PaperFree检测报告简明打印版

相似度：22.13%
编　号：AOSQHQIKUEDSADJP
标　题：面向重复记录检测的数据清洗算法的研究
作　者：张攀
长　度：10378字符
时　间：2018-04-10 21:01:10
比对库：中国学位论文全文数据库；中国学术期刊数据库；中国重要会议论文全文数据库；英文论文全文数据库；互联网资源；自建比对库

相似资源列表(学术期刊，学位论文，会议论文，英文论文等本地数据库资源)
1. 相似度：1.23%　篇名：《基于QPSO-LSSVM的数据库相似重复记录检测算法》
来源：《计算机科学》　年份：2016　作者：梁雪
2. 相似度：1.10%　篇名：《海量数据相似重复记录检测的研究》
来源：《桂林电子科技大学硕士论文》　年份：2011　作者：张平
3. 相似度：0.91%　篇名：《基于xml数据清洗的应用研究》
来源：《重庆大学硕士论文》　年份：2007　作者：谭亚竹
4. 相似度：0.73%　篇名：《多种群遗传神经网络在股指预测中的应用》
来源：《统计与决策》　年份：2014　作者：汪劲松
5. 相似度：0.73%　篇名：《构建数据仓库过程中的数据清洗研究》
来源：《图书与情报》　年份：2013　作者：刘喜文
6. 相似度：0.73%　篇名：《国际证券市场信息化基础设施发展趋势及启示》
来源：《证券市场导报》　年份：2013　作者：孙辰健
7. 相似度：0.58%　篇名：《浅析数据清洗》
来源：《计算机光盘软件与应用》　年份：2017　作者：田伟
8. 相似度：0.57%　篇名：《基于聚类树的相似重复记录检测算法改进研究》
来源：《合肥工业大学硕士论文》　年份：2015　作者：戴颖
9. 相似度：0.51%　篇名：《基于聚类算法的数据清洗的研究与实现》
来源：《华北电力大学（河北）硕士论文》　年份：2015　作者：张燕
10. 相似度：0.29%　篇名：《海量数据的相似重复记录检测算法》
来源：《计算机应用》　年份：2013　作者：周典瑞
11. 相似度：0.26%　篇名：《BP神经网络模型与学习算法》
来源：《软件导刊》　年份：2011　作者：樊振宇
12. 相似度：0.22%　篇名：《如何指导学生自学生物》
来源：《读写算（教师版）：素质教育论坛》　年份：2017　作者：王瑞
13. 相似度：0.20%　篇名：《统计数据质量评估方法研究述评》
来源：《统计与信息论坛》　年份：2011　作者：许涤龙
14. 相似度：0.19%　篇名：《基于遗传神经网络的相似重复记录检测方法研究》
来源：《舰船电子工程》　年份：2011　作者：肖蕾
15. 相似度：0.18%　篇名：《浅析中国兵器工业信息化发展趋势》
来源：《信息通信》　年份：2013　作者：李娜
16. 相似度：0.18%　篇名：《一种有效检测汉语相似重复记录的方法》
来源：《科技情报开发与经济》　年份：2011　作者：邹亚会
17. 相似度：0.17%　篇名：《基于SNM算法的大数据量中文地址清洗方法》
来源：《计算机工程与应用》　年份：2016　作者：郭文龙
18. 相似度：0.16%　篇名：《可达矩阵的Warshall算法实现》
来源：《安徽大学学报：自然科学版》　年份：2011　作者：叶红
19. 相似度：0.15%　篇名：《微纳传动系统的BP神经网络非线性控制》
来源：《重庆大学学报：自然科学版》　年份：2011　作者：林超
20. 相似度：0.13%　篇名：《基于相似光谱匹配预测土壤有机质和阳离子交换量》
来源：《农业工程学报》　年份：2014　作者：魏昌龙

