学校编码：10384　　分类号_____密级_____
学　号：24320101152248　　UDC_____

厦门大学

硕　士　学　位　论　文

数据挖掘中的重复记录检测算法研究

Research of Duplicate Record Detection Algorithm in Data Mining

何　玲

指导教师姓名：　廖　明　宏　教　授
专　业　名　称：　计算机软件与理论
论文提交日期：　2013　年　04　月
论文答辩日期：　2013　年　06　月
学位授予日期：　2013　年　　月

指　导　教　师：_____
答辩委员会主席：_____

2013 年 04 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

2013 年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

2013 年　　月　　日

摘 要

随着信息化的高速发展和客观上硬件技术的有效支撑，使得数据集中的海量数据不免存在冗余、缺失、不确定数据和不一致数据等诸多情况，这些统称为“脏数据”。人们需要从数据集中获取真实可靠的数据就必须进行数据清洗。而重复记录检测是数据清洗领域中的研究热点。本文首先介绍了数据仓库、数据清洗以及重复记录检测的相关知识，包括数据清洗的原理、方法、基本流程和工具以及重复记录检测匹配算法和重复记录检测清除算法。在此基础上，本文提出了三个改进算法。分别是基于传递闭包的重复记录检测算法，基于属性分析的重复记录检测算法和基于完全子图的重复记录检测算法。基于传递闭包的重复记录检测算法在 SNM 算法的基础上提出了三个方面的改进，分别是在排序步骤进行多趟排序，引入判断机制和引入特定权值和有效权值。基于属性分析的重复记录检测算法是在基于传递闭包的重复记录检测算法的基础上，提出两个方面的改进，通过属性分析，并引入过滤机制。在保证正确率的同时，有效的提高了算法的效率，减少算法的运行时间。基于完全子图的重复记录检测算法是针对前两个算法中因为应用传递闭包而产生误识别的问题而提出的改进算法。算法的解决方法是将相似记录集视为一个完全子图，将合并相似重复记录的问题转换为在连通图中寻找完全子图。最后，论文通过实验验证，表明改进的算法取得了良好的效果。

关键词：重复记录检测；SNM；传递闭包；完全子图

Abstract

Approximately duplicate records' cleaning is important in the field of data cleaning. Duplicate records detection is the process of identifying multiple records that refer to one unique real-world entity or object. However, due to different data representations in different data sources or errors because of various reasons, determining whether two records are equal is not a simple arithmetic predicate. Based on the existing duplicate records identification algorithms SNM and MPN, this paper proposes three improved algorithms about duplicate record detection algorithm. The first one is based on the transitive closure which analyzes attributes and sorts the dataset multiple times to make duplicate records more clustered considering that the sliding window size *w* is hard to select in SNM algorithm. Secondly, it gives a special weight to each attribute based on its contribution in the comparison and introduces the concept of effective weight so that to make the comparison more accurate and finally it merge the duplicate records by the method of transitive closure. The second algorithm proposed in this paper is based on attributes analysis, it analyze the attributes and orders them by their weights, then to improve the efficiency of detection by introducing filtering mechanism based on the analysis mentioned above. The third algorithm is based on complete sub-graph. By analyzing the MPN algorithm, it is clear that transitive closure in the merge step will cause higher false-positive rate. Our improved method treats a similar dataset as a complete sub-graph, and therefore the problem of duplicate records detection is converted to finding complete sub-graphs from an association graph where the vertexes represent data records and the edges reflect the similarity between records. At the same time, our algorithm effectively avoids the redetection of some parts of an already detected sub-graph. Further more, another improved algorithm based on the assumption that any two complete sub-graphs only have one common vertex in the association graph is proposed, as any two duplicate record sets only contain one same record, because after analyzing we find that it is a common phenomenon. And finally, the experimental results show that these advanced algorithms solve problem well.

