大雅相似度分析

论文标题： 正文0415v1
检测日期: 2018年04月15日
正文字符数： 46754
正文字数： 35238
检测范围： 大雅全文库

一、总体结论

文献相似度7.15%	重复字符数3343	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	7	20	9	14	5

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	1.12%
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	0.98%
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	0.92%
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	0.86%
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	0.82%
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	0.34%
《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%
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	0.14%
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%
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	0.09%
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%
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%
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%
郭勇	酶工程原理与技术	北京：高等教育出版社，2005.09	0.08%
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	0.07%
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%
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	0.07%
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%
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%
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%

蒲晓蓉;陆庆	计算机互联网络技术教程	成都：电子科技大学出版社，2001.08	0.06%
杨森;杨守印	送电线路电气及机械计算	工人出版社，1986.04	0.06%
张世和;徐继延	数据结构习题解析与实训	北京：清华大学出版社，2008.08	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%
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%
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%
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%
骆淑波;彭景	烹饪营养与卫生	沈阳：东北财经大学出版社，2008.08	0.04%
刘立柱;王刚;丁志鸿	信息理论与编译码技术	北京：国防工业出版社，2013.08	0.04%
赵山林;高媛	C语言程序设计	北京：人民邮电出版社，2012.10	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%
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%
张燕红	计算机控制技术 第2版	南京：东南大学出版社，2014.02	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%
郝春生	人文素质教程	北京：清华大学出版社；北京交通大学出版社，2004.06	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%
李继承	医学细胞生物学	杭州：浙江大学出版社，2005.08	0.04%
杨颖秀	教育法学	北京：中国人民大学出版社，2014.09	0.04%
陈艳	人文素质教程	北京：北京交通大学出版社，2008.02	0.04%
李文军;钟家龙	财贸类专业教材 中国经济地理	北京：中国环境科学出版社，2002.09	0.04%
顾伟驷	现代电工学	北京：科学出版社，2005.06	0.04%
张顺清;李金山	中华人民共和国文化史	哈尔滨：黑龙江教育出版社，1992.06	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%
	震荡寻底，寻找上海经济新推力	解放日报，2009.05.02	0.04%
	美术馆建设的“展出季效应”	中国文化报，2013.08.25	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%
	中南集团操盘商业地产	浙江市场导报，2008.02.05	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%
	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%
朱紫强	增速或超7% 降息降准空间大	东莞日报，2015.07.20	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%
瞿学忠;赵龙	世界上最长的狂欢节	兰州晚报，2010.10.11	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%
	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%
	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%
	推动文化产业健康发展	柴达木日报，2014.05.19	0.04%
	市政府市政协召开第七次联系会议	鄂州日报，2012.08.02	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%

	国家住建部专家组来淮	淮北日报 , 2010.09.17	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	1.29%
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	1.27%
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	1.26%
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	1.24%
苟素	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	1.24%
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	1.24%
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	1.23%
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	1.23%
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	1.22%
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	1.21%
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	1.21%
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	1.21%
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	1.2%
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	1.2%
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	1.2%
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	1.19%
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	1.19%
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	1.19%
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	1.19%
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	1.18%
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	1.18%
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	1.18%
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	1.18%
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	1.18%
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	1.17%
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	1.17%
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	1.17%
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	1.17%
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	1.16%
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	1.16%
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	1.16%
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	1.16%
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	1.15%
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	1.15%
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	1.15%

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	1.14%
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	1.14%
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	1.13%
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	1.13%
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	1.13%
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	1.12%
史晓楠	FDM中支撑设计规则研究	技术与创新管理 , 2009 , 第5期	1.12%
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	1.12%
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	1.11%
	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	1.11%
程国建;彭中亚;王莹	基于独立成分分析和核向量机的虹膜识别方法	计算机工程与设计 , 2010 , 第5期	1.11%
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	1.1%
陈金广;李洁;高新波	双重迭代变分贝叶斯自适应卡尔曼滤波算法	电子科技大学学报 , 2012 , 第3期	1.1%
毋涛;李原	基于离散过程跟踪的自动识别处理中间件	计算机工程 , 2011 , 第19期	1.1%