21. 相似度：0.12% 篇名：《不确定数据的重复检测及清洗研究》
来源：《南京航空航天大学硕士论文》 年份：2012 作者：邓慧挺
22. 相似度：0.12% 篇名：《优化中职体育几种常见的教学方法》
来源：《祖国：教育版》 年份：2015 作者：王楠
23. 相似度：0.12% 篇名：《人脸识别技术探究》
来源：《硅谷》 年份：2011 作者：杨斌
24. 相似度：0.12% 篇名：《基于dbscan算法的相似重复记录检测方法研究》
来源：《哈尔滨工程大学硕士论文》 年份：2007 作者：崔亮
25. 相似度：0.11% 篇名：《浅析BP神经网络基本模型的C语言实现》
来源：《通信技术》 年份：2013 作者：赵朝凤
26. 相似度：0.11% 篇名：《BP神经网络的算法及改进》
来源：《中国西部科技》 年份：2017 作者：付海兵
27. 相似度：0.11% 篇名：《基于BP神经网络的我国企业人力资源需求预测分析》
来源：《对外经贸》 年份：2013 作者：潘珠
28. 相似度：0.10% 篇名：《人工神经网络在企业盈利能力评价与预测中的应用——以沪市52家上市公司为例》
来源：《商业会计》 年份：2014 作者：张振
29. 相似度：0.07% 篇名：《基于R2R的关联数据词汇异构问题研究》
来源：《情报科学》 年份：2014 作者：赵龙文

相似资源列表(百度文库，豆丁文库，博客，新闻网站等互联网资源)

1. 相似度：2.54% 标题：《——关于基于可变滑动窗口的相似重复记录检测算法研究与设计 - ...》
来源：http://www.doc88.com/p-1847569456102.html
2. 相似度：1.53% 标题：《基于可变滑动窗口的相似重复记录检测算法研究与设..._爱问共享资料》
来源：http://ishare.iask.sina.com.cn/f/j5Uup066IX.html
3. 相似度：1.36% 标题：《【公益-免费】-》相似重复记录清理方法研究综述 - 道客巴巴》
来源：http://www.doc88.com/p-253295025910.html
4. 相似度：0.73% 标题：《信息时代_ryxxrky_新浪博客》
来源：http://blog.sina.com.cn/s/blog_16aa015440102wywt.html
5. 相似度：0.60% 标题：《匹配算法JaroWinkler_百度文库》
来源：http://wenku.baidu.com/view/bf76ab8f70fe910ef12d2af90242a8956becaa7a.html
6. 相似度：0.51% 标题：《对基于MPN数据清洗算法的改进 -- 万方数据中小学数字 ...》
来源：http://edu.wanfangdata.com.cn/Periodical/Detail/jsjyyyrj200802089
7. 相似度：0.51% 标题：《对基于MPN数据清洗算法的改进 - 计算机应用与软件 - 金月芽期刊网...》
来源：http://www.jinyueya.com/magazine/10339130.htm
8. 相似度：0.48% 标题：《数据清洗研究综述》
来源：http://www.360doc.com/content/15/1016/14/28334341_506054329.shtml
9. 相似度：0.35% 标题：《一种检测多语言文本相似重复记录的综合方法pdf下载_爱问共...》
来源：http://ishare.iask.sina.com.cn/f/36041493.html
10. 相似度：0.24% 标题：《Python文本相似性计算之编辑距离详解_python_脚本之家》
来源：http://www.jb51.net/article/98449.htm
11. 相似度：0.22% 标题：《近似重复记录的增量式识别算法_图文_百度文库》
来源：http://wenku.baidu.com/view/6fd82ed78762caaedd33d4a7.html
12. 相似度：0.22% 标题：《梯度下降(Gradient Descent)小结 - 刘建平Pinard - 博客园》
来源：https://www.cnblogs.com/pinard/p/5970503.html
13. 相似度：0.17% 标题：《详细介绍一下smith waterman算法和Pfrap算法-CSDN论坛》
来源：https://bbs.csdn.net/topics/330239023
14. 相似度：0.14% 标题：《卷积神经网络(CNN)基础介绍 - CSDN博客》
来源：https://blog.csdn.net/fengbingchun/article/details/50529500
15. 相似度：0.14% 标题：《《现代经济信息》2016年第07期电子版 - 金月芽期刊网 免费论文下载》
来源：http://www.jinyueya.com/magazine/17727/2016_07/
16. 相似度：0.13% 标题：《梯度下降法 - CSDN博客 》
来源：https://blog.csdn.net/sql__learning/article/details/72638630
17. 相似度：0.13% 标题：《数据质量研究综述 - shuoma的日志 - 网易博客》
来源：http://blog.163.com/anby1314125@126/blog/static/288754042007411361439 5/

