**参考文献**

【1】相似重复记录清理方法研究综述。叶焕倬 吴 迪 (中南财经政法大学信息与安全工程学院信息系 武汉 430073) ：56-57

【2】Galhardas H, F lorescu D, Sh asha D, et a.l Declarative Data C lean ing: Language , M odel and A lgorithm s[ C ]. In: P roceedings of the 27th International Conference on Very Large Data Bases, R om e, Ita ly. 2001: 371- 380

【3】B itton D, D e W itt D J . Duplicate Record E li m in ation in Large Data Files[ J]. ACM Transactions on Database System s, 1983, 8 ( 2 ): 255- 265.

【4】H ernandezM, Stolfo S. TheMerge /Pu rge Prob lem for LargeDatabases [ C]. In: Proceedings of theACM SIGMOD InternationalConference on Management of Data, S an Jose, California . 1995: 127- 138.

【5】H ernandezM A, Stolfo S J. Real- W orld Data is D irty : Data C leansing and th e M erge /Purge Prob lem [ J]. D ata M ining and K now ledge D iscovery, 1998, 2( 1): 9- 37.

【6】梁雪，任剑锋，景丽． 基于 ＱＰＳＯＬＳＳＶＭ 的数据库相似重复记录检测算法［Ｊ］． 计算机科学，２０１２，３９（１ １ ）：１５７ １５９．

【7】马翔． 粒子群优化 ＢＰ 神经网络用于重复记录检测［ Ｊ］． 辽宁工程技术大学学报：自然科学版，２０１０，２９（５）：９５９ ９６３．

【8】BiRon D，De Witt D J．Duplicate Record Elimination in Large Data Files [J]. ACM

Transactions on Database Systems，1 983，8(2)：255-265．

【9】Elma garmid A K，Ipeirotis P G, Verykios V S．Duplicate Record Detection：A Survey [J]. IEEE Transactions on Knowledge and Data Engineering，2007，19(1)：1-16．

【10】Jin L，Li C，Mehrotra S．Efficient record linkage in large data sets[C]．Eighth InternationalConference on DatabaseSystems forAdvancedApplications．Kyoto，2003：137-148．

【11】Wangy Y R，Madnick S E．The inter-database instance identification problem in integrating autonomous systems[A]．In：Wang Y R, Madrtick S E，eds．Proceedings of the Fifth International Conference on Data Engineering[C]．IEEE Computer Societh，1989：46-55．

【12】Galhardas H, Florescu D. An Extensible Framework for Data Cleaning[C]. In: Proceedings of the 16th IEEE International Conferenceon Data Engineering. San Diego, California, 2000: 312.

【13】陈伟 数据清理关键技术及其软件平台的研究与应用 [D]. 南京 :南京航空航天大学 ,2004:15-24.

【14】Flanagan T.,Safdie E. Apractical Guide to Achieving Enterprise Data Quality [DＢ /OL] .http://www.techguide.com,1999.

【15】Monge AE, Elkan C. An efficient domain-independent algorithm for detecting approximately

duplicate database records[C].In:Workshop on Research Issues on Data Mining and Knowledge

Discovery,1997:211-217.

【16】C. L. Forgy. 0PS5 User’s Manual. Technical Report CMU-CS-81-135, Carnegie Mellon University, July 1981.

【17】Newcombe HB,Kennedy JM,Axford SJ,etal.Automatic linkage of vital records[J].

Science,1959(130):954-959.

【18】Qiu Y F,Tian Z P,Ji W Y,et al.An efficient approach for detecting approximately duplicate

database records[J].Chinese J of Computers, 2001 (1):69-77.

【19】Gianni Costa, Giuseppe Manco, Riccardo Ortale.An incremental clustering scheme for data

de-duplication [C]. Data Mining and Knowledge Discovery, 2010(20):152-187.

【20】Alfredo Ferro, Rosalba Giugno, Piera Laura Puglisi, Alfredo Pulