大雅相似度分析

论文标题： 正文0416v1
检测日期: 2018年04月16日
正文字符数： 46068
正文字数： 34493
检测范围： 大雅全文库

一、总体结论

文献相似度6.41%	重复字符数2951	最密集相似处	密集相似处	非密集相似处	前部相似度	中部相似度	尾部相似度
		1	6	20	8	12	7

二、相似片段分布

三、典型相似文献

相似图书

作者	题名	出处	相似度
中国人工智能学会	中国人工智能进展 2009	北京：北京邮电大学出版社，2009.12	1.14%
中国科学技术协会学会学术部	2005-2006中国科协系统学术会议文献题录 1	北京：中国科学技术出版社，2008.06	1%
姚穆;姜寿山	2006中国国际毛纺织会议暨IWTO羊毛论坛论文集	北京：中国纺织出版社，2007.05	0.93%
中国科学技术协会学会学术部	2005-2006中国科协系统学术会议文献题录 四	北京：中国科学技术出版社，2008.06	0.87%
李国斌;贾宗璞;付子义	2008信息技术与环境系统科学国际学术会议论文集 卷2	北京：电子工业出版社，2008.04	0.83%
曹江华	Red Hat Enterprise Linux 5.0服务器构建与故障排除	北京：电子工业出版社，2008.09	0.35%
《网管员世界》杂志社	《网管员世界》2005超值精华本	北京：电子工业出版社，2006.02	0.25%
《程序员》杂志社	程序员2004合订本 上	北京：电子工业出版社，2005.01	0.23%
孙文江	PHP应用程序开发教程	北京：中国人民大学出版社，2013.05	0.14%
邱林	滦河流域水库群联合调度及三维仿真	北京：中国水利水电出版社，2010.11	0.13%
李兴业	非智力因素与创造力的培养	武汉：湖北教育出版社，2002	0.11%
吕雪峰;彭文波	嵌入式Linux软件开发从入门到精通	北京：清华大学出版社，2014.09	0.1%
张有为	电子数字计算机原理 机器组织与程序设计	北京航空学院，1985.02	0.09%
林万祥	成本会计学	成都：西南财经大学出版社，1994.04	0.09%
李忠	迎接大运的深圳青年 2010深圳青年发展报告	深圳：海天出版社，2011.01	0.08%
弗瑞曼;桑德森	精通ASP.NET MVC 3框架 第3版	北京：人民邮电出版社，2013.05	0.08%
刘云;北京市教育委员会组	数据库及其应用	北京：高等教育出版社，1999	0.08%
李金海	误差理论与测量不确定度评定	北京：中国计量出版社，2003.11	0.08%

原菊梅	复杂系统可靠性Petri网建模及其智能分析方法	北京：国防工业出版社，2011.09	0.08%
洪汉鼎;傅永军	中国诠释学 第7辑	济南：山东人民出版社，2010.06	0.07%
高铁红;曲云霞	控制工程基础	北京：中国计量出版社，2010.07	0.06%
胡可云;田凤占;黄厚宽	数据挖掘理论与应用	北京：北京交通大学出版社，2008.04	0.06%
罗斯·加兰（RossGarland）;汪小金;易洪芳	云大项目管理实用译丛 项目治理	北京：中国电力出版社，2014.05	0.06%
温州市政协七届五次会议秘书处	温州市政协七届五次会议文件 中国人民政治协商会议第七届温州市委员会第五次会议各次大会执行主席名单	政协第七届温州市委员会第五次会议秘书处，2002.03	0.06%
周建中;张勇传;李超顺	水轮发电机组动力学问题及故障诊断原理与方法	武汉：华中科技大学出版社，2013.12	0.06%
黄永斌	区域物流信息平台理论与实证	杭州：浙江大学出版社，2010.09	0.06%
俞涔;叶红玉	企业进出口贸易 流程与实务	杭州：浙江大学出版社，2010.07	0.06%
中国地球物理学会	中国地球物理 2008	北京：中国大地出版社，2008.10	0.05%
中国测绘学会	中国测绘学会第二届综合性学术年会论文选编 第1卷 大地测量	北京：测绘出版社，1982.07	0.05%
黄流兴;牛胜利	蒙特卡罗方法及其应用 第4卷	西安：陕西科学技术出版社，2004.10	0.05%
本书编写组	小学教育学	西安：西北大学出版社，2011.02	0.04%
郭小聪	政府经济学	北京：中国人民大学出版社，2003.11	0.04%
郭小聪	政府经济学 第4版	北京：中国人民大学出版社，2015.02	0.04%
张岳	中国水利发展战略文集 1996-2004	北京：中国水利水电出版社，2004.12	0.04%
国际生物多样性计划中国委员会	中国生物多样性保护与研究进展 9 第九届生物多样性保护与持续利用研讨会论文集	北京：气象出版社，2012.04	0.04%
中国就业培训技术指导中心	秘书国家职业资格培训教程 四级秘书 国家职业资格四级 2014版	北京：中央广播电视大学出版社，2013.12	0.04%
郭小聪	政府经济学 第3版	北京：中国人民大学出版社，2003.11	0.04%
任效乾	环境保护及其法规	北京：冶金工业出版社，2002.05	0.04%
郭小聪	政府经济学 第2版	北京：中国人民大学出版社，2008.06	0.04%
周英	中国水利发展报告 2007	北京：中国水利水电出版社，2007.09	0.04%
任效乾;王荣祥	环境保护及其法规 第2版	北京：冶金工业出版社，2002.05	0.04%
崔炳建	河南国家助学贷款研究	郑州：河南大学出版社，2012.09	0.04%
薛文山;曾北危	环境监测分析手册	太原：山西科学教育出版社，1988.06	0.04%
景晖;崔永红;孙发平	2007-2008年青海经济社会形势分析与预测	西宁：青海人民出版社，2008.01	0.04%
孙乃民	2001年吉林省农村经济形势分析与预测	长春：吉林人民出版社，2000.12	0.04%

宋钰	健康教育与健康促进	沈阳：辽宁大学出版社，2010.03	0.04%
全国知名中学科研联合体实施素质教育的途径与方法课题组	素质教育新教案 语文 初中第4册 第3版 初二下学期用	北京：西苑出版社，2003.12	0.04%
李桢	课堂教学与交流	昆明：云南科学技术出版社，2008.01	0.04%

相似报纸

作者	题名	出处	相似度
	[互联网]4大收费邮箱横向评测	电脑报，2007.09.28	0.1%
	市政府市政协召开第七次联系会议	鄂州日报，2012.08.02	0.04%
	市教育局以人为本为农村青年教师营造温馨家园——“青年教师之家”开创教育工作新局面	滕州日报，2012.06.27	0.04%
	美术馆建设的“展出季效应”	中国文化报，2013.08.25	0.04%
冯琦	六星国际汽车生态文化产业园落哈	黑龙江经济报，2014.01.15	0.04%
朱紫强	增速或超7% 降息降准空间大	东莞日报，2015.07.20	0.04%
	推动文化产业健康发展	柴达木日报，2014.05.19	0.04%
	国家住建部专家组来淮	淮北日报，2010.09.17	0.04%
	中南集团操盘商业地产	浙江市场导报，2008.02.05	0.04%
	我市首家旅游产品展示中心正式投用	德阳日报，2012.09.04	0.04%
	践行社会责任 培育市场人才	北京商报，2015.03.27	0.04%
齐冬梅	依靠群众破解权力监督和制约难题	学习时报，2013.01.14	0.04%
	震荡寻底，寻找上海经济新推力	解放日报，2009.05.02	0.04%
	外商投资企业为西安经济持续发展注入活力	西安日报，2016.05.17	0.04%
	两路街道摊点实名登记助推市场规范管理	重庆晚报，2013.12.20	0.04%
	创业创新双轮驱动开发开展多擎拉动	兴安日报，2016.10.24	0.04%
瞿学忠;赵龙	世界上最长的狂欢节	兰州晚报，2010.10.11	0.04%
	双井富力城店、望京一店本周盛大庆典 大兴一店、天通苑店、上地店同贺 国美会员九周年庆典8月28日盛大开幕...	北京晨报，2009.08.20	0.04%

相似期刊

作者	题名	出处	相似度
李金玉;王曙燕;孙家泽	基于加权D-S的软件易用性评估模型	计算机工程与设计，2016，第1期	1.3%
陈金广;李洁;高新波	基于UT变换的单步滞后无序量测算法	中国科学:信息科学，2011，第5期	1.29%
王富	井下光纤光栅温度压力传感器的研制	西安石油大学学报(自然科学版)，2011，第1期	1.28%
荀素	Smarandache kn数列与除数和函数的混合均值	西安石油大学学报(自然科学版)，2011，第2期	1.26%
朱欣娟;薛惠锋	基于需求分解的知识系统建模方法	计算机应用，2003，第6期	1.26%
张乃禄;付龙飞;任源;孙国鹏;赵歧	基于模糊PID的焊枪伺服控制系统研究	西安石油大学学报(自然科学版)，2011，第2期	1.26%
朱欣娟;薛惠锋	一种概念本体的共享方案	计算机工程与应用，2004，第27期	1.25%
李炀;翟社平	改进的SIFT图像匹配算法	计算机技术与发展，2016，第11期	1.25%
田泽;郭海英	RapidIO传输性能测试分析研究	电脑知识与技术：学术交流，2010，第10期	1.24%
张克旺;潘煜;张琼;张德运	e-MAC:一种面向Ad Hoc网络的高吞吐量MAC协议	软件学报，2010，第10期	1.23%
石小松;程国建	BP神经网络在剩余油分布预测中的应用研究	电脑知识与技术，2008，第36期	1.23%
王学龙;张璟	P2P在制造资源信息共享中的应用	计算机工程与应用，2012，第11期	1.23%