Key Words：Duplicate Records Detection; SNM; Transitive Closure; Complete Sub-Graph

目 录

Contents

第一章 绪论

1.1 课题目的及意义

随着信息技术领域的不断发展和信息化建设的不断深入，当面对海量数据的同时，现实中的数据集也变得越来越错综复杂。数据中不可避免的存在冗余数据、缺失数据、不确定数据和不一致数据等诸多情况,这样的数据统称为“脏数据”。据统计，一些具有代表性的大公司的数据错误率预期在1%-5%左右，个别公司可能更高[1]。据报道，美国商业公司每年在处理“脏数据”上的花费都在上亿美元[48]。普化永道会计事务所在纽约的研究也表明，75%的被调查公司存在因“脏数据”问题而造成经济损失的现象，只有35%的被调查公司对自己的数据充满信心[49]。根据“垃圾进，垃圾出”的原理，错误的数据不仅将影响从数据集中抽取模式的正确性和导出规则的准确性，使得系统产生错误的分析结果，导致错误的决策[20]。而且从效率上来说，还将导致昂贵的操作费用和漫长的响应时间。因此，如何将“脏数据”有效的转化成高质量的干净数据是需要解决的首要问题，这涉及到数据清洗技术。

数据清洗也称为数据清理(Data Cleaning，Data Cleansing 或 Data Scrubbing)，目的是检测数据中存在的错误数据和不一致数据，并将它们剔除或者改正，以提高数据的质量[2]。假如输入数据的质量很差，那么数据挖掘返回的结果多半也将令人失望。数据清洗作为数据仓库技术中的重要一环，其实施效果如何直接决定了进入数据仓库的数据的质量，而数据质量是影响数据挖掘成功与否的重要因素，因其将进而影响决策支持系统的正确分析。因为数据仓库需要频繁地从各式各样的数据源中进行装载和刷新，而这些数据中不可避免地存在很多异常、冗余和错误，这就要求进入数据仓库前对数据进行清洗[3]。通过清洗，可以对残缺的数据进行修复、对错误的数据进行纠正和对多余的数据进行清除，将不正确的数据格式转换为所要求的格式，从而达到数据类型相同化、数据格式一致化、数据信息精练化和数据存储集中化的效果[4]。

“脏数据”按其不同的表现形式可具体概括为不完整数据、相似重复数据和错误数据三种类型[5]。其中由于多数据源合并而造成的信息重复是最关键的问题，因此重复信息的检测和清除成为一个研究的热点[6,7]。

1.2 重复记录检测研究现状

1.2.1 国外研究现状

国外对数据清洗技术方面的研究最早出现在美国，是从对全美社会保险号的错误进行纠正开始的。美国信息业和商业的发展促进了这方面工作的相关研究。识别并消除数据集中的相似重复对象，也就是重复记录的检测与清洗是数据清洗技术的一项重要研究内容。

相似重复记录识别，也称字段匹配，即选用合适的算法检测出标识同一现实实体的重复记录。这是重复记录检测中的核心步骤。现已有的算法多事针对不同错误类型的此类算法。大致分为以下几类。一是基于字符的相似性度量，能够很好的处理印刷、字符排序上的错误。主要有编辑距离算法(edit distance)[12]，仿射距离算法(affine gap distance)[13],Smith-Waterman算法[14]，Jaro距离算法(Jaro distance metric)[15]，和Q-gram算法[16, 17]。二是基于令牌的相似性度量，主要用来处理词语错位重排的问题，比如“Richard Smith”和“Smith，Richard”。主要有原子字符串算法(Atomic Strings)[18]和WHIRL算法[19]。三是语音相似性度量。基于字符和令牌的相似性度量都是主要针对基于字符串表示的数据集记录。但也有一种类型的字符串语音上相似但就字符而言并不相似。比如单词Kageonne与Cajun在语音上是相似的，尽管它们的字符表示相距甚远。语音相似性度量就是用来处理这类问题并进行字符串匹配。主要有探测法(Soundex),纽约模式识别与智能系统(New York State Identification and Intelligence System)[22]，名字压缩算法(Oxford Name Compression Algorithm)[23]，音位算法(Metaphone Algorithm)[24]，双音位算法(Double Metaphone Algorithm)[25]。