18. 相似度：0.13% 标题：《Levenshtein Distance算法（编辑距离算法） - 熊仔其 ...》
来源：http://www.cnblogs.com/xiongzaiqiren/p/4997947.html
19. 相似度：0.12% 标题：《大连理工大学信息检索研究室(DUTIR)-搜人搜物搜信息,重情重义重认知》
来源：http://ir.dlut.edu.cn/news/detail/482
20. 相似度：0.12% 标题：《两字符串的编辑距离 (Eidt Distance)_小虫..._新浪博客》
来源：http://blog.sina.cn/dpool/blog/s/blog_6f611c300101f72q.html
21. 相似度：0.12% 标题：《R语言与机器学习学习笔记(分类算法)(5)神经网络_instin..._新浪博客》
来源：http://blog.sina.com.cn/s/blog_5d29ee450101gmj6.html
22. 相似度：0.11% 标题：《深入理解FFM原理与实践 - 止战 - 博客园》
来源：http://www.cnblogs.com/zhizhan/p/5238415.html
23. 相似度：0.10% 标题：《损失函数-经验风险最小化-结构风险最小化 - CSDN博客》
来源：https://blog.csdn.net/qq_18343569/article/details/49824643
24. 相似度：0.10% 标题：《(六) 6.2 Neurons Networks Backpropagation Algor...》
来源：https://www.cnblogs.com/ooon/p/5284676.html
25. 相似度：0.10% 标题：《BP网络中误差反向传播过程是怎么样的? - 知乎》
来源：https://www.zhihu.com/question/36874225
26. 相似度：0.09% 标题：《BP神经网络的学习_百度文库》
来源：http://wenku.baidu.com/view/80c7ad745a8102d276a22fed.html
27. 相似度：0.09% 标题：《人工神经网络评价法_燕子_新浪博客》
来源：http://blog.sina.com.cn/s/blog_51e178450100jpu7.html
28. 相似度：0.09% 标题：《谁能说一下“感知器神经网络的应用”?_百度知道》
来源：https://zhidao.baidu.com/question/937985096843271932.html

全文简明报告

{100%：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。 }{81%：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。{97%：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，{ 59%：是近年来数据挖掘的研究重点。 }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”。{ 58%：脏数据的类型主要包括重复数据、不完整数据、错误数据等[ ][ ]。 }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，{ 55%：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[ ]。

{ 57%：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[ ][ ]。

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：

{ 64%：（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。 }

{ 74%：（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。 }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。

{ 71%：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。{ 63%：将出自不同数据源的数据集进行整合被认为是一个困难而且极为重要的问题，最早的研究主要是从数据连接[ ]、数据实体识别[ ]、对象识别等问题来展开，是商业保险、医疗、等领域中的研究重心之一。 }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[ ]。

数据清洗的研究重点包括：{ 56%：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。{ 59%：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，主要的工作有两个方面——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[ ]。

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Jaro提出的“排序&合并（Merge/Purge）”算法[ ]。这种算法存在明显的缺点，许多研究人员在此基础上提出了各种各样的改进思路和算法实现，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。