曹庆年;赵博;孟开元	基于ARM9的嵌入式Linux网络通信系统设计与实现	西北大学学报(自然科学版) , 2009 , 第1期	1.22%
林浒;王久远;杨海波;贾正峰	SIP服务器系统	计算机系统应用 , 2016 , 第5期	1.22%
行燕;高荣芳	基于角色访问控制机制在Web信息系统中的应用	现代电子技术 , 2007 , 第16期	1.21%
王魁生;庄杰;李泽辉	基于共享用户界面的笔式交互系统的设计与实现	科学技术与工程 , 2013 , 第20期	1.21%
张留美;潘少伟	微观剩余油仿真研究综述	技术与创新管理 , 2010 , 第6期	1.21%
郭超;刘烨	多色彩空间下的岩石图像识别研究	科学技术与工程 , 2014 , 第18期	1.21%
石争浩;黄士坦;冯亚宁;李长河	遗传算法和BP算法相结合进行图像匹配	武汉大学学报(工学版) , 2003 , 第3期	1.21%
张玢;孟开元;田泽	嵌入式TCP／IP协议栈中ARP的详细分析与实现	电脑知识与技术：学术交流 , 2010 , 第9期	1.2%
刘东	基于BS模式的粮仓检测信息系统	科学技术与工程 , 2013 , 第3期	1.2%
花蕾;温超;仇涵	过滤器技术在企业级Web开发上的研究与应用	电子科技杂志 , 2008 , 第8期	1.2%
谢晓燕;石晓龙	机车综合无线通信设备语音单元的设计与实现	计算机技术与发展 , 2016 , 第2期	1.2%
李维乾;解建仓;李建勋;李莉	突发水污染事件中遥感瓦片大数据存储系统lt;supgt;①lt;supgt;	计算机系统应用 , 2016 , 第2期	1.2%
李瑜;李真;刘建刚	基于BS模式的客流预测系统的设计与实现	现代电子技术 , 2006 , 第14期	1.2%
王家华;全斐;杜延宁	基于面向对象技术的对遗留系统GASOR的重构研究	现代计算机(专业版) , 2007 , 第11期	1.19%
李小康;高荣芳;陈江	基于Vxworks的视频记录仪控制软件设计与实现	电脑知识与技术 , 2009 , 第5期	1.19%
赵红毅;刘利坚	一种分布式系统进程调度方法研究	电子科技杂志 , 2010 , 第6期	1.19%
CHENG;GUOJIAN;Etc.	The Probability Neural Networks for Lithology Identification	微计算机信息 测控自动化 上旬刊 , 2007 , 第6期	1.18%
曹俊侠	单片机串行外围接口电路的三线式结构设计	陕西能源职业技术学院学报 , 2007 , 第2期	1.18%
舒新峰;段振华	有穷时间投影时序逻辑的完备公理系统	软件学报 , 2011 , 第3期	1.18%
李建勋;解建仓;高阳;李维乾	基于复杂agent的水污染运移仿真模拟	计算机应用研究 , 2015 , 第4期	1.18%
朱欣娟;厍向阳;薛惠锋	层次案例规划在知识系统中的应用研究	西安建筑科技大学学报(自然科学版) , 2005 , 第1期	1.18%
薛涛;马腾	基于资源权重最大资源利用率的动态资源调度算法	计算机应用研究 , 2016 , 第5期	1.17%
王倩;刘建华;咎林萍	公众网络身份生态系统研究	计算机技术与发展 , 2015 , 第6期	1.17%
王松伟;石美红;张正;郭仙草	基于熵和变异度的织物疵点图像分割方法	西安工程大学学报 , 2014 , 第2期	1.16%
蔡磊;程国建;潘华贤	基于球向量机的图像分割	计算机工程与应用 , 2011 , 第16期	1.16%
杨凯峰;牟莉;许亮	基于离散小波变换和RBF神经网络的说话人识别	西安理工大学学报 , 2011 , 第3期	1.15%
爨莹;王中	基于模糊多属性决策的射孔方案选择模型研究	计算机工程与应用 , 2009 , 第25期	1.15%
刘天时;肖敏敏;李湘眷	自适应的Haar型LBP纹理特征提取算法研究	计算机工程与科学 , 2015 , 第7期	1.15%
马骋;张玉梅	高速公路入口匝道控制算法综述	计算机仿真 , 2009 , 第1期	1.14%
史晓楠	FDM中支撑设计规则研究	技术与创新管理 , 2009 , 第5期	1.14%
毋涛;张帆	云计算下基于属性的访问控制方法lt;supgt;①lt;supgt;	计算机系统应用 , 2016 , 第2期	1.14%
李俊锋;方明	基于编码的TreeView控件节点生成算法	电脑知识与技术 , 2009 , 第4期	1.14%
兰军芳;黄伯虎	飞邻物联智能系统关键技术在智慧园区中的应用	物联网技术 , 2013 , 第7期	1.13%
	Evaluation of China shale gas from the exploration and development of North America shale gas	西安石油大学学报：自然科学版 , 2011 , 第2期	1.13%
刘天时;赵嵩正	一种分层式2PC协议通信算法研究	计算机工程 , 2004 , 第6期	1.13%
毋涛;何科训	基于SaaS模式的服装生产管理系统研究	计算机技术与发展 , 2016 , 第2期	1.12%
苑庆涛;曹小鹏;王奇峰	NCRE在线报考系统的设计与实现	计算机技术与发展 , 2016 , 第10期	1.11%

其他网络文档

作者	题名	相似度
	2010 年第三期 2010 年第三期 2010 年第三期	1.38%

	二氧化碳增能解堵技术在延长油田的应用	1.24%
	Improved Quadrature Kalman Filter with Large Numbe……	1.21%
	非高斯系统下卡尔曼滤波算法误差性能分析(1)	1.21%
	A RFID Data Cleaning Method Based On Improved M-Ke……	1.12%
	基于离散过程跟踪的自动识别处理中间件	1.11%
	Performance Analysis of the Affine Projection CM A……	1.05%
	A CSCW Architecture Oriented to Textile Industry I……	1.04%
	基于多流多状态动态贝叶斯网络的音视频连续语音识别	1.02%
	APPLICATION OF IMPROVED DIJKSTRA ALGORITHM IN SELE……	1.02%
	图书馆信息咨询简报	1%
	e-MAC:一种面向Ad Hoc网络的高吞吐量MAC协议	0.99%
	Clothing Marketing Based on Consumer Lifestyle	0.98%
严劲	分布式综合入侵防御系统的研究和初步实现	0.97%
	基于ARM 的嵌入式Web 服务器的研究与实现Embedded Web Server Researc	0.97%
	有穷时间投影时序逻辑的完备公理系统	0.96%
	Delay and Energy Analysis in Sparse Mobile Network……	0.95%
	ISBN 9036518342 THE CHEMISTRY OF CHINESE LANGUAGE	0.94%
	游戏算法分析在C语言教学中的应用	0.88%
	一种分布式并行服务器模型的性能分析与改进	0.85%
	多模式匹配算法的性能分析	0.84%
	基于WEBGIS的区域水资源信息系统的设计与实现	0.76%
	ISPRS Workshop on Updating Geo-spatial Databases with Imagery The 5th ISPRS Workshop on DMGISs CHANGE DETECTION BASED ON SPATIAL DATA MINING	0.6%
	A novel heuristic Error-Driven learning for recogn……	0.54%
	ActionScript3.0从零基础学习类	0.32%
	天嵌科技出品-Linux移植之Step By Step_V4.2_20100125	0.27%
卢金伟	土壤团聚体水稳定性及其与土壤可蚀性之间关系研究	0.26%
柳仁川	基于电流变和DSP技术的轨道车辆减振器研究	0.26%
郭阳明	虚拟现实的实时性研究	0.21%
金山	筒式减振器油封的研究与开发	0.2%
邢安	空间高分子材料表面SiO2类薄膜的制备及其原子氧防护研究	0.19%
龚治国	框排架结构设计软件的研制（II）	0.19%
徐淑周	喹诺里西丁类生物碱lasubine关键中间体的合成研究	0.19%
蒋巍	人脸识别方法的研究	0.19%
曹杰	广安须4气藏气井产能评价研究	0.18%
宁力	搜索引擎中网页查重方法的研究	0.17%
宋佳	Effects of Written Output on Foreign Language Vocabulary Acquisition:A Case Study	0.17%

四、全文相似情况

摘要

I 摘要 在现今社会的信息发展过程中，各种来源的数据不断累积，但是原始累积的数据往往含有脏数据，例如错误的、相似重复的和缺失的数据等，对于脏数据进行清洗的一个关键点在于去除数据集中的重复数据。 本文主要对相似重复记录检测的相关算法进行了研究与创新。相似重复记录检测是指准确检测出源数据集中的重复数据，以达到清洗数据的目的。真实情景中，数据规模庞大，数据来源多样，这都增加了重复数据检测的难度。虽然存在解决这类问题的优秀算法

，例如 SNM 算法和 MPN 算法等，但是已有的算法在解决实际应用中的重复记录检测问题时，仍存在不足之处。本文首先研究了传统的多趟近邻排序算法，并对该算法的缺点进行改进，提出了改进的多趟近邻排序算法（OMPN），以适用于实际问题；然后，通过研究基于遗传神经网络（GAANN）求解重复检测问题的算法，将 OMPN 算法与 GAANN 相结合，得到准确度更高的 AOMPN 算法和 BPOMPN 算法；最后，将本文提出的 OMPN 算法应用于实际的“航天情报信息管理系统”的数据清洗模块。本文的主要内容如下：1. 提出了基于 MPN 算法思想的 OMPN 算法。多趟近邻排序算法首先对数据集中的数据记录依据预先选取的排序关键字进行排序，使得相似重复记录排序后位置相近，然后使用滑动窗口对排序后的数据进行判等。但是，该过程需要依赖专家经验知识进行关键字的选取和判等字段的选取，同时，真实的数据可能存在数据缺失，然而，MPN 算法并没有考虑这种情况。MPN 算法所使用的固定大小的滑动窗口不仅会导致对重复数据的检测不全面，而且会导致对非重复数据的冗余检测。本文在多趟近邻排序算法的基础上，提出基于字段区分度的关键字选取方法，根据数据特点进行关键字的选取，同时，在判等过程中，同样根据字段区分度为字段赋予不同权值，避免人为因素；然后，使用所提出的公式计算得到滑动窗口的大小，由该公式得到的大小是不固定的，可以根据数据情况自动调整，减少了漏检记录数量和冗余操作；最后，对源数据中存在缺失值的记录进行标记和单独检测。通过实验验证，本文所提出的改进的多趟近邻排序算法具有较高的查全率，且更适用于真实问题场景。2. 基于神经网络的多趟近邻排序算法。基于遗传神经网络进行相似重复记录检测的算法效果较好，但是该算法不仅训练过程耗时严重，而且在检测过程中存在冗余操作。本文在多趟近邻排序算法与遗传神经网络这两种算法的基础上，使用遗传神经网络对 MPN 算法中滑动窗口内的记录进行判等，将这个算法记作 AOMPN，使得神经网络可以仅对同一个滑动窗口内的记录进行判等，避免了传统的遗传神经网络对数西安电子科技大学硕士学位论文 II 据全集上的任意两个不同记录进行判等，极大地提高了算法运行效率。同时，考虑到遗传神经网络训练速度慢的缺点，本文尝试使用单一的神经网络执行判等操作，得到了基于单一神经网络的多趟近邻排序算法（AOMPN）。实验结果表明，本文所提出的这两种算法准确度和运行效率较高。3. 本文所提出的算法在“航天情报信息管理系统”中的应用。本文主要完成了该系统的数据清洗模块和 APP 模块的开发。在真实业务场景中，航天情报管理系统的数据清洗模块需要实现对源数据的去重和清洗，该系统所使用的数据是真实的不带标签的数据，且数据规模相对较小，综合分析 OMPN 算法、AOMPN 算法与 AOMPN算法的优势与适用场景，最终采用 OMPN 算法实现系统的数据清洗模块。关键词：相似重复检测，数据清洗，多趟近邻排序，神经网络，遗传算法 ABSTRACT III ABSTRACT With the development of the information technology and the information construction, the size of the data becomes larger and larger. Variety of dirty data are inevitable, such as wrong data, reduplicative data and halfbaked data and so on. As a result, effective algorithms are necessary for data cleaning. The duplicate records detection problem is one of the most important problem in data cleaning. In this paper, we have researched and improved the algorithms for the duplicate records detection problem. The duplicate records detection problem is to find the reduplicative records for a given dataset. In real world, it ’ s difficult to design effective

algorithms for the problem since the large size and the different sources of the data.

Although there are some algorithms for solving this problem, such as the

SortedNeighborhood Method (SNM) and the MultiPass SortedNeighborhood Method

(MPN), they all have shortcomings when tackle the realworld duplicate records detection

problems.

The effectiveness of the SNM and the MPN relies on the expert knowledge of the

dataset. So it ’ s hard to solve dataset with no priori knowledge. With the goal of

overcoming the shortcomings of the SNM and the MPN, we proposed the Optimized

MultiPass SortedNeighborhood Method (OMPN). In addition, we make a combination of

the OMPN and the geneticbased artificial neural network to solve the problem and

propose the Advanced MultiPass SortedNeighborhood Method (AOMPN) and the BP

network based MultiPass SortedNeighborhood Method (AOMPN). The AOMPN and

the AOMPN are superior to the other algorithms. Finally, we apply the proposed

algorithm to the spaceflight information management system to accomplish the

datacleaning in the realworld problem. The main contributions of this paper are as

follows:

1. The Optimized MultiPass SortedNeighborhood Method (OMPN) is proposed.

The MPN first sort all the records and then use a scalefixed sliding window to check the

duplicate records. However, it need the expert knowledge to select the key and to check the

duplicate records in a sliding windows. In the OMPN, the field distinction degreebased

method is proposed to select the key without the expert knowledge. In the meantime, the

OMPN uses the scalable sliding window to make the checking process more precise. The

OMPN also take the halfbaked data into account by prelabel scheme. Compared with

other algorithms, the OMPN performs well and it ' s suitable for solving the realworld

duplicate records detection problem. 2. The Advanced MultiPass SortedNeighborhood Method(AOMPN) is proposed.

The geneticbased artificial neural network that used to solve the problem should select

two different records in the whole dataset to check whether they are duplicate or not. It ' s

very timeconsuming and the check stage can be simplified. The AOMPN makes a

combination of the geneticbased artificial neural network and the OMPN to select records

only in a sliding windows. It can not only improve the precision ratio and the recall ratio

but also reduce the runtime compared with the geneticbased artificial neural network.

However, to train an appropriate geneticbased artificial neural network is still

timeconsuming. We also do experiments with the signal BP network and then generate the

BP network based MultiPass SortedNeighborhood Method(AOMPN). Experimental

results show that the AOMPN and the AOMPN all perform well.

3. We apply the proposed algorithm to the spaceflight information management

system. The data cleaning module is one of the most important modules in this system. We

do experiments by the OMPN, the AOMPN and the AOMPN with the given aerospace

craft dataset. Finally, we choose the OMPN to accomplish this module.

Keywords: Duplicate Record Detection, Data Cleaning, SortedNeighborhood, Neural

Network, Genetic Algorithm

第一章 绪论

1

第一章 绪论

1.1 研究的背景和意义

在现今的信息时代，为了在激烈的市场竞争中占据先机，保险、金融等各种行业纷纷加快了信息化的步伐，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[1]。

从规模庞大的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，是数据挖掘研究领域的一个重点方向。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，从而产生各种类型的“脏数据”，例如不可避免的数据重复、缺失、错误等[2][3]。如果这些数据不能被正

确清洗，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，需要高效的对数据进行清洗的算法。

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，只有借助计算机技术，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[4]。

在众多降低数据质量的原因中，最突出的一类原因是数据记录的重复，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[5][6]。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[5]：

（1）资源浪费[5]：重复的记录需要更多的存储资源。

（2）数据的一致性被破坏[5]：数据集中重复的记录之间可能存在互补、自相矛盾和多余等关系。它们共同对应的现实中的实体发生变化会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。

对数据集检测与消除重复记录，既可以节省存储资源，又可以保证数据的一致性，是数据清洗的重要环节[6]。

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。现实中，数据的来源是西安电子科技大学硕士学位论文

2

不确定的、多种多样的，针对这些数据，对它们进行清洗是一种极其复杂的问题，在数据连接[7]、数据实体识别[8]、对象识别等问题上，已经存在较早的研究结果，这些

数据处理的研究对于商业保险、医疗等领域来说，具有非常重大的意义。美国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[9]。数据清洗的研究重点包括：检测重复记录、检测异常数据、有效处理缺失数据。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[9]。因而数据清洗领域的一个重点研究方向即为相似重复记

录检测。

在重复记录清洗方面，国外展开了大量的研究，主要的工作有两个方面——属性匹配[10]和重复检测[11]。

相似重复记录检测领域一种主流的算法是“排序归并”法，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。最基本的算法是“排序合并”算法[11]。这

种方法有很大的不足，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，主要的改进方向包括对字段相似度匹配算法的改进和对相似记录判断方法的改进。

用于属性匹配问题的方法主要有编辑距离算法[12]和递归属性匹配算法[10]等。

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[12]。Hernandez 等

提出了一种有效地多趟近邻排序算法（MultiPass Sorted Neighborhood，记为 MPN）

[13][14]，该算法的操作过程是对 SNM 算法（Sorted Neighborhood Method，记为 SNM）

独立执行多次，MPN 算法中的每一次 SNM 操作都使用不同的关键字对记录进行排序，

在检测过程中使用固定的滑动窗口，最后使用 C 语言重写的 OPS5[15]规则编程判定记

录是否相似。SC Hong 等提出了基于优先级队列的方式进行相似重复记录检测[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，解决了大文本库中的相似检测问题[17]。Alfredo Ferro 使用了基于 qgrams 的相似

度衡量函数[18]，可以避免许多不必要的比较和判断，提高了时间效率。国内的针对重复数据的检测问题的研究主要是改进已有的效果较好的算法，实现对算法精度和效率的提高。复旦大学周傲英等比较早开始数据清理的研究工作[19]。另

外还有基于 NGram 进行检测的算法[20]，该算法以一条数据的 NGram 值作为排序键，

这种基于 NGram 的算法在对因为拼写错误而造成的重复记录进行检测时表现良好。

另外，还有基于权重的检测方法[21]，具体实现过程是，首先按照字段的等级划分进行

权重赋值，然后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，也存在着相关的可以进行数据清洗的软件，这些软件同样需要高效的算法的支持[22][23]。

虽然国内外科研人员已经在重复记录检测问题上探索到了很多有效算法，但仍旧或多或少存在适用局限性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

相似重复记录检测领域的发展虽然已经取得诸多有效成果，但是还有一些不足之处，比如[4]：

（1）检测效率与查全率存在提升空间，特别的，当数据规模非常庞大时，现有的算法尚需改进。

（2）大多数数据清理方法具有适应局限性，只针对特定的业务场景而设计，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。在本文中，首先，针对传统的多趟近邻排序算法 MPN 在时间消耗和检测精度的不足，提出了改进的 OMPN 算法。OMPN 算法有三个改进点：（1）通过统计字段区

分度改善了传统的 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。其次，在 OMPN 算

法的基础上，本文使用反向传播神经网络对 OMPN 算法中的的滑动窗口内的记录进

行判等操作。传统的基于遗传神经网络进行重复记录检测的算法时间复杂度高，该算法的判断操作是针对待处理的数据全集的，而 OMPN 算法将数据全集缩小至可伸缩

的滑动窗口内进行判断。在本文中，将两种算法的优势相结合，提出基遗传神经网络判等的AOMPN算法和基于BP神经网络判等的AOMPN算法，与OMPN算法相比，

这两种算法在查准率性能方面得到很大提升。

1.4 论文结构

论文一共分为六章，每一章的主要内容如下：

第一章是绪论。针对数据清洗的研究展开了介绍，并介绍了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。

第二章介绍已存在的有关算法，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。然后介绍了最基本的近邻排序算法西安电子科技大学硕士学位论文

4

和多趟近邻排序算法，对算法原理、设计以及执行流程和算法的优缺点都进行了介绍。

除此之外还介绍了其它常用算法包括优先级队列算法、NGram 算法等。然后对本文

用到的 BP 神经网络理论基础进行说明。最后介绍了在该领域中衡量算法效果的几个

常用标准及其计算方法。

第三章，首先阐述 OMPN 算法的创新灵感来源，然后详细介绍了 OMPN 的操作

思路，并采用 SNM 算法和 MPN 算法作为对比，通过实验结果，证明了 OMPN 算法

查全率较高的优势，并对算法的优缺点进行了详细分析。

第四章首先介绍了 GAANN 算法的主要思路，通过分析该算法的优缺点，在结

合 OMPN 算法的基础上，提出 AOMPN 算法和 BPOMPN 算法这两种针对 OMPN

算法的改进算法，详细介绍了算法思路，最后给出了详细的实验结果。

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 OMPN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、OMPN 算法在系统中的应用方式以及该算法对数据质量的提高等。

第六章，总结了本文的研究内容并对未来的研究方向进行了分析。

Equation Section (Next)

第二章 重复记录检测相关算法概述

5

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，对于两条记录

1R 、 2R ，它们的内容相同或者相似，且都对

应着同一个现实实体 E ，则记录对 1 2,R R 互相重复

[24]。实际数据库中可能存在多

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 19930102

School of Computer Science, Xi ’ an University of Electronic Science and Technology

1802002 Jack Panda Female 19900720 School of Artificial Intelligence, Xi ’ an University of Electronic

Science and Technology

1801003 S. Water Male 199312

Schol of Computer Science, Xi ’ an University of Electronic Science and Technology

1802004 Jack Panda Female 19900720 School of Artificial Intelligence, Xi ’ an University of Electronic

Science and Technology

1801005 Mr.Sam W Male 19930102 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 ”，是由“ Sam Water ”采用了不同的书写方式而产生的；Gender 字段值

西安电子科技大学硕士学位论文

6

“ 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“ 19930102 ”；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。

有众多原因造成数据重复，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[4][19]。

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[11][13]：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。数据库中的每条记录均由不同的字段组成，如果不同记录的字段取值十分相近，那么它们极有可能是重复的。所以，可以通过计算记录在不同字段的相似性实现检测。目前主要有两类字段相似度匹配的算法，本节分别对其进行阐述。

2.2.1 基于单字段的相似度匹配

基于单字段的相似度匹配算法在相似重复记录检测中的应用过程是，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程度。编辑距离算法[25]、

SmithWaterman 算法[26]、Jaro 算法[27]等都是用于这类问题的有效算法。

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[25]。两个字符串 1S 和 2S 的编辑距离是指： 1S 变成 2S 需要对其单个字

符进行各种操作的操作数，例如插入、替换、删除等操作。编辑距离越小代表 1S 和 2S

越相似[25]。

如图 2.1 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 L 距离的

经典解法是使用动态规划方法。L距离算法在应对字母书写错误、缩写等场景下效果

较好。

SmithWaterman 算法（简记 SW 算法）[26]最早是在生物学序列比对领域被提出

的，是一种用于匹配遗传序列的动态规划算法。它的主要思路是通过罚分和空位计算第二章 重复记录检测相关算法概述

7

不同字段内容的相似度。SW 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。

c h a n g e

c h a l l e n g e

l l e s

插入插入插入删除 图 2.1 编辑距离示意图

Jaro 算法[27]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

距离也是字符串相似度的一种评价方式，对于给定的字符串

1S 和 2S，两者的 Jaro 距

离如下公式(21)所示[27]：

1 2

1

( )