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	1.36%
	二氧化碳增能解堵技术在延长油田的应用	1.22%
	Improved Quadrature Kalman Filter with Large Numbe……	1.19%
	A RFID Data Cleaning Method Based On Improved M-Ke……	1.1%
	基于离散过程跟踪的自动识别处理中间件	1.09%
	Performance Analysis of the Affine Projection CM A……	1.03%
	A CSCW Architecture Oriented to Textile Industry I……	1.02%
	基于多流多状态动态贝叶斯网络的音视频连续语音识别	1.01%
	APPLICATION OF IMPROVED DIJKSTRA ALGORITHM IN SELE……	1%
	图书馆信息咨询简报	0.99%
	e-MAC:一种面向Ad Hoc网络的高吞吐量MAC协议	0.98%
严劲	分布式综合入侵防御系统的研究和初步实现	0.96%
	基于ARM 的嵌入式Web 服务器的研究与实现Embedded Web Server Researc	0.96%
	有穷时间投影时序逻辑的完备公理系统	0.95%
	Delay and Energy Analysis in Sparse Mobile Network……	0.94%
	ISBN 9036518342 THE CHEMISTRY OF CHINESE LANGUAGE	0.93%
	游戏算法分析在C语言教学中的应用	0.87%
	一种分布式并行服务器模型的性能分析与改进	0.84%
	多模式匹配算法的性能分析	0.83%
	基于WEBGIS的区域水资源信息系统的设计与实现	0.75%
	ISPRS Workshop on Updating Geo-spatial Databases with Imagery The 5th ISPRS Workshop on DMGISs CHANGE DETECTION BASED ON SPATIAL DATA MINING	0.59%
	A novel heuristic Error-Driven learning for recogn……	0.54%
	ActionScript3.0从零基础学习类	0.31%
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%
	天嵌科技出品-Linux移植之Step By Step_V4.2_20100125	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%
张涛	大尺度环境移动机器人同时定位与地图构建算法的实现	0.2%
金山	筒式减振器油封的研究与开发	0.2%
徐淑周	喹诺里西丁类生物碱lasubine关键中间体的合成研究	0.18%
曹杰	广安须4气藏气井产能评价研究	0.18%
胡忠煜	织物图像自动拼接算法研究	0.18%
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	0.18%
龚治国	框排架结构设计软件的研制（II）	0.18%
蒋巍	人脸识别方法的研究	0.18%
唐小松	烟叶数据分析系统的设计与实现	0.17%
宁力	搜索引擎中网页查重方法的研究	0.16%
崔永刚	企业战略联盟内的信任机制构建研究	0.16%
宋佳	Effects of Written Output on Foreign Language Vocabulary Acquisition:A Case Study	0.16%

四、典型相似内容对比

1

当前位置: 0% **本段(页)重复比例:** 68.50%

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do exp*eriments by the* OMPN,the A-OMPN and the A-*OMPN with t*he gi*ven aerospacecraft dataset.Finally,we choose the OMPN to accomplish this module.Keywords:Duplicate Record Detection,Data Cleaning,Sorted-Neighborhood,NeuralNetwork*,Gen*etic Algorithm*第一章绪论第一章绪论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。

2

当前位置: 7.21% **本段(页)重复比例:** 18.11%

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。本文在分析常用相似*重复记录检测算法的基础上,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,需要对所有待处理的数据进行判等,*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。

3

当前位置: 8.11% **本段(页)重复比例:** 64.53%

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A-OMPN算法和基于BP神经网络判等的A-OMPN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。主要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。*首先简单介绍了衡量字段相西安电子科技大学硕士学位论文似度的*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。*

4

当前位置: 10.81% **本段(页)重复比例:** 82.31%

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 School1801001 Sam Water M 1993/01/02School of Computer Science,Xi ’ anUniversity of Electronic Science andTechnology1802002 Jack Panda Female 1990/07/20School of Artificial Intelligence,Xi ’ an University of ElectronicScience and Technology1801003 S.Water Male 1993/1/2Schol of Computer Science,Xi ’ anUniversity of Electronic Science andTechnology1802004 Jack Panda Female 1990/07/20School of Artificial Intelligence,Xi ’ an University of ElectronicScience and Technology1801005 Mr.Sam W Male 1993-01-02College of Computer Science,Xi ’ anUniversity of Electronic Science andTechnology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。

5

当前位置: 18.92% **本段(页)重复比例:** 11.66%

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 算法滑动窗*口的过程如图2.3 所示:.....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 算法的优点在于使用简单,滑动窗口判重过程效率较高,运行速度较快。

6

当前位置: 19.82% **本段(页)重复比例:** 50.93%

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)滑动窗口的大小w较难选择。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,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。西安电子科技大学硕士学位论文2.3.2 多趟近邻排序算法MPN 算法是在SNM 算法的*基础上的一种改进算法。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。

7

当前位置: 21.62% **本段(页)重复比例:** 17.24%

*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。第二章重复记录检测相关算法概述含有两趟SNM 过程的MPN *算法的流程图如图2.3 所示:*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 M*PN 算法流程图正是由*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。

8

当前位置: 50.45% **本段(页)重复比例:** 10.92%

*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第三章改进的OMPN 算法有三种[45]:(1)以缺失数据的均值、中位数、众数等统计计算结果填充缺失值;(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。

9

当前位置: 52.25% **本段(页)重复比例:** 25.70%

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.读取数据集,得到待检测的数据;西安*电子科技大学硕士学位论文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N 组排序关键字1 2{,,...}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条记录提取其排序关键字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ID 加入到缺失关键字记录集合*__iincomplete*set wi h keyt 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_idup set ;8.将_idup set 与*__iincomplete*set wi h keyt 进行重复归并,然后计算此集合的传递闭包itrans*itive_closure_set ;9.将N 次SNM 重复检测得到的itrans*itive_closure_set 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*_total dup set* 。