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[ ]。{ 73%：Hernandez等提出了多趟近邻排序算法[ ][ ]，即MPN（Multi-Pass Sorted Neighborhood），}{ 74%：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，最后使用C语言重写的OPS5[ ]规则编程判定记录是否相似。Qiu首先计算每条记录的N-gram统计值，{ 55%：然后根据这个N-gram值对数据集进行排序，}最后再用用优先级队列的方式聚类检测重复记录[ ]。Gianni Costa等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[ ]。Alfredo Ferro使用了基于q-grams的相似度衡量函数[ ]，可以避免许多不必要的比较和判断，提高了时间效率。

国内的相关研究主要是对已知算法的改进和创新以实现更高的精度和效率。复旦大学周傲英等比较早开始数据清理的研究工作[ ]。{ 57%：邱越峰等提出了基于N-Gram的相似记录检测算法[ ]。 }算法以一条数据的N-Gram值作为排序键，该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。{ 66%：陈伟提出了基于权重进行相似重复记录检测的方法[ ]，}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，{ 63%：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[ ][ ]。 }

{ 55%：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

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，{ 72%：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，{ 56%：尤其是较大数据量的相似重复检测问题。 }

{ 60%：（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。

{ 60%：（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。

{89%：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。 }{94%：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IMPN算法。IMPN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。

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。{ 57%：本文将反向传播神经网络应用于相似重复记录检测，}{ 62%：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，{ 59%：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算法的查准率。

1.4 论文结构

论文的结构如下所示：

{ 64% : 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。

{80% : 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。{ 58% : 然后对本文用到的BP神经网络理论基础进行说明。 }{ 58% : 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，{ 57% : 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。

{ 55% : 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 ，{ 71% : 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，{ 56% : 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。

{ 77% : 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：{ 59%：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。{ 59%：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：{ 58%：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[ ]。{ 73%：定义两个字符串 和 的编辑距离： }{ 55%：变成 需要对其单个字符进行插入、替换、删除操作的次数。 }{ 58%：编辑距离越小代表 和 越相似。 }

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。

{ 58%：Smith-Waterman算法[ ]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，{ 65%：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[ ]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。{ 63%：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：

其中，{ 71%：代表匹配的字符个数， 代表换位的数目。 }{ 59%：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：

其中， 是Jaro距离，{ 55%：是前缀的匹配长度， 是一个常数， }{ 60%：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，{ 66%：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[ ]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： ；

{ 61%：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，{ 66% : 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， }{93% : MPN）、优先队列算法（Priority Queue Strategy， }{ 63% : 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，{ 59% : 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 ，{ 63% : 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。 }

{ 58% : 滑动窗口的过程如图2.3所示： }

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。

{ 69% : （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

{ 69% : 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，{ 67% : 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。

{ 55% : 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。 }{ 71% : 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：

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：{ 63%：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[ ]。N-Gram值由记录中每个单词出现的概率综合计算得出，{ 65%：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。{ 63%：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。{ 63%：韩京宇等人提出了一种基于N-Gram层次空间的聚类算法DGHS[ ]，记录 、的N-Gram相似性计算公式如下： }

其中 表示记录 的所有字段组成的集合。{ 56%：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：

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[ ]，每个组成单元的结构功能并不复杂，{ 67%：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[ ]。{ 55%：典型的神经元模型如图2.6所示： }

其中，{ 64%：~ 是输入向量的不同分量， }{ 58%：~ 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，{ 56%：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

{ 59%：梯度下降法[ ]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。

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，{ 58%：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 。

{ 60%：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。{ 65%：是梯度下降法的学习速率，一般情况下， }{ 55%：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。

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”。{ 57%：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。

{ 62%：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，{ 59%：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

{ 74%：一个典型的三层BP网络如图2.7所示， }{ 59%：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，{ 58%：第三层是输出层，共有 个输出神经元。 }{ 57%：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，{ 56%：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，{ 69%：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。

如图2.7所示，前向传播过程如公式2-12和公式2-13所示，{80%：设 使用sigmoid函数。 }

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。

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。

在样本 上，参数的更新如公式2-19所示。

{ 57%：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

{ 61%：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。

{ 60%：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。