3

Jaro

m m m t

D

S S m

(21)

在公式(21)中， m 和 t 分别是匹配了的字符数和字符的换位数。而另一种

JaroWinkler 相似度匹配算法在 Jaro 算法的基础上，使相同的字符串的分数更高，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。JaroWinkler 距离的计算如公式(22)所示：(1)J W jaro jaroD D L P

D (22)

在公式(22)中， jaroD 是 Jaro 距离， P 可以调整前缀匹配的权值， P 0.25。

2.2.2 基于多字段的相似度匹配

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[28]。

余弦相似度（cos ）[28]是一种基于 TFIDF 加权算法的多字段相似度匹配方法。

算法的步骤如算法 2.1 所示：

算法 2.1：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w (0,1,)i i n ；

2.对每个单词wi 分配权重， w wlog(1)log( )i iweight tf idf ，其中单词出现的次数（词

频）用 witf 表示， wiidf 表示记录总数除以包含wi 的记录个数（逆文档频率）；

3.将待匹配的字段转化成向量a 和b ；

4.计算向量的cos 值： 1

2 2

1 1

( )

cos

( ) ( )

n

i ii

n n

i ii i

A B

A B

；

5. cos 值越接近 1，则说明相似度越高，将cos 与所给阈值相比较，判断记录的相似性。

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[29]。使用基于

单个字段的相似度匹配算法进行相似重复记录检测时，拥有不同的权值的字段对记录相似与否的影响程度也不同，字段之间相似到记录整体的相似度关系是非线性的，这是一种适合使用基于训练样本的有监督学习的场景。因此，存在使用神经网络进行判定的方法，通过带标签的数据集（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，相当于将重复记录检测问题转化为二类分类问题，如果网络输出的结果大于所给阈值，则判定它们重复，否则不重复[29]。

2.3 相似重复记录检测算法

相似重复记录检测领域最直接的方法是对数据集中的数据进行一对一地判等，这种做法简单，查重效果好，但是时间复杂度为 2( )O N，很不高效。“ 排序归并 ” 是目

前相似重复记录检测算法所采用的主要思想，其主要过程是：首先选取排序关键字，关键字可以使预先设定的，也可以按照相应算法进行计算；然后，根据关键字取值，对数据集进行排序，将相似记录汇聚到邻近位置；最后，使用相似度匹配算法进行重复检测。常见的基于“ 排序归并 ” 思想算法有近邻排序算法（SortedNighborhood

Method，记为 SNM）[11]、多趟近邻排序算法（MultiPass SortedNeighborhood，记为

MPN）[13]、优先队列算法（Priority Queue Strategy，记为 PQS）[16]、NGram 算法[20]

等。

2.3.1 近邻排序算法

SNM 算法的设计思路是：首先，根据数据领域的专家知识经验指定数据集排序所使用的关键字的生成方式；其次，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，设第 i 条记录 iR 的排序关键字为 ikey ；然后，按照记录按照各自的的 ikey 完

成排序，根据相似记录对应的关键字内容也相似的原理，不同的重复记录在排序完成后理论上会处于邻近的位置；最后在每一个固定大小的滑动窗口内，判断与当前窗口内的数据集重复与否。SNM 算法步骤如算法 2.2 所示：

算法 2.2：SNM 算法

1.确认排序关键字的生成方案，滑动窗口大小为w；

2.对每条记录

iR 生成排序关键字 ikey ；

3.按照

ikey 对数据集的记录进行排序（只考虑内部排序）；

4.对同一个滑动窗口内的记录进行判断。

SNM 算法的滑窗如图 2.3 所示：

N

N

当前窗口的记录下一个窗口的记录.....

.....

图 2.2 SNM 算法滑动窗口过程示意图

若数据集大小为 N，使用 SNM 算法生成排序关键字过程的时间复杂度为 ( )O N，

本文使用快速排序，其时间复杂度为 (log)O N N，对滑动窗口内记录的判等操作的时

间复杂度为 ( )O wN，其中w为窗口的固定大小。SNM 算法十分简单，滑动窗口判重

过程效率较高，运行速度较快。但它也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，不相似的记录却处于邻近位置，这就导致算法的检测效果大打折扣。

（2）很难选择合适的滑窗大小。若w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w太小，则检测不全，导致算法查全率下降。 2.3.2 多趟近邻排序算法

MPN 算法改进了 SNM 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，多趟近邻排序的关键字不同，西安电子科技大学硕士学位论文

10

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。

传递闭包的定义[30]是：设 R 是一种关系，定义在集合A 上， P 代表传递性，满

足下列所有条件的关系

1R 称为R 的传递闭包：

（1）

1R R ；

（2）

1R 满足性质P ；

（3）如果存在集合A 上的关系 'R， 'R 满足性质P 并且 'R R，则 '1R R 。

多采用 Warshall 算法[31]计算传递闭包，算法 2.3 给出了 Warshall 算法的伪代码：

算法 2.3：Warshall 算法

1.W := MR

2.FOR k:= 1 to n

FOR i := 1 to n

FOR j := 1 to n

[,] [,] ([,] [,])W i j W i j W i k W k j

3.Output W ；

MPN 算法对多趟检测结果可以计算传递闭包的理论基础是相等的传递性[14]：若

记录 1R 和 2R 重复，记录 2R 和 3R 重复，则 1R 和 3R 也重复。

MPN 算法的步骤如算法 2.4 所示：

算法 2.4：MPN 算法

1.确认m 个排序关键字的生成方案，独立重复地执行步骤 2~4 m 次； 2.对数据集生成排序关键字；

3.按照 (1, ,)ikey i m 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 (1, ,)iw i m，每次比较时将 新进入窗口的记录 wR 与窗口内剩余

的 1w 条记录进行相似性判断得到重复记录集合 (1, ,)iD i m ；

5.对m 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 算法的流程图如图 2.3 所示： 第二章 重复记录检测相关算法概述

11

生成关键字方案1

生成每条记录

的排序关键字

按照记录的关键字

进行快速排序

滑动窗口

进行重复检测

检测到重复记录？

滑动窗口

过程结束？

加入到重复记录集1中

是是否生成关键字方案2

生成每条记录

的排序关键字

按照记录的关键字

进行快速排序

滑动窗口

进行重复检测

检测到重复记录？

滑动窗口

过程结束？

加入到重复记录集2中

是是否合并检测结果并

计算传递闭包

读取输入数据

开始

结束

第1趟SNM得到的

重复记录合集1

最终检测结果

第2趟SNM得到的

重复记录合集2

图 2.3 MPN 算法流程图

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 MPN 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，缩短了滑动丬并的执

行时间。但是 MPN 算法也存在缺点：依旧没有克服 SNM 算法的缺点，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

NGram 算法是一种基于聚类思想的算法[32]。NGram 值由记录中每个单词出现

的概率综合计算得出，重复记录的 NGram 值相似，算法的优势在于对常见拼写错误

西安电子科技大学硕士学位论文

12

表现较好，例如基于 NGram 层次空间的 DGHS 算法效果较好[33]。

记录

1R 、 2R 的 NGram 相似性计算方法如公式(23)所示[20]：

1 2

1 2

1 2

( ) ( )

(,)

( ) ( )

q q

q

q q

G R G R

sim R R

G R G R

(23)

公式(23)中, 1 2(,)qsim R R 表示记录 1R 、 2R 的 NGram 值， ( )qG R 表示记录R 的所

有字段组成的集合。

另一种优先级队列算法的思想是：用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，若队列中不含有当前记录则赋予其

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。采用多趟优先级队列算法进行相似重复记录检测的过程如图 2.4 所示[16]：

输入数据集

排序 排序数据集1 队列扫描

Key1 匹配模型

结果1

排序 排序数据集2 队列扫描

Key2 匹配模型

结果2

第一趟

PQS

第二趟

PQS

...

...

...

...

最终的重复

记录聚类

...

图 2.4 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[34]，整体表达能力强，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，

然后加上阈值(偏置)，最后经过非线性函数的处理，得到输出结果[35]。图 2.5 展示了

一个神经元[34]：

第二章 重复记录检测相关算法概述

13

1a

2a

na

1w

2w

nw

SUM f t

1

b

.

.

.

.

图 2.5 神经元模型

其中， ia 是输入向量的第 i 个分量， iw 是第 i 个输入分量连接到该神经元的权值，

b 表示偏置， SUM 是加权求和操作， f 表示激活函数。则对于输入向量1 2(, ,...,)ninput a a a 经过此神经元时，得到：

1

( )

n

i i

i

output f w a b

。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[36]是一种经典的优化方法，其主要思想是不断沿着负梯度方向进行搜

索。给定如公式(24)所示的目标函数， 0 1, ..., d 是要学习的参数，

j

ix 是第 j 个输

入特征向量的第 i 个分量，

0

Ijx ，

0 表示偏置，共有 d 维特征。

1 20 1 2

0

( ) ...

d i

d

j j j j j

d i

i

h x x x x x

(24)

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。也就是说，在当前

7

内，即将离开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w窗口内的重复记录在窗口内的位置依次为 0、、…w，其中w

是刚滑入的数据，0 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

据对于的尺寸影响越大。本文提出了动态计算大小的计算公式，如公式()所示：

m max m

dex

ext idex

w

i index

w w w w

i index

(3)

公式()中，常数

maxw 表示滑动窗口的大小可以取的最大值，常数 minw 表示滑动窗口的最小尺寸，w表示当前滑动窗口的大小， index 表示即将滑出滑动窗口的记录

在数据集中的索引位置，

iB 代表数据集中索引为 i的记录是否与 index 位置的记录互为重复，若它们重复，则 iB =，否则 iB =0。通过分析公式(3)得到，若w内的记录都

是重复的，则滑动窗口大小更新为最大值 maxw，相反，若w内的记录互不重复，则滑

动窗口大小更新为最小值

minw ；并且距离 index 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 index位置的记录互为重复记录时）。

R0 R0_0 R0_ RR0_ R0_3 R0_4 R2 R3..... .....

