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（GA-ANN）求解重复检测问题的算法，将 OMPN 算法与 GA-ANN 相结合，得到准确度更高的 A-OMPN 算法和 BP-OMPN 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 A-OMPN，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

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（A-OMPN）。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、A-OMPN 算法与 A-OMPN 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

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 half-baked 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 Sorted-Neighborhood Method (SNM) and the Multi-Pass Sorted-Neighborhood Method (MPN), they all have shortcomings when tackle the real-world 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 Multi-Pass Sorted-Neighborhood Method (OMPN). In addition, we make a combination of the OMPN and the genetic-based artificial neural network to solve the problem and propose the Advanced Multi-Pass Sorted-Neighborhood Method (A-OMPN) and the BP network based Multi-Pass Sorted-Neighborhood Method (A-OMPN). The A-OMPN and the A-OMPN are superior to the other algorithms. Finally, we apply the proposed algorithm to the spaceflight information management system to accomplish the data-cleaning in the real-world problem. The main contributions of this paper are as follows:

1. The Optimized Multi-Pass Sorted-Neighborhood Method (OMPN) is proposed. The MPN first sort all the records and then use a scale-fixed 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 degree-based 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 half-baked data into account by pre-label scheme. Compared with other algorithms, the OMPN performs well and it’s suitable for solving the real-world duplicate records detection problem.

2. The Advanced Multi-Pass Sorted-Neighborhood Method(A-OMPN) is proposed. The genetic-based 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 time-consuming and the check stage can be simplified. The A-OMPN makes a combination of the genetic-based 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 genetic-based artificial neural network. However, to train an appropriate genetic-based artificial neural network is still time-consuming. We also do experiments with the signal BP network and then generate the BP network based Multi-Pass Sorted-Neighborhood Method(A-OMPN). Experimental results show that the A-OMPN and the A-OMPN 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 A-OMPN and the A-OMPN with the given aerospace craft dataset. Finally, we choose the OMPN to accomplish this module.

Keywords: Duplicate Record Detection, Data Cleaning, Sorted-Neighborhood, Neural Network, Genetic Algorithm