当前位置: 53.15% **本段(页)重复比例:** 36.67%

10 *含有两趟SNM 过程的OMPN 算法流程如图3.4 所示:第三章改进的OMPN 算法计算字段区分度生成*键值方案1生成每条记录的排序键值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1中合并标记记录集合1及重复记录集合1生成键值方案2生成每条记录的排序键值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2中合并标记记录集合2及重复记录集合2合*并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1标记的记录集合1合并后的重复记录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 算法流程图西安电子科技大学硕士学位论文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。

当前位置: 55.86% **本段(页)重复比例:** 62.24%

11 *第三章改进的OMPN 算法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*据生成器“ febrl ”[46](开源地址:https://sourceforge.net/projects/febrl/)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*说明字段名称字段描述举例1 举*例2rec_id 记录*ID rec-*454-org rec-454-dup-0culture 文化pak paksex 性别f fage 年龄30 30date_of_birth 出生日*期198*70221 19870221title 头衔hon hongiven_name 名字sophie sophiesurname 姓氏bozdar bozdaastate 州suburb 郊区holsworthy holsworthypostcode 邮编street_numbe*r 街*道号码46 46address_1 地址1 thurgood court thurgood courtaddress_2 地址2phone_number* 电话号码08 42167414 08 42167414soc_*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、重复比例、字段特征、错误类型、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。

当前位置: 62.16% **本段(页)重复比例:** 12.19%

12 *图3.5 SNM、MPN、*OMPN 算法查全率折线图西安*电子科技大学硕士学位论文图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,并作出如图3.8 所示的F1 值折线图。表3.*13 SNM、MPN、*OMPN 算法F1-measure 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 99.14910 87.712 95.725 98.967第三章改进的OMPN 算法20 87.648 94.336 98.76550 87.597 93.262 97.41380 86.537 90.126 97.114100 86.237 89.304 96.709200 84.847 88.672 94.166500 84.159 88.095 92.334图3.*8 SNM、MPN、*OMPN 算法F1-measure 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。

13

当前位置: 64.86% **本段(页)重复比例:** 25.24%

*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,比传统的SNM 算法和MPN 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 算法第四章改进的遗传神经网络算法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,对不同的字段分配相应的权值,这种方式表现不稳定,随着数据量的增加算法的查准率下降。

当前位置: 69.37% **本段(页)重复比例:** 41.43%

14 *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(4-1) *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],函数表*达形式如公式(*2-9)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,在本文中,初始权值取(1,1)之间的随机数。

当前位置: 78.38% **本段(页)重复比例:** 16.46%

15 对数据全集中的、非1D 中的其余记录进行排序;6.采用bestBP 进行可伸缩大小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 ;BP-OM*PN 算法流程图如图4.5 所示:第四章改进的遗传神经网络算法计*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BP-trained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80 98.58 98.53 98.39100 98.73 98.76 98.78200 99.10 99.

当前位置: 81.98% **本段(页)重复比例:** 8.35%

16 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 神经网络的学习速率设置为0.15,动量系数设置为0.9[60],最大训练次数为5000BPT ;遗传算法种群初始大小50sizem ,变异概率0.05mutaP ,种群最多迭代=300GAT 次。A-OMPN 算法和BP-OMPN *算法训练时间的结果如表4.4 所示:表4*.4 A-OMPN 和BP-OMPN 算法训练时间对比表数据集大小(条)A-OMPN 算法(s)BP-OMPN 算法(s)100 742.260 38.665200 2934.840 156.076500 18513.945 968.7421000 59564.372 3970.9582000 335082.875 16484.161由表格4.4 画出A-OMPN 算法和BP-OMPN *算法训练时间的结果对比*折线图如图4.9 所示:图4.9 A-OMPN、BP-OMPN 算法训练时间折线图由图4.9 可知,BP-OMPN 算法训练过程的时间消耗较低,且相对于A-OMPN 算法优势明显。

当前位置: 83.78% **本段(页)重复比例:** 20.32%

17 *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A-OMPN 算法,A-OMPN 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A-OMPN 算法训练网络的过程耗时严重,所以针对A-OMPN算法,提出了简化的BP-OMPN 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A-OMPN算法和BP-OMPN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BP-OMPN 算法训练较快。

18

当前位置: 84.68% **本段(页)重复比例:** 33.13%

第五章航天情报系统中的相似重复记录检测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、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。

19

当前位置: 85.59% **本段(页)重复比例:** 28.03%

*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 8iOS 操作系统为iOS*8.0 及其以上Andr*oid 操作系统为An*droid4.0 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。

20

当前位置: 93.69% **本段(页)重复比例:** 12.57%

(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 *getMyColle*ction GET返回当前登录账户收藏的航天器列表7 getSpace*craftDetail*ByID GET 返回某个航天*器的详细信息8 addToCollection*ByID 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。