2

3

4

5

6

3w

5w

4w

图 3.3 自适应大小的滑动窗口

图 3.3 展示了自适应大小的滑动窗口检测过程，在图 3.3 中，R0、R0_0、R0_、

R0_2、R0_3、R0_4 互为相似重复记录，记录 R0、R、R2、R3 互不重复。滑动窗口

内的归并过程从左向右进行。设定最大窗口为 5，最小窗口为 3，①表示初始时，滑

动窗口大小为 4；由公式(32)计算得到第②步所采用的滑动窗口大小变为 5，因为此

时窗口内的记录和 R0_2 全部重复，所以第②步的窗口大小被扩大；然后由公式(32)

计算得到第③步窗口大小取最小值 3，因为此时滑动窗口内的数据和 R1 互不重复，

所以窗口大小被缩小；同理得到第④⑤⑥的滑窗大小。图 3.3 简明生动地表现了自适

应滑动窗口的变化过程，展现了 O 算法所使用的自适应大小的滑动窗口检测方

法的优势。

3.3 基于预标记处理排序关键字不完整的方法

3.3.1 MPN 排序方法的缺陷

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，表 3.5 给出了一种数据缺失情况的示例。

表 3.5 缺失数据及不完整数据示例

Recrd irst Name Last Name ddress

1R Jack tolo 123 irst Street 1278

2R Jack 123 First Street 1233 3R Jack Stolpho 123 First Street 12345678

4R Jacon Stiles 123 Forest Street 1234

在表 3.5 中，记录 R1 的 Last Name 字段原本应该是“ Stolo ”，出现字段不完整

而变成“ Sto ”， 在生成排序关键字时若采用 3.1.1中的第 3种方式，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，本来重复记

录的 R1 和 R2 由于排序关键字的首字母的差异而无法聚集在近邻的位置，使得被检

测为互为重复记录的概率减小。

传统的 MPN 算法在处理这种带有不完整数据和缺失数据的数据集时，就会遇到这种问题，从而使检测精度降低。为了克服这一缺点，本文提出了针对排序关键字不完整的改进方法，详细介绍在 3.3.2 节。

3.3.2 改进的基于预标记的方法

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[45]：

29

（1）使用缺失数据的一些统计计算进行填充；

（2）根据业务和经验选择合适的值进行填充；

（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（2）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（3）

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，因此，这几种处理缺失值的方法不能较好的适用于相似重复记录检测问题。

针对相似重复数据检测问题，考虑到不完整数据和缺失数据会造成记录的排序关键字缺失或不完整，进而会对记录排序后的位置产生影响，所以本文针对缺失数据使用“ 标记处理法 ”，该方法的主要操作过程

（1）对所有关键字不完整的记录的 进行标记；

（2）从数据全集中去除第（1）步所标记的数据，只对排序关键字完整的记录进行排序和归并；

（3）处理被标记的带有缺失值的记录，分别对这些记录进行检测，将其一一聚类到第（2）步得到的重复数据簇中。

本文这种基于预标记处理缺失值做法能弥补 MPN 算法在排序关键字缺失的情况检测效果差的缺点，同时，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，时间耗费在合理的范围内，在真实数据集中，缺失值往往只占有较小的比例，因此该方法是可行的。

3.4 OMPN 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，OMPN 算法的改进思想在于以下三点：

（1）基于字段区分度选取排序关键字，避免了对专家经验的依赖性。

（2）采用可伸缩的滑动窗口检测方法，根据数据特点动态调整检测窗的大小，减少不必要的比较次数。

（3）预标记含有不完整排序关键字的记录，更适用于真实应用场景。

有N 趟 SNM 过程的 OMPN 算法步骤如算法 3.3 所示： 算法 3.3：OMPN 算法

1.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；

30

3.优先选取区分度较大的字段去生成 N 组排序关键字 { , ,... }nkey key key ；

4.独立地执行步骤 5~8 N 次；

5.按照排序关键字的产生方式对每提取其排序关键字

ikey ；

6.对数据集按照关键字

ikey 排序，如果某条记录的 key 不完整或者为空则将该记录的 加入到缺

失关键字记录集合 _ _ _ iincompleteset wi h keyt 中，完整则正常排序；

7.进行可伸缩大小的滑动窗口得到重复集合 _ idup set ；

8.将 _ idup set 与 _ _ _ iincompleteset wi h keyt 进行重复归并，然后计算此集合的传递闭包

itransitive_closure_set ；

9.将 N 次 SNM 重复检测得到的 itransitive_closure_set 集合进行归并，然后得到

最终的 _ _total dup set 。

含有两趟 SNM 过程的 OMPN 算法流程如图 3.4 所示：

31

生成键值方案1

滑动窗口

加入到重复记录集1中

是否是是否合并标记及

重复记

生成键值方案2

生成每条记录

的排序键值

检测

滑动窗口

过程？

加入到重复记录集2中

是否是是否合并标记及

重复记

合并检测结果并

计算传递闭包

记录集合1

录集合1

重复记录1

第1趟SNM得到的

重复记录合集1

最终检测结果

第2趟SNM得到的

重复记录合集2

合并后的

重复记录2

检测到的重复

记录集合2

标记的记

录集合2

图 3.4 含有 2 趟 SNM 过程的 OMPN 算法流程图

32

3.4.2 时间复杂度分析

OMPN 算法是在 MPN 算法的基础上进行的改进与创新，因此主要对这两种算法

的时间复杂度进行分析。同时，理想状态下所有的数据都可以在内存中处理，不考虑磁盘 IO 的情况。

设N 是记录总个数，是滑动窗口尺寸，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 ( )O N ；排序过程采用快速排序，

算法时间复杂度为 (log)O N N ；滑动窗口的归并检测过程需要进行 N 次比较，所以

产生 ( )O N 的时间复杂度；在传递闭包的计算过程中，假设重复记录数据集的大小

为d，则该阶段的时间复杂度为 3( )O d 。所以对于 MPN 算法来说，总的时间复杂度

为：

3( ) (log)+ ( )+ ( )mpnT O N O N N O N O d ()

由公式()可以看出，OMPN 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 k，则区分度统计阶段的时间复杂度为 ( )O kN ；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 ( )O N 、 (log)O N N 、 ( )O N ；设因为排序关键字为空而被标注的数据集

包含记录数为m，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

( )O mN ；传递闭包过程中的时间复杂度同理为 3( )O d，d 为检测后

的大小。所以 OMPN 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m 满足m ＞k 并且m ＞）：

3( ) (log)+ ( )+ ( )+ ( )IMPNT O kN O N N O N O mN O d ()

观察公式(33)和公式(34)组成部分可以发现，两者复杂度在同一数量级，所以两

者的时间复杂度在特定的数据集上是一致的。若用a 代表常数，两个公式都可以简化

成下面的式子：

3( ) (log)+ ( )T O aN O N N O d ()

当重复记录较多时，即d 的值较大，此时 3( ) ( ) (log)O d O aN O N N ，MPN 和

OMPN 的时间复杂度都变为 3( )O d，这体现出了 MPN 算法和 OMPN 算法的时间消耗

受重复数据比例的影响均较大。

33

3.5 SNM、MPN、OMPN 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“ febrl ”[46]

（开源地址：https:sourceforge.netproectsfebrl）生成的。“ febrl ”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 字段描述 举例 1 举例 2

rec_id 记录 ID rec4org rec454dup0

culture 文化 pak pak

sex 性别 f f

age 年龄 30 30

date_of_birth 出生日期 19822l 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 421674 08 42167414

soc_sec_id 社保 ID 20942

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、重复比例、字段特征、错误类、重复记录的概率分布等，由该生成器得到的数据集非常接近现实数据；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。“ 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 参数说明

34

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 算法的评价指标

针对相似重复记录检测问题，常用的评价指标有查全率（）、查准率（）

以及 Fmeasure[13][14]。

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（）：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，如公式(36)所示：

recall

N

()

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，如公式(37)所示： P

precision

TP

TP F

(39)

Fmeasure 的计算公式如(38)：

35

2

2

(1)

F

precision recall

(310)

其中参数α取值为 1 时，Fmeasure 即为最常见的 F1measure：

1

2 precision recall

F

reecision recall

(311)

假设有 8 条记录

1R、2R、3R、4R、5R、6R、7R、8R，其中{1R，2R，3R}和

{4R，5R，6R}互为相似重复记录，若通过算法检测出的重复结果为{1R，2R，3R，

7R}和{4R，5R，6R，8R}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)=.67，

F1measure=210.6667(1+0.6667)=80.00。

3.5.3 实验结果分析

本文使用了多个数据集进行算法仿真，对 OMPN 算法、MPN 算法和 SNM 算法

的查全率和查准率进行综合对比。这三种算法的实验参数为：

OMPN 算法的实验参数：最大滑动窗口最大为

max =20，最小为 min =3，进行单

趟 SNM 的次数 t =3；

MPN 算法的实验参数：固定滑动窗口大小为 =5，进行单趟 SNM 的次数 t =3；

SNM 算法的实验参数：固定滑动窗口大小为 =5。

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。

数据集的规模与生成数据的参数取值如表 3.8 所示：

表 3.8 测试数据集参数取值说明

数据集 1p 2p 3p 4p 5p 6p 7p

dataset1 00 1000 3 1 1 uniform phonetic

dataset2 10000 2000 3 1 1 uniform phonetic

dataset3 20000 00 3 1 1 uniform phonetic

dataset4 50000 10000 3 1 1 uniform phonetic

dataset5 80000 100 3 1 1 uniform phonetic

dataset6 100000 20000 3 1 1 uniform phonetic

36

dataset7 200000 40000 3 1 1 uniform phonetic

dataset8 500000 100000 3 1 1 uniform phonetic

表 3.8 中的参数

1p ~ 7p 代表的含义分别为： 1p 原始记录数、 2p 重复记录数、 3p

单个记录最多重复数、

4p 单个字段最多修改数、 5p 单个记录最多修改字段数、 6p重复记录的概率分布、

7p 错误类型。

本实验的实验环境配置如表 3.9 所示：

表 3.9 实验环境配置

操作系统 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）

在表 3.8 所示的 8 种不同规模的据集上，分别使用 SNM、MPN 和 OMPN 进行实

验，统计每种算法在每种数据集上的 recall 、 precision和运行时间（：秒），实

验结果分别如表 3.10 至表 3.12 所示。

表 3.10 SNM、MPN、OMPN 算法查全率对比表

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

表 3.11 SNM、MPN、OMPN 算法查准率对比表

5 100.00 99.89 99.50

10 99.94 100.00 99.54

37

20 99.97 100.00 98.78

50 99.87 99.98 96.93

80 99.81 99.92 95.

100 99.92 99.87 94.77

200 99.73 99.89 89.70

500 99.53 99.70 88.17

表 3.12 SNM、MPN、OMPN 算法运行时间对比表

数据集大小（千条） SNM 算法（s） MPN 算法（s） OMPN 算法（s）

5 0.503 0.938 1.274

10 1.196 2.320 2.9

20 3.540 6.985 10.327

50 17.169 26.3 .420

80 47.2 239.434 197.439

100 67.173 372.245 290.903

200 237.805 1668.995 1391.415

500 1493.305 104.894 11296.066 根据表 3.10 至表 3.12 的实验结果，分别作出查准率、查全率和运行时间的折线

图，如图 3.5 至图 3.7 所示。

图 3.5 SNM、MPN、OMPN 算法查全率折线图

38

图 3.6 SNM、MPN、OMPN 算法查准率折线图

图 3.7 SNM、MPN、OMPN 算法运行时间折线图

根据表 3.10 至表 3.12 计算得到三种算法的每种数据集上的 F1 值，如表 3.13 所

示，并在图 3.8 中作出 F1 值的折线图，。

表 3.13 SNM、MPN、OMPN 算法 F1measure 值对比表

数据集大小（千条） SNM 算法（） MPN 算法（） OMPN 算法（）

5 89.258 97.278 99.149

10 87.712 95.725 98.967

第三章 改进的 OMPN 算法

39

20 87.6 94.336 98.7

50 87.597 93.2 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

图 3.8 SNM、MPN、OMPN 算法 F1measure 折线图 由图 3.5 可以看出：改进的 OMPN 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 OMPN 算法查全率的数值稳定在 96以上。所以证明了改进的 OMPN 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 OMPN 算法的查全率较为稳定。

