM（Spring、SpringMVC、MyBatis）三层技术框架，使用Maven添加依赖。数据持久化层由MyBatis实现，它起到了对JDBC的封装作用。服务端的业务逻辑由Spring控制，Spring框架起衔接SpringMVC和MyBatis框架作用。移动端采用经典的MVC技术路线，Model层负责沙盒内数据的封装与维护，View层负责“空间瞭望”App的页面展示与用户交互、包括搜索、统计、航天器信息分类浏览等，ViewController负责处理逻辑业务，如更新航天器列表、收藏航天器、统计信息提取等，并调用Model的接口更新数据库内容。

（3）接口设计

服务端与前端以及移动端的数据传输采用JSON的数据格式，JSON数据更加简便易读易操作，前后端交互采用Http通信协议。核心的接口设计如表5.6所示：

表5.6 数据接口设计表

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 接口名称 | 请求方式 | 接口说明 |
| 1 | getHasLaunchedSpacecraft | GET | 请求已发射航天器列表 |
| 2 | getSpacecraftByCountry | GET | 按照国家分类返回  航天器列表 |
| 3 | getCountBySpacecraftType | GET | 请求某类型所有航天器数量 |
| 4 | getCountByCountry | GET | 请求某国家所有航天器数量 |
| 5 | getSearchResult | GET | 按照关键字返回检索结果 |
| 6 | getMyCollection | GET | 返回当前登录账户  收藏的航天器列表 |
| 7 | getSpacecraftDetailByID | GET | 返回某个航天器的详细信息 |
| 8 | addToCollectionByID | POST | 收藏某颗航天器 |
| … | … | … | … |
|  | login | POST | 登录 |

## 5.3 数据清洗模块

### 5.3.1 “脏数据”产生原因

“航天情报信息管理系统”的数据采集方式有两种，包括人工在线填报数据以及从现存的Excel表格数据批量导入到系统中。人工操作的出错是难以避免的，再加上现存Excel数据来自于不同的子部门由不同的研究人员维护，没有统一的标准，以上即是现存数据集中的“脏数据”产生的主要原因。再加上数据库中的航天器信息多来自于不同的渠道，这就使得数据集中并不存在一个能唯一标识航天器的字段。

表5.7 航天器重复记录举例

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 航天器  名称 | 发射场 | 发射  结果 | 发射时间 | 国家 | 研制单位 | …… | 运载火箭 |
| A | Cape Canaveral | 成功 | 2010/8/14 11:07 | 美国 | 洛克希德-马丁 | …… | 宇宙神-5 |
| A1 | 卡纳维  拉尔角 | 成功 | 2010.08.14  11:07 | US | 洛马 | …… | Atlas-5 |
| B | 卡纳维拉尔  角发射场 | 失败 |  | United.States | Lockhead Martin | …… | 猎鹰-9 |

表5.7展示了三条航天器记录A、A1、B（因真实数据涉及商业机密故数据略有修改）的部分信息，其中A和A1对应着同一颗航天器，B对应另外的一颗航天器。由上表可以看出待处理的数据主要有以下特征：

（1）中英文格式不统一，如“洛克希德-马丁”和“Lockhead Martin”、“宇宙神-5”和“Atlas-5”等。

（2）存在缺失数据，如B的发射时间信息缺失。

（3）中英文缩写与全拼格式不统一，如“United States”和“US”、“洛克希德-马丁”和“洛马”等。

（4）时间格式不统一，如A的“2010/8/14”和A1的“2010.08.14”。

### 5.3.2 重复记录检测算法的应用

数据清洗模块的调用发生在数据录入模块的过程中。用户首先将数据由Web端录入，由浏览器发送到服务器端进行处理。服务器端首先进行预处理，包括时间格式的统一等操作，然后调用重复记录监测模块。该模块的业务流程如图5.4所示：



图5.4 数据清洗业务流程图

由于“航天情报信息管理系统”中总的数据量不超过8000条，所以数据重复检测模块可以采用IMPN算法实现。在两条记录进行相似重复判断时，考虑到两条航天器同一时刻发射的概率非常小，所以在判重过程中可以人为设定“发射时间”占有较高的权重。数据预处理过程主要是按照固定的规则对数据进行检查，如表5.7所示：