聚类算法能够辅助相似度的计算，在相符重复记录的识别中得到很好的应用。主要算法有，基于DBSCAN算法的自适应的密度聚类算法[34]；空间局部密度聚类算法[35]；基于网格和密度的聚类算法[37,38]；基于蚁群的聚类算法，主要应用在多维数据空间中[39]；自适应的粒子群优化聚类算法[40]。

对于一组检测出的相似重复记录有两种处理的方法。一是清除，即把一条记录看成是正确的，而其他记录则是含有错误信息的重复记录；二是合并，即把每一条检测出的重复记录看成是数据源的一部分，对这些记录进行合并，产生一条具有更完整信息的新记录。常用的算法有近邻排序法(Sorted

Neighborhood Method)[41,42]，多趟近邻排序法(Multi-pass Sorted Neighborhood)[41,42]，Delphi 算法(Delphi Algorithm)[43]，优先队列算法(Priority Queue Strategy, PQS)[44]等。

1.2.2 国内研究现状

由于中西文本身的差异性，国外数据清洗技术不能完全适用于中文数据的清洗。国内对于数据清洗的研究较晚，并且直接针对中文的数据清洗研究的成果也不多。尽管在一些学术期刊及会议上也能看到一些有关这方面的文章，但直接针对数据清洗，特别是中文清洗的研究还很少。国内现在主要是在数据仓库、决策支持、数据挖掘等方面的研究中做了些简单的阐述，而对于商业性的数据清洗工作主要是针对各自的具体应用，理论性不强。银行、证券等对数据的准确性要求较高的行业，都针对自己的具体应用开发相应的数据清洗软件，但很少公布理论性的东西。中文数据清洗在理论研究上的欠缺，也使市场上几乎看不到关于中文数据清洗的软件和工具[26]。然而随着数据仓库、客户关系管理系统等在企业中的大量应用，必然要求高质量的数据集支持，同时也将带动对中文数据清洗技术的研究和提高数据质量方法的研究，以及中文数据清洗工具的开发。

中文字段匹配指的是基于中文数据的字段匹配，中文字段匹配方法主要包括以下三类[68]：

一是字符串匹配方法。字符串匹配方法主要分为五种：单个字符的匹配方法、汉语自动分词方法、特征词匹配方法、词法分析得到的字符串匹配方法和中文缩写的回归字段匹配方法。单个字符的匹配的主要思想是逐步抽取某字符串中的单个字符与另一个字符串中的所有字符进行逐一比较，并将匹配成功的字符个数记录下来进行相似度计算，从而判断相似性。汉语自动分词方法是利用汉语的分词技术对字符串进行分词处理，得到单个分词的字符串，再用分词字符串作为匹配单位，进行匹配。特征词匹配方法是指只使用能够代表字段语言的关键词进行匹配，而不考虑其他的字符匹配。词法分析字符串匹配方法是指，先在字段中查找出一些具有特殊标识的字或词，然后将其从字段中取出，用这些字符串作为匹配单位，并根据这些具有特殊标识的字或者词的重要程度进行权值设定，最后通过匹配结果与权值的乘积值来判断相似性。中文缩写的回归字段匹配方法使用了文本值字段的回归结构进行匹配操作。

二是拼音匹配方法。中文常会出现同音字的现象。为了增大匹配的几率，有些时候需要用匹配单位的字符拼音进行匹配。其目的是解决汉语中一音多字的问题，具有实际意义，可作为字符串匹配方法的一种辅助方法，提高匹配精确度。主要步骤是，首先得到一张汉字与拼音的对照表，其中每个汉字和其对应的拼音组成数据库的一条记录，然后，使用字符型匹配算法进行匹配，将分词或汉字变成对应的拼音表示。接着用拼音串和字符或者分词作为匹配单位进行匹配。最后对于匹配结果通过事先定好的规则或策略进行筛选和抉择。