由图 3.6 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；OMPN 算法采用基于字段区分度的加权判等方式，在检测大小从

5000 到 200000 的数据集时，查准率均在 90以上，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 OMPN 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。 由图 3.7 可以看出：SNM 算法时间消耗最低，因为 MPN 和 OMPN 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，OMPN 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 OMPN 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。

综合以上结果可以看出，OMPN 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。但是由图3.8 可以看出，综合考虑查全率和查准率，OMPN 算法的整体表现要好于 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OMPN，并从三个方面介绍了传统 MPN 的处理方法的不足以及 OMPN 算法的对其缺

点的改进。之后介绍了 OMPN 算法的设计与流程步骤，最后通过采用数据生成器 febrl

生成的不同规模的数据集进行对比实验，实验验证了 OMPN 算法的查全率和 F1 值都

较高，比 SNM 和 MPN 这两个算法更优，但是也反映了 OMPN 算法查准率不够理想

的问题，因此，本文继续对 OMPN 算法进行改进，详细介绍在第四章。 Equation Section (Next)

第三章 改进的 OMPN 算法

41

43

第四章 基于遗传神经网络改进的 OMPN算法

OMPN 算法采用“ 排序归并 ”的思想，在归并过程中，需要采用有效的方式对滑动窗口内的记录进行判等，OMPN 采用基于字段区分度的加权判等方式，避免了

人工参与，但是这种方式表现不稳定，当数据量的增加时，OMPN 的查准率下降。

传统的 MPN 算法主要根据人工经验选择有效的字段对不同的记录进行判断，因此，

受到人为干扰。为了避免这个弊端，同时提高 OMPN 算法的查准率，本章提出了基

于遗传算法和神经网络相结合的、有监督地学习判等方法，并与 OMPN 的算法思想

相结合，提出增强的MPN算法（Advanced Optimized MultiPass Sorted Neighborhood，

AOMPN）。AOMPN 算法的查全率和查准率都优于现有算法，但是 GAANN 算法的

训练耗时较大，所以，我们在 AOMPN 的基础上进行简化，只训练一个 BP 网络，

用这个训练好的网络进行判等操作，由此得到基于 BP 网络的 MPN 算法（BPbased

Optimized MultiPass Sorted Neighborhood，BPOMPN）。

本章首先介绍了遗传神经网络在相似重复记录检测问题上的操作过程，然后分别给出 AOMPN 算法和 BPOMPN 算法的具体操作过程，并通过实验验证了本文提出

的 AOMPN 算法和 BPOMPN 算法的性能。

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

BP 网络使用 BP（Back Propagation）算法进行训练。 1x

2x

3x

4x

1

1h

2h

3h

4h

5h

6h

o

图 4.1 进行相似重复记录检测的 BP 网络结构

BP 神经网络一般由 3 层或者 3 层以上神经元组成。一方面，由引理[47]可以看出

44

三层网络的较好的逼近性能。另一方面，网络层数的增加虽然可以增强表达能力，但是也会带来较高的训练复杂度。因此，对于重复记录检测，GAANN 采用三层 BP 网

络，图 4.1 表示了用于相似重复记录检测的 BP 网络结构，网络各层所代表的含义及

其计算方法分别为：

（1）输入层的结点数由数据集中记录的字段数目确定，输入层的取值是两条记录的相似度向量。若数据集中的m 个字段，则输入层结点数为m 。任取两

条记录

xR 和 yR，计算这两个记录在第 i 个字段下的相似度值为 isim， 1,2,...i m，则

这xR 和 yR 的相似度值组成的向量为 1 2(, ,...,)msim sim sim，BP 网络的输入层各结点的

取值也随之确定，即第 i 个结点的输入为 isim 。例如，对于表 3.6 所示的字段说明数

据，每条记录都有 16 个字段，其中，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，不属于数据集的特征，所以这个数据集共有 15 个字段，因此，输

入层应当有 15 个结点（不含一个偏置结点）。

（2）输出层的结果代表 BP 网络对于两条记录相似度的计算。结点个数为 1，输

出的结果介于 0 到 1 之间，结果越接近 1 表示两条记录相似程度越高。

（3）隐含层连接输入层与输出层，主要进行 BP 网络的前向计算和误差反向传播的过程。结点数目由公式[48]计算得到。

2

1

2

1 2

0.43 0.12 2.54 0.77 0.35 0.51

0.618( ),

max(,)

M NK K N K

N N K N K

M

N K N N K

M M M

(41)

其中M 是隐含层结点数， N 是输入层结点数，K是输出层结点数，计算过程的结果需要四舍五入进行修整。例如，使用表 3.6 所示的字段说明表进行 BP 网络结构

的设计时，输出数值个数为1，所以K =1；15个字段对应着输入神经元的个数为N =15，

通过公式(41)计算得到 1M =7， 2M =24，最终得到适用于该数据集的 BP 网络的隐层

神经元个数为M =24。

本文使用的激活函数为 Sigmoid 函数[39,40,41]，如公式(29)所示。

神经网络的初始权值对网络最终的收敛情况具有重要的作用，同时，初始权值也会影响训练速度的快慢 [49]。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 sigmoid 激活函数的导数取最大值的地方进行下一轮权值的调

节与变化。

权值更新的学习速率决定了每次训练后权值更新的幅度，学习速率较大会使得训练过程震荡，过小的学习速率又会使得权值在每一次更新过程中的改变量较小，

45

进而减慢训练过程，耗时较大[50]。通常，可以采用小学习速率，以保证系统稳定性而

不会导致修改幅度过大，本文取值为 0.15。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。

BP 神经网络训练终止的条件是训练误差达到预先设定的阈值或权值更新次数达到预先设定的迭代次数，本文在仿真实现过程中，同时采用这两种终止条件，只要满足其中一条，则训练终止。

针对相似重复检测问题，将已知记录是否重复的数据作为训练集，用于训练神经网络，得到训练完成的网络后，用该网络对测试数据进行相似重复检测。使用 BP 网

络进行相似重复记录检测的算法操作过程如图 4.2 所示：

训练好的

固定权值和阈

值的

对测试数据集进

重复检测

随机初始化

BP网络

否是 46 图 4.2 BP 网络检测重复记录的流程图 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。但训练好的 BP 神经网络可能达到局部最优状态。GA算法[51]模拟生物自然进化的操作，它利用自然界的进化思想，通过选择、交叉、变异等操作实现种群的进化过程，并在演化过程中淘汰适应性较差的个体，经过数代进化之后筛选出适应度较高的个体，得到全局的最优解[51]。GA 的全局搜索能力较强，所以可以将其应用到 BP 神经网络中，以解决神经网络难以跳出局部最优的问题[52][53]。 GAANN 算法框架的元素主要有： 1.染色体 采用实数编码[54]的编码方式，每一个实数对应着一个 BP 网络的权值或阈值，并作为一个独立的染色体，设种群规模为P，则共有P 组神经网络权值和阈值，即需要训练P 个 BP 神经网络。 以图 4.1 所示的网络结构为例，首先训练好一个 BP 神经网络，将训练得到的权值以及阈值作为一个个体，输入层有 4 个神经元，1 个偏置单元，隐含层与 6 个神经元，所以输入层到隐含层之间的权值共有(4+1)6=30 个，输出层有 1 个神经元，所以隐含层到输出层之间共有 6+1=7 个权值，因此，该神经网络共有 37 个权值，1 个输出层阈值，当采用实数编码方式时，染色体的长度为 38。 2.适应度函数 在重复检测问题中，适应度函数用于衡量相似重复检测结果和真实结果的之间的差距，差距越小，说明这个染色体所代表的的BP神经网络对数据的检测结果越准确，所以适应度值越大。本文使用公式(42)[55]计算第 i 个个体的适应度值。 11iifE (42) 公式(42)中， iE 代表使用第 i 个个体对应的 BP 神经网络进行重复记录检测时，在记录数据集上所得到的总误差， iE 的值是用公式(43)进行计算。 2

1

( )

N

i j j

j

E g a

(43)

第四章 改进的遗传神经网络算法

47

公式(43)中，N 代表训练数据集的总大小，

jg 代表第 j 条记录的期望输出结果，

ja 代表第 j 条记录的实际输出结果。

3.选择

采用轮盘赌[56]方法进行选择操作，使得适应度函数值较大的个体被选中的概率较

大。个体被选中的概率用公式计算：

1

i

i M

i

i

f

P

f

(44)

公式(44)中，M 是种群中个体总个数， if 是第 i 个个体的适应度值。

4.交叉、变异[44]

因为一个网络权值向量构成一个染色体，所以在该问题上，采用实数编码，因此对任意不同的两条个体上的对应位置执行数值交换，即完成一次交叉操作，本文采用两点交叉操作。

变异算子模仿的是基因突变的过程：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。

使用遗传神经网络进行相似重复记录检测的执行过程图如图 4.3 所示。其主要操

作为：首先，进行有监督的训练学习得到P 组 BP 网络；然后，用这P 组 BP 网络的

权值产生初始状态下的种群，种群中含有P 条染色体，种群经过遗传操作不断得到优

化；最后，从进化后的种群中获取最优的 BP 网络，使用这组最优的网络权值所

代表的网络对待检测的数据集进行重复性检测。

GAANN 算法操作过程中，需要对数据集中的任意两个不同的记录进行比较，判断他们是否相似或重复，带来较高的时间复杂度，为 2( )O N，存在多余的判断过程。

所以需要对其进行改进，在保留遗传神经网络所具有的优势的同时，减小算法的时间复杂度。

48

读取数据

提取训练数据集

和测试数据

求隐含层、输出

层个单元输出

求目标值和实际

值的偏量e

e在允许的

范围内？

训练结束，得到

训练好的BP网络

计算网络中的神

经元误差

求误差梯度

更新权值

随机初始化N组

BP网络

结束

否是提取N组权值与阈值初始化的种群选择算子轮盘赌操作交叉算子变异算子进化代数达到上限？计算个体的适应度产生新一代种群产生最优的BP网络对测试数据集进行重复记录检测否是 图 4.3 遗传神经网络检测相似重复记录流程图 4.2 基于神经网络的 AOMPN 算法和 BPOMPN 算法 AOMPN 算法和 BPOMPN 算法的操作过程都包括两部分：（1）根据 OMPN 算法进行记录排序，包括基于字段区分度提取排序关键字、依据关键字对所有记录进行排序、选取自适应大小的滑动窗口；（2）使用某种判等方法，对同一个滑动窗口内的数据执行判等操作，最终得到重复记录。与 GAANN 算法相比，AOMPN 算法和BPOMPN 算法首先通过 OMPN 算法对数据集进行预处理，以避免大量多余的判等操作，因此时间复杂度降低。AOMPN 算法使用 GAANN 算法的思想进行判等，BPOMPN 算法仅使用 BP 神经网络进行判等。 4.2.1 基于遗传神经网络的 AOMPN 算法 传统 SNM 算法和 MPN 算法使用的是基于专家经验知识的规则产生式系统（OPS5），这种判等方法主要依赖人工操作以选择进行比较的字段。第三章提出了基于字段区分度的加权判等方式，这种判等方法对小数据量的数据表现较好，当数据规第四章 改进的遗传神经网络算法 49 模不断增大时，算法的查准率逐渐变差，性能不稳定。传统的基于遗传神经网络进行相似重度检测的算法不存在排序操作，所以需要在数据全集对数据进行两两判断，时间复杂度为 2( )O N 。综合 GAANN 算法和 OMPN 算法，本文进一步提出了 AOMPN 算法，该算法主要使用训练完成的神经网络判断滑动窗口内的记录是否重复，训练过程采GAANN的训练思路。AOMPN 算法实现了根据数据集的特点训练专门的用于判等的网络，减少了人工干预，同时可以提高算法对多种数据集的适应性。 算法对于给定的训练集，共训练