表5.7 数据预处理规则表

|  |  |
| --- | --- |
| 规则名称 | 规则描述 |
| 主键判空 | 如果记录的主键为空，则忽略该记录 |
| 日期格式检查 | 标准格式为yyyy-MM-dd HH:mm:ss，将非标准格式的日期统一成标准格式 |
| 日期非法内容检查 | 年份超过当前年份或者当月天数超过31天，对该记录进行标记，由人工检查处理 |
| 国家字段检查 | 国家字段的标准格式为英文全称，将中文内容和英文缩写内容统一成标准格式 |

### 5.3.3 实验设计与结果分析

本节主要目的是通过实验设计说明IMPN算法对“航天情报信息管理系统”数据质量的提升。实验采用3个不同大小的航天器数据集进行测试，数据集均来自于航天情报信息管理系统，如下表所示：

表5.8 航天器测试数据集

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集名称 | dataset1 | dataset2 | dataset3 |
| 数据量（条） | 500 | 1000 | 2000 |
| 重复记录数（条） | 100 | 200 | 400 |
| 字段数目（个） | 24 | 24 | 24 |
| 单条记录最多重复数（个） | 5 | 5 | 5 |

实验采用的IMPN算法的滑动窗口最小值=3，最大值设为=10，算法过程中进行单趟SNM的次数=3。由于系统中的数据量较小（不超过8000条记录），而且没有唯一的ID去区分不同的记录，所以若采用基于遗传神经网络改进的IMPN算法的话，需要人为提取一定量的重复记录进行训练，需要花费大量的时间而且不一定能取得良好的效果，所以这里采用第三章介绍的IMPN算法（未采用遗传神经网络进行改进）。检测结果如下表：

表5.9 IMPN算法检测结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集名称 | 数据量（条） | 查准率（%） | 查全率（%） | 运行时间（s） |
| dataset1 | 500 | 96.08 | 98.00 | 0.118 |
| dataset2 | 1000 | 97.04 | 97.50 | 0.308 |
| dataset3 | 2000 | 97.53. | 98.75 | 0.735 |

从以上结果可以看出，IMPN算法对航天器记录的重复检测取得了良好的效果：IMPN算法的查全率和查准率均达到了95%以上，且运行时间在可接受的范围内。

## 5.4 本章小结

本章首先介绍了“航天情报信息管理系统”的需求分析以及系统设计，然后介绍了一些关键技术的核心实现。接着以数据清洗模块为重点进行展开，介绍了系统中的“脏数据”的来源，数据清洗算法在系统中的应用，并设计了对比实验验证算法对数据质量的提升。

# 第六章 总结与展望

## 6.1 总结

随着信息化技术的发展与应用，各式各样的信息管理系统支撑着企业的数据管理与维护。网络上积累了大量的数据，但这些数据的质量参差不齐。由于数据来源不同、存储于不同的操作系统以及硬件平台、人为错误等导致了包括错误数据、相似重复数据和缺失数据等“脏数据”的产生。其中相似重复记录的检测是数据清洗领域的一个研究重点。本文主要工作内容如下：

（1）研究和介绍了相似重复记录检测领域的相关技术。包括基于单字段和多字段的相似度匹配算法、常用的相似重复记录检测算法以及BP神经网络的理论基础等。

（2）基于现有MPN算法提出了IMPN算法。MPN算法主要有两个缺点：排序关键字的选择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；MPN采用固定大小的滑动窗口，这种方式不够灵活，存在许多无效的比较判重。IMPN主要是从以下三点进行了改进：1.优先选择区分度较大的字段提取排序关键字，从一定程度上减少了人为的参与，减弱算法对于专家知识的依赖；2.采用根据窗口内容可伸缩大小的自适应滑动窗口，减少了部分不必要的判等比较，从而提高了算法的效率；3.对排序关键字为空或者不完整的记录进行标记，初步检测完成之后再对这些标记记录进行归并处理，提高了算法的鲁棒性。