主要的记录匹配算法有，基于 Q-gram 层次空间的检测算法，主要用于大数据量的相似重复记录检测[27]；PCM 重复记录检测算法，它不仅可以应用在英文字符集合中，在中文字符集中效果也很好[28]；基于 N-Gram 的检测算法，主要适用于检测常见的拼写错误造成的重复记录[29]；基于优先队列的改进算法[30]；文献[31]中提出的基于 Canopy 聚类技术[32]的聚类算法，融合倒排检索对重复记录进行聚类以减小计算量；基于整体相似性的序列聚类算法(global similarity clustering)以及基于局部相似性的序列聚类算法(local similarity clustering)[33]，该算法具有较快的处理速度，并能取得较好的聚类质量；结合空间索引结构树与网格密度聚类算法的基于网格-密度与空间划分树的聚类算法，它能够仅在耗费线性时间复杂度的情况下发现任意形状的树[36]。

1.3 本文主要研究内容与结构安排

本文共分为六章。内容如下

第一章为概述。主要介绍课题的研究意义和目的。以及国内外重复记录检测的研究现状。

第二章主要介绍重复记录检测的相关理论知识。包括数据挖掘过程介绍；数据清洗的定义、原理、方法、基本流程和相关应用工具；重复记录检测知识介绍以及字段匹配算法和重复记录清洗算法介绍。

第三章提出基于传递闭包的重复记录检测算法，它是基于 SNM 算法的改进算法，主要改进的方面有排序步骤，引入判断机制和引入特定权值和有效权值的概念，在合并步骤，通过传递闭包对不同窗口中检测出的重复记录进行合并。最后通过实验验证算法的有效性。

第四章提出基于属性分析的重复记录检测算法，它在基于上一章算法的基

础上，提出了两个方面的改进，通过属性分析，并引入过滤机制。在保证正确率的同时，有效的提高了算法的效率，减少算法的运行时间。最后通过实验验证算法的有效性。

第五章提出基于完全子图的重复记录检测算法，它是针对前两章算法中因为应用传递闭包而产生误识别的问题提出的改进算法。算法的解决方法是将相似记录集视为一个完全子图，将合并相似重复记录的问题转换为在连通图中寻找完全子图。最后通过实验对算法进行验证。

第六章总结了本文所做的研究工作，并展望下一步的工作。

第二章 重复记录检测相关理论介绍

2.1 数据挖掘过程介绍

2.1.1 数据挖掘介绍

在信息爆炸的时代，人们面对以指数级增长的数据信息，渴望能够去粗取精、去伪存真的能将浩瀚无垠的数据转换成知识的技术，同时在客观条件上计算机硬件技术稳定进步，数据库技术日趋成熟。在这样的大背景下数据挖掘应运而生。

数据挖掘是按照既定的业务目标从海量的数据中提取出潜在的，有效的并能够被人理解的模式的高级处理过程。它是一门交叉性学科，融合了数据库技术、人工智能、机器学习、模式识别、统计学和数据可视化等多个领域的技术和理论。在较浅的层次上，它利用数据库管理系统的查询、检索和报表功能,并结合多维分析、统计分析方法进行联机分析处理(OLAP)，得出可供决策参考的统计分析数据。从深层次上来说,数据挖掘则从数据库中发现隐含的、潜在的知识。OLAP的概念最早是由关系数据库之父E. F. Codd于1993年提出的[45]。OLAP和数据挖掘都是从数据库中抽取有用的信息,就决策支持的需要而言两者是相辅相成的，OLAP旨在简化和支持联机分析,而数据挖掘的目的是尽可能使这一过程自动化。

图 2-1 显示了数据库中知识发现的过程[46]。我们可以看出，数据挖掘是数据库中知识发现(knowledge discovery in database, KDD)的核心步骤。

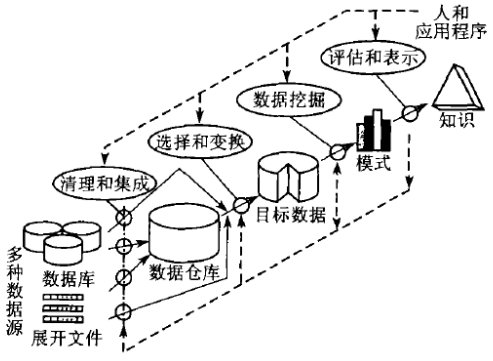

图 2-1 数据库中的知识发现

2.1.2 数据挖掘的方法

根据数据挖掘方法所属领域的不同大致可以分为以下几类。一是数学统计方