sizem 个 BP 神经网络，然后用训练得到的 sizem 个

BP 神经网络的权值作为初始的 sizem 个个体，再进行相关的遗传算法方面的操作，最后

从sizem 个 BP 神经网络中选择最优的结果作为最终训练得到的神经网络，然后用这个

神经网络进行判等操作，AOMPN 算法的操作过程如算法 4.1 所示： 算法 4.1：AOMPN 算法

1.读取待检测的数据集以及训练数据集，设定执行 SNM 的次数为 SNMt，最大遗传进化次数 maxg ；

2.随机产生 sizem 个初始 BP 神经网络；

3.分别对

sizem 个 BP 网络进行训练，用训练完成的 sizem 个 BP 网络的权值向量产生初始种群；

4.种群进行遗传操作直到达到 maxg，得到最优的个体，即对应着最优网络 bestBP ；

5.根据字段区分度选取排序关键字；

6.对每一条待检测记录：

7.对数据全集中的、非 1D 中的其余记录进行排序；

8.采用 bestBP 进行可伸缩大小的滑动窗口重复检测，得到重复集合 2D ；

9.在集合 2D 中检测 1D 中的记录的重复记录，得到最终的重复记录集合D；

10.独立地执行步骤 5~9 SNMt 次，得到 SNMt 个重复记录集合，并对 SNMt 个重复记录集合进行传递

闭

使用 AOMPN 算法进行相似重复记录检测的流程图如图 4.4 所示。

50

使用BPbest检测

记录含有

滑动窗口

重复记录集合

结束

检测到的重复记录集合标记的记录集合最终检测结果获取训练数据初始化m个BP神经网络对m个神经网络进行训练由m个训练好的BP网络初始化种群进行遗传算子操作（选择、交叉、变异）最优BP网络BPbest生成键值方案生成每条记录的排序键值 图 4.4 AOMPN 算法流程图 与 OMPN 算法的基于字段区分度的加权判等操作相比，使用遗传神经网络进行判等可以有效地提高算法的查准率。同时，与传统的使用遗传神经网络求解该问题的算法相比，OMPN 算法对数据集进行了预处理，只需要比较滑动窗口范围内的记录，减少了不必要的判断计算，减少了时间消耗。 4.2.2 基于 BP 网络的 BPOMPN 算法 AOMPN 算法的需要较长的时间进行网络的训练，因此对该算法进行简化，提出基于 BP 神经网络的 BPOMPN 算法。设 In 表示 BP 神经网络的输入层节点数，隐含层节点个数为 Hn，则一次训练过程（包括前向计算输出和误差反向更新过程）的时间复杂度为 ( )I HO n n，设 BP 网络终止训练过程的最大次数为 BPT，遗传算法的种群第四章 改进的遗传神经网络算法 51 数量为sizem，则需要训练 sizem 个 BP 网络以产生初始种群，所以种群初始化过程需要( )size I HO m n n 的时间复杂度，种群演化过程的时间复杂度为2( )sizeO m，所以 AOMPN算法训练过程的总时间复杂度如公式(48)： _ 2( ),( ), size I H I H sizeA IMPN trainsize I H sizeO m n n n n n mTO m n n m (45) 可以看出算法训练过程的时间复杂度为平方级别。除此之外，BP 神经网络优化算法属于高维优化问题，算法训练的结果往往是收敛于某个鞍点附近，而不是局部最小值[]。在鞍点和局部极小值的梯度都等于零，大量鞍点的存在才是神经网络优化困难的真正原因，而基于 GA 对 BP 网络进行改进的算法只有在问题拥有较多局部极值

的时候效果较好，所以适用较为局限[58]。因此，本节提出了简化的 BPOMPN 算法，

仅训练单一的 BP 神经网络，代替 AOMPN 算法中遗传伸进网络的训练。BPOMPN

步骤如算法 4.2 所示：

算法 4.2：BPOMPN 算法

1.读取待检测的数据集以及训练数据集，设定执行 SNM 的次数为 SNMt ；

2.使用训练数据集训练出一个 BP 神经网络；

3.根据字段区分度选取排序关键字；

4.对每一条记录：

如果该记录的关键字存在缺失：

将该记录的 ID 加入到缺失关键字记录集合 1D 中；

5.对数据全集中的、非 1D 中的其余记录进行排序；

6.采用 bestBP 进行自适应尺寸的滑动窗口重复检测，得到重复集合 2D ；

7.在集合 2D 中检测 1D 中的记录的重复记录，得到最终的重复记录集合D；

8.独立地执行步骤 5~9 SNMt 次，得到 SNMt 个重复记录集合，并对 SNMt 个重复记录集合进行传递闭

包计算，得到最终重复记录集合 FD ；

图 4.5 是 BPOMPN 算法的流程图。

52

计算字段区分度

按照记录的键值

进行快速排序

标记记录，放

到缺失键值集

合中

使用可伸缩的滑动窗

口进行重复检测

用BPtrained检测

到重复记录？

记录含有

不完整键值？

滑动窗口

过程结束？

加入到重复记录集中

是否是是否合并标记记录集合及

重复记录集合

合并多趟SNM过程的检测结果

并计算传递闭包

读取输入数据

开始

结束

检测到的重复记录集合标记的记录集合最终检测结果获取训练数据初始化BP神经网络对神经网络进行训练训练好的BP网络BP-trained生成键值方案生成每条

记录的排序键值 图 4.5 BP-OMPN 算法流程图 4.3 OMPN、A-OMPN、BP-OMPN 综合对比实验 本节对OMPN算法、A-OMPN算法和BP-OMPN算法在不同数据集上进行实验，并统计实验结果。实验环境设置和数据集设定与第三章 3.5 节相同，均采用 8 个规模不同但生成方式相同（生成器“ febrl ”的参数一致）的数据集，对于神经网络的训练，训练过程采用的数据集大小为 500，即含有 500 条记录，使用“ febrl ”生成，生成参数与实验数据集一致，详情见表 3.7 与表 3.8。第四章改进的遗传神经网络算法 53 对于 OMPN 算法，最大滑动窗口大小max =20，最小滑动窗口大小 min =3，算法过程中进行单趟 SNM 的次数SNMt =3。对于 BP 神经网络的训练，学习速率设置为 0.05，动量系数设置为 0.9[59]，网络最大训练次数 =5000BPT ；GA 的种群初始大小 =50sizem，变异概率 =0.05mutap，种群最多迭代 =300GAT 次。OMPN 算法、A-OMPN 算法、BP-OMPN 算法的查全率、查准率、运行时间（单位：秒）的实验结果分别如表 4.1 至表 4.3 所示，对于 A-OMPN 算法和 BP-OMPN 算法，这里统计的运行时间是使用训练好的网络进行检测的时间，不包括训练网络的时间。表 4.1 OMPN、A-OMPN 和 BP-OMPN 算法查全率对比表 数据集大小（千条） OMPN 算法（%） A-OMPN 算法（%） BP-OMPN 算法（%） 5 98.80 98.90 98.80 10 98.40 98.35 98.15 20 98.75 98.90 98.90 50 97.90 97.85 97.65

80 98.58 98.53 98.39

100 98.73 98.76 98.78

200 99.10 99.15 99.50

500 96.91 97.23 97.43

表 4.2 OMPN、AOMPN 和 BPOMPN 算法查准率对比表 数据集大小（千条） OMPN 算法（） AOMPN 算法（） BPOMPN 算法（）

5 99.50 100.00 100.00

10 99.54 100.00 100.00

20 98.78 100.00 99.98

50 96.93 99.87 98.95

80 95.69 99.85 98.80

100 94.77 99.88 98.81

200 89.70 99.68 97.81

500 88.17 99.21 97.69

表 4.3 OMPN、AOMPN 和 BPOMPN 算法运行时间对比表

54

数据集大小（千条） OMPN 算法（s） AOMPN 算法（s） BPOMPN 算法（s）

5 1.274 21.110 20.800

10 2.961 83.169 77.989

20 10.327 336.404 328.788

50 59.420 2172.766 1928.253

80 197.439 5510.201 5287.167

100 290.903 8505.129 8119.437

200 1391.415 32842.004 30901.765 500 11296.066 204281.991 193198.802 根据表 4.1 至表 4.3 所示的查全率、查准率和运行时间（单位：秒），分别作出

OMPN 算法、AOMPN 算法和 BPOMPN 算法的查全率折线图、查准率折线图和运

行时间折线图，如图 4.6 至 4.8 所示。

图 4.6 OMPN、AOMPN、BPOMPN 算法查全率折线图 第四章 改进的遗传神经网络算法

55

图 4.7 OMPN、AOMPN、BPOMPN 算法查准率折线图 图 4.8 OMPN、AOMPN、BPOMPN 算法运行时间折线图 由图 4.6 所示的查全率对比结果可以看出，AOMPN 算法、BPOMPN 算法与

OMPN 算法相比，查全率非常接近，且均在 96以上，证明基于 OMPN 改进的

AOMPN 算法和 BPOMPN 算法能够保证较好的查全率，这与理论上的结果一致，

因为基于神经网络改进的算法和 OMPN 算法的主要区别在于判等过程采用了不同的

方法，而这一区别几乎不影响查全率的指标。

由图 4.7 所示的查准率对比结果可以看出，AOMPN 算法和 BPOMPN 算法相对

于 OMPN 算法，查准率均有较大的提升；由图 4.8 所示的运行时间对比结果可以看

56

出，AOMPN 算法和 BPOMPN 算法的检测时间相近，且数倍于 OMPN 算法。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，AOMPN 算法和 BPOMPN 算法采用 BP 神经网络进行判等，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 OMPN 算法采用基于权重的判等过

程时间复杂度高，OMPN 算法具有常数级的 (1)O 的时间复杂度，而改进的 AOMPN

算法和 BPOMPN 算法，对于

In 个输入结点、 Hn 个隐含结点的三层网络，检测过程

的时间复杂度为 ( )I HO n n，因此，AOMPN 算法和 BPOMPN 算法的检测时间较长。

接下来对 AOMPN 算法和 BPOMPN 算法在多个不同数据集上进行训练，并统

计训练时间，分析这两种算法的训练效率。

实验采用 5 个规模不同但生成方式相同（生成器“ febrl ”的参数一致）的数据集，

训练数据集的规模与生成数据的参数取值如表 4.4 所示。数据集的大小分别为 100、

200、500、1000、2000，重复记录所占的比例均为 20。BP 网络训练时的参数为：

0.05 ，动量系数取 0.9[60]，最大训练次数为 5000BPT ；GA 种群初始大小 50sizem ，

变异概率 0.05mutaP ，种群最多迭代 =300GAT 次。 AOMPN 算法和 BPOMPN 算法训练时间的结果如表 4.4 所示：

表 4.4 AOMPN 和 BPOMPN 算法训练时间对比表 数据集大小（条） AOMPN 算法（s） BPOMPN 算法（s） 100 742.260 38.665