（3）将遗传神经网络应用于IMPN算法的判重过程，提高了算法的查准率。经过实验验证，IMPN算法的优点在于实现了较高的查全率，但却随着数据量的增大查准率逐渐下降，这是算法第三个改进点带来的副作用。为了提高IMPN算法的查准率，第四章将遗传神经网络应用到IMPN算法的重复判断过程中：首先利用训练数据集训练多组三层BP神经网络，然后将其初始化为一个种群，采用遗传算法进行优化，可以克服其收敛慢和易陷入局部最优的缺点，最后将最优的BP神经网络用到IMPN算法的滑动窗口判重过程中，两条记录各字段的相似度值作为输入，输出和阈值进行大小判断以确定它们是否互为相似重复记录。经过对比实验可以验证这一改进提高了IMPN算法的查准率。

（4）最后的章节首先介绍了航天情报信息管理系统的需求分析、概要设计、技术实现，然后重点介绍了系统中的数据清理模块，包括数据清理模块的设计、重复记录产生的原因、IMPN算法在系统中的应用等，最后通过设计实验验证IMPN算法对系统中数据质量的提高。

## 6.2 展望

本文主要研究了数据清洗领域的相似重复记录检测问题，针对研究过程中发现的问题与不足之处，今后仍然需要在以下几个方面进行深入研究：

（1）算法对于中文数据库的处理能力较弱，因为在生成排序关键字以及大小排序时对字符的处理是基于值的，中文的相似重复记录检测一般先进行分词处理，所以在中文重复记录检测方面有待于更广泛和深入地研究。

（2）BP神经网络的隐含层结点个数的确定方式目前尚未有严格唯一的定论。隐层结点数过少会导致网络网络性能表现很差；隐层结点数过多，会增加网络训练时间，且可能使训练陷入局部极小值，导致“过拟合”的出现。

（3）实验部分使用的数据量大小还远未到海量数据的标准，当数据量大小超出内存所能容纳的上限时，提取数据集和排序过程均需要作出调整。可能的解决方案包括将数据集划分为多个更小的单位，采用外部排序等。数据量增大时，算法的效率问题也是亟待考察和解决的。

参考文献

致谢

三年的时间转瞬即逝，这段时间的经历让我成长了许多，也学到了很多东西。

首先我要特别感谢我的研究生导师刘志镜教授。刘老师为人和蔼可亲，科研态度严谨，做事专注，知识渊博，为学生树立了优秀的榜样。在我研究生期间，刘老师耐心指导我进行科研，引领我走进科研的大门，每当我遇到个人无法解决的问题时，老师总是耐心地与我共同分析，通过与老师进行讨论，可以帮助我分析产生问题的原因并想到切实可行的解决问题的建议，这都有助于我及时克服科研瓶颈。研究生生活即将结束，但是刘老师严谨的科研态度和认真的敬业精神一直影响着我，指导我在今后的工作和生活中永葆脚踏实地、认真求实的态度，在此我衷心感谢刘老师这几年来对我的指导与帮助。

其次我要感谢姚勇老师。姚老师在我做毕业设计的过程中给予了专业的指导和帮助，姚老师做人做事的态度令我非常欣赏，是我们学习的榜样。

再者还要感谢实验室李金洋博士、王炳华、朱群、黄辉煌、张沐杰等师兄在我研究生期间给予的指导和帮助；感谢同项目组实验室同学张亚超、严欢以及师弟毛彦淇、崔森、李天霖，大家一起为了同一个目标共同奋斗的日子令我难忘。感谢实验室同学陈仲元、王斌、陈磊、薛登峰以及其它师弟师妹对我学习和生活上的支持与帮助。

感谢我的家人对我的支持和关爱，感谢我的女朋友申星的陪伴和支持，你们是我背后最温柔和强劲的支撑力量。感谢舍友们三年来带来的欢乐和笑声。

最后，感谢所有评审老师，感谢您们抽出宝贵的时间对我的论文进行审阅！

作者简介

##### 基本情况

张攀，男，山东菏泽人，1993年7月出生，西安电子科技大学计算机学院计算机技术专业2015级硕士研究生。

##### 教育背景

2011.09～2015.07 西安电子科技大学，本科，专业：软件工程

2015.09～2018.07 西安电子科技大学，硕士研究生，专业：计算机技术

##### 攻读硕士学位期间参与的科研项目

1. 陕西省慧农科技“农掌门”项目，2015.10~2016.04，负责该项目iOS APP农民端的开发和维护。
2. 北京空间科技信息研究所航天智库平台“航天情报信息管理系统”中后台数据清洗模块的开发以及“空间瞭望”App的iOS端研发与维护。

1. [] Rahm E, Do H H. Data cleaning: Problems and current approaches[J]. IEEE Data Eng. Bull., 2000, 23(4): 3-13. [↑](#endnote-ref-2)
2. [] Raman V, Hellerstein J M. An interactive framework for data cleaning[M]. Computer Science Division, University of California, 2000. [↑](#endnote-ref-3)
3. [] Bitton D, DeWitt D J. Duplicate record elimination in large data files[J]. ACM Transactions on database systems (TODS), 1983, 8(2): 255-265. [↑](#endnote-ref-4)
4. [] Elmagarmid A K, Ipeirotis P G, Verykios V S. Duplicate record detection: A survey[J]. IEEE Transactions on knowledge and data engineering, 2007, 19(1): 1-16. [↑](#endnote-ref-5)
5. [] Jin L, Li C, Mehrotra S. Efficient record linkage in large data sets[C]//Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings. Eighth International Conference on. IEEE, 2003: 137-146. [↑](#endnote-ref-6)
6. [] Wang J R, Madnick S E. The inter-database instance identification problem in integrating autonomous systems[C]//Data Engineering, 1989. Proceedings. Fifth International Conference on. IEEE, 1989: 46-55. [↑](#endnote-ref-7)
7. [] Galhardas H, Florescu D, Shasha D, et al. Extensible framework for data cleaning[C]//2000 IEEE 16th International Conference on Data Engineering (ICDE'00). IEEE, 2000. [↑](#endnote-ref-8)
8. [] 陈伟. 数据清理关键技术及其软件平台的研究与应用[D]. 南京航空航天大学, 2004. [↑](#endnote-ref-9)
9. [] Flanagan T, Safdie E. A practical guide to achieving enterprise data quality[J]. Webpage, Date Accessed, 1998, 12: 01-99. [↑](#endnote-ref-10)
10. [] Monge A, Elkan C. An efficient domain-independent algorithm for detecting approximately duplicate database records[J]. 1997. [↑](#endnote-ref-11)
11. [] Hernández M A, Stolfo S J. The merge/purge problem for large databases[C]//ACM Sigmod Record. ACM, 1995, 24(2): 127-138. [↑](#endnote-ref-12)
12. [] Hernández M A, Stolfo S J. Real-world data is dirty: Data cleansing and the merge/purge problem[J]. Data mining and knowledge discovery, 1998, 2(1): 9-37. [↑](#endnote-ref-13)
13. [] Forgy C L. OPS5 user's manual[R]. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 1981. [↑](#endnote-ref-14)
14. [] Newcombe H B, Kennedy J M, Axford S J, et al. Automatic linkage of vital records[J]. Science, 1959: 954-959. [↑](#endnote-ref-15)
15. [] Qiu Y F, Tian Z, Ji W, et al. An efficient approach for detecting approximately duplicate database records[J]. CHINESE JOURNAL OF COMPUTERS-CHINESE EDITION-, 2001, 24(1): 69-77. [↑](#endnote-ref-16)
16. [] Costa G, Manco G, Ortale R. An incremental clustering scheme for data de-duplication[J]. Data Mining and Knowledge Discovery, 2010, 20(1): 152. [↑](#endnote-ref-17)
17. [] Ferro A, Giugno R, Puglisi P L, et al. An efficient duplicate record detection using q-grams array inverted index[C]//International Conference on Data Warehousing and Knowledge Discovery. Springer, Berlin, Heidelberg, 2010: 309-323. [↑](#endnote-ref-18)
18. [] 郭志懋, 周傲英. 数据质量和数据清洗研究综述[D]. , 2002. [↑](#endnote-ref-19)
19. [] 邱越峰, 田增平, 季文贇, 等. 一种高效的检测相似重复记录的方法[J]. 计算机学报, 2001, 24(1): 69-77. [↑](#endnote-ref-20)
20. [] 陈伟, 王昊, 朱文明. 一种提高相似重复记录检测精度的方法[J]. 计算机应用与软件, 2006, 23(10): 29-30. [↑](#endnote-ref-21)
21. [] Hassanien A E, Azar A T, Snasel V, et al. Big Data in Complex Systems[M]. Springer, Heidelberg, 2015. [↑](#endnote-ref-22)
22. [] 靳丹, 张磊, 王洪军,等. 基于Hadoop的大数据清洗框架设计与应用[J]. 网络新媒体技术, 2015, 4(5):33-38. [↑](#endnote-ref-23)
23. [] Levenshtein V I. Binary codes capable of correcting deletions, insertions, and reversals[C]//Soviet physics doklady. 1966, 10(8): 707-710. [↑](#endnote-ref-24)
24. [] Hassanzadeh O, Miller R J. Creating probabilistic databases from duplicated data[J]. The VLDB Journal—The International Journal on Very Large Data Bases, 2009, 18(5): 1141-1166. [↑](#endnote-ref-25)
25. [] Jaro M A. UNIMATCH, a Record Linkage System: Users Manual[M]. Bureau of the Census, 1980. [↑](#endnote-ref-26)
26. [] Cohen W W. Integration of heterogeneous databases without common domains using queries based on textual similarity[C]//ACM SIGMOD Record. ACM, 1998, 27(2): 201-212. [↑](#endnote-ref-27)
27. [] Hylton J A. Identifying and merging related bibliographic records[J]. 1996. [↑](#endnote-ref-28)
28. [] 韩京宇, 徐立臻, 董逸生. 一种大数据量的相似记录检测方法[J]. 计算机研究与发展, 2005, 42(12): 2206-2212. [↑](#endnote-ref-29)
29. [] Kohonen T. An introduction to neural computing[J]. Neural Networks, 1988, 1(1):3-16. [↑](#endnote-ref-30)
30. [] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. The bulletin of mathematical biophysics, 1943, 5(4): 115-133. [↑](#endnote-ref-31)
31. [] Sutskever I, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning[C]//International conference on machine learning. 2013: 1139-1147. [↑](#endnote-ref-32)
32. [] 李军. 一种相似重复记录检测算法的改进与应用[J]. 成都工业学院学报, 2017 (2017 年 02): 17-20. [↑](#endnote-ref-33)
33. [] Christen P, Pudjijono A. Accurate synthetic generation of realistic personal information[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2009: 507-514. [↑](#endnote-ref-34)
34. [] Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records[C]//Kdd workshop on data cleaning and object consolidation. 2003, 3: 73-78. [↑](#endnote-ref-35)
35. [] 孟祥逢, 鲁汉榕, 郭玲. 基于遗传神经网络的相似重复记录检测方法[J]. 计算机工程与设计, 2010 (7): 1550-1553. [↑](#endnote-ref-36)
36. [] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural networks, 1989, 2(5): 359-366. [↑](#endnote-ref-37)
37. [] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural networks, 1989, 2(5): 359-366. [↑](#endnote-ref-38)
38. [] Zhang Y, Gao X, Katayama S. Weld appearance prediction with BP neural network improved by genetic algorithm during disk laser welding[J]. Journal of Manufacturing Systems, 2015, 34: 53-59. [↑](#endnote-ref-39)
39. [] Asadi E, da Silva M G, Antunes C H, et al. Multi-objective optimization for building retrofit: A model using genetic algorithm and artificial neural network and an application[J]. Energy and Buildings, 2014, 81: 444-456. [↑](#endnote-ref-40)
40. [] Pal S K, Wang P P. Genetic algorithms for pattern recognition[M]. CRC press, 2017. [↑](#endnote-ref-41)
41. [] Sutskever I, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning[C]//International conference on machine learning. 2013: 1139-1147. [↑](#endnote-ref-42)