200 2934.840 156.076

500 18513.945 968.742

1000 59564.372 3970.958

2000 335082.875 16484.161

由表格 4.4 画出 AOMPN 算法和 BPOMPN 算法训练时间的结果对比折线图如

图 4.9 所示：

图 4.9 AOMPN、BPOMPN 算法训练时间折线图 由图 4.9 可知，BPOMPN 算法训练过程的时间消耗较低，且相对于 AOMPN 算

法优势明显。这主要是由于 AOMPN 算法基于种群进化对训练好的多个 BP 网络进

行优化，而 BPOMPN 算法只进行单一网络的训练，且没有遗传操作等寻优过程，因

第四章 改进的遗传神经网络算法

57

此 AOMPN 算法的训练时间较长。

综上，基于遗传神经网络的 AOMPN 算法和基于 BP 神经网络的 BPOMPN 算

法在保证了 OMPN 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

以牺牲检测时间为代价。另一方面，AOMPN 算法的训练过程的时间消耗过大。简

化的 BPOMPN 算法缩减了训练过程的时间。通过实验结果可以看出，AOMPN 算

法的效果与 BPOMPN 算法的效果差距较小，因此在实际问题中，可以综合考虑精度

要求和时间要求，选择最适用于实际情况的算法。

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：使用训练数据集训练出合适的BP网络，然后针对其可能陷入局部最小值的缺点，

引入遗传算法对其进行改进。然后将遗传神经网络与 OMPN 算法相结合提出了AOMPN 算法，AOMPN 算法在提高查准率的同时缩减了遗传神经网络进行重复记

录检测的复杂度，但是 AOMPN 算法训练网络的过程耗时严重，所以针对 AOMPN

算法，提出了简化的 BPOMPN 算法，缩减了训练过程的时间消耗。最后通过对比实

验证明了AOMPN算法和BPOMPN算法都可以得到较高的查全率和查准率，最后，

给出了这两个算法的训练时间，可以看出 BPOMPN 算法训练较快。

59

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、概要设计以及技术实现，然后重点介绍了数据清理模块，包括数据清理模块的设计、重复记录产生的原因、OMPN 算法在系统中

的应用以及该算法对数据质量的提高。

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

需求

Web 操作系统为 Windows XP，浏览器为 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；从 xls 文件导入。（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） 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：开发 iOS 手机端和 Android 手机端，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。

60

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

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：

航天情报信息系统

数据录入模块 数据清洗模块 数据检索模块 服务支撑模块 移动应用模块 数据应用模块

数据批量导入在线录入数据预处理相似重复记录检测数据查询数据检索用户管理下载服务监视综合营销平台数据统计数据搜索数据统计可视化展示 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：

（1）和移动端的页面设计；

（2）服务逻辑功能设计；

（3）数据化的实现。

持久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。

61

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。图 5.2 展示了

“ 航天情报信息管理系统 ” 的总体架构。

注册登录界面、检索界面、

统计与可视化界面、用户管

理界面等

注册登录界面、搜索界面、

统计界面、综合营销界面等

Web App

Shiro鉴权

业务接口API

前端

交互

服务

器端

持久

化层

MyBatis

MySQL数据库

数据缓存

MVC

发送请求

返回数据或

JSP

Spring

会话管理

数据标准化、

数据清洗等预

处理操作

CRUD操

作查询结果

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、轨道信息、、、航天国家与机构、故障信息等。其中，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：

西安电子科技大学硕士学位论文

62

航天器

航天器

航天器

交付周期

航天器故障

卫星平台

航天机构 航天国家

航天场

运载火箭

故障ID

航天器ID

故障等级

故障后果

平台ID

平台

单位

平台图片

机构ID

机构 研制单位ID 国家ID 国家名称

ID

场

名称

简介

运载火箭ID

火箭名称

所属系列

分类

发生 1n

研制

1

属于

n

1

属于

n

1

发射

n

1

1运载n

属于n 1 属于

n

1

n

图 5.3 实体关系结构图

由图 5.3 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 ”。

表 5.2 至表 5.5 是较为核心数据库字段设计表：

（1）航天器表

表 5.2 航天器字段设计表

spacecraft_id 航天器 ID int(11) 否 主键 spacecraft_name_cn

航天器中

spacecraft_name_en

航天器英

文名称

varchar(40) 否 无

spacecraft_launch_num

航天器发

射编号

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

country 所属国家 varchar(20) 否 无

institution_id

所属机构

ID

institution_name 所属机构

operator_id

运营单位

ID

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

country_id 国家 ID int(11) 否 主键

country_name_cn

国家中文

名称

country_name_en

国家英文

名称

varchar(60) 否 无

budget_per_year_gov

政府年度

西安电子科技大学硕士学位论文

64

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

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 5.5 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 第五章 航天情报系统中的相似重复记录检测

65

5.2.2 系统实现

（1）系统开发与运行环境

表 5.6 给出了本文所搭建的航天情报信息系统的开发环境与运行环境。 表 5.6 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

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

iOS 8.0

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 BS 加移动端 CS 的综合技术方案进行实现。其中服务端采用较为成熟的 SSM 三层技术框架实现，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，而 SpringMVC 和 MyBatis 框架被 Spring 框架衔接起来。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、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

西安电子科技大学硕士学位论文

66

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 5.7 所示：

表 5.7 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，除此之

外现存 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

2010814

11:07

美国

洛克希德

马丁

…… 宇宙神-5

A1 卡纳维 成功 2010.08.14 US 洛马 …… Atlas-5

第五章 航天情报系统中的相似重复记录检测

67

拉尔角 11:07

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“ Atlas5 ”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“ 2010814 ”和 A1 的“ 2010.08.14 ”。

5.3.2 重复记录检测算法的应用

对于“ 航天情报信息管理系统 ”所读取的原始数据，需要对其进行清洗，以去除重复记录，该系统中的数据清洗模块即为实现这一目标而设计。用户首先从网页端将原始数据录入系统；然后，浏览器向服务端发送数据；服务器端收到原始数据后，对其进行时间格式的统一等简单操作；最后，服务器端调用数据清洗模块，实现对系统数据的去重。本文所实现的数据清洗模块的业务流程如图 5.4 所示： 开始

结束

用户Web端录入数据IMPN算法进行重复检测待入库数据数据预处理去除重复记录的数据集归并到数据库中用户在线录入数据推送到审核人员进行审核审核通过？是否合并删除重复记录 图 5.4 数据清洗业务流程图 在判重的时候，考虑到两条航天器同一时刻发射的概率非常小，所以在判重过程西安电子科技大学硕士学位论文 68 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，该系统所使用的预处理规则如表 5.8 所示：表 5.8 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 日期统一用 yyyyMMdd HH:mm:ss 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 本文共提出了 OMPN 算法、AOMPN 算法和 AOMPN 算法，它们都在重复检测问题上具有较好的性能。在第三章与第四章的实验分析部分，分别给出了这三种算法在数据生成器产生的带标签数据集上的实验结果，通过算法的查全率和查准率验证了算法的性能。但是，数据清洗模块实际采用 OMPN 算法而不是 AOMPN 算法或者AOMPN 算法，其原因是：（1）“ 航天情报信息管理系统 ”中总的数据不超过 8000条，数据量比较小，OMPN 算法可以满足系统对于查全率和查准率的需求。（2）数据库中的航天器信息源自于不同的渠道，并不存在一个能唯一标识航天器的字段。所以若采用基于遗传神经网络的AOMPN算法或者基于BP神经网络的AOMPN算法，需要提取出训练数据集对网络进行迭代训练，而没有唯一标识数据记录的字段导致提取带有标签的训练集只能由人工完成，工作量大效率较低。 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，并基于本文提出的 OMPN 算法设计数据清洗模块，在真实的航天器重复数据检测中取得的良好效果。

69

第六章 总结与展望

6.1 总结

信息时代的发展带来了剧增的数据量，由于来源的多样性、存储数据的方式和硬件设备不同、人为错误难以避免等，导致我们获得的数据存在缺失或存在相似重复数据，因此需要研究出高效的算法，实现对脏数据的清洗操作。在数据清洗方面，相似重复记录检测是一个研究热点，本文在研究了传统解法的基础上对其进行改进和创新，

并使用本文所提出的算法实现了实际项目的数据清洗模块，达到了系统的设计要求。本文的主要工作主要有：

（1）提出了改进的多趟近邻排序算法 OMPN。在 OMPN 算法中，首先，提出了

基于字段区分度的关键字选取方法和基于字段区分度加权的判等方法，这种方法可以避免传统的 MPN 需要依赖专家经验知识的弊端，避免了人工干扰，可以根据数据特

点提取有效地关键字。其次，提出了自适应大小的滑动窗口处理方法，在传统的 MPN

算法中，使用的滑窗大小是固定的，带来灵活度不够、存在漏检和冗余检测的问题，而在 OMPN 中，通过当前滑窗内记录的重复程度计算得到新的滑窗大小，避免了相

似度低的记录的重复检测，也增加了对相似度高的数据的检测次数，更加灵活，效果更好。最后，传统的 MPN 算法对存在缺失值的数据效果较差，因此，OMPN 算法对

缺失数据进行预标记，避免了数据缺失带来的不准确性。实验结果表明 OMPN 算法

相对于 MPN 算法具有一定的优势。

（2）结合遗传神经网络和 OMPN 算法，提出增强的多趟近邻排序算法 AOMPN

和基于 BP 神经网络的多趟近邻排序算法 BPOMPN。遗传神经网络进行检测时，准

确度较高，但是需要对任意两个不同的记录的相似度向量进行检测，所以，检测过程繁杂，存在冗余操作。OMPN 算法提出了基于字段区分度的判等方法，但是该方法

对数据量敏感，在大规模数据上，算法的查准率下降。因此，本文综合遗传神经网络和 OMPN 算法，对于 OMPN 算法的滑动窗口内的记录，使用遗传神经网络进行判等，

既避免了遗传神经网络的冗余操作，又提高了判等操作的准确度，最终得到查准率和查全率都较高的 AOMPN 算法。另外，针对遗传神经网络训练速度慢的缺点，本文

提出使用单一 BP 神经网络执行判等操作，得到 BPOMPN 算法。通过进行实验，验

证了 AOMPN 算法和 BPOMPN 算法的查准率、查全率均较高，且 BPOMPN 算法

的训练耗时明显小于 AOMPN 算法。在实际问题中，可以综合考虑时间和精度要求，

选择合适的算法。

（3）在真实的航天情报信息管理系统中，运用本文提出的 OMPN 算法进行数据

清洗模块的搭建。该系统所提供的真实数据量为 8000 条数据，且真实数据没有标签，

所以使用 OMPN 算法进行构建。

6.2 展望

本文主要研究了重复记录检测问题，针对研究过程中发现的问题与不足之处，得到以下两点未来工作方向：

（1）对含有中文的记录，算法性能较差，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ASCII值的，而中文的检测方法需要分词操作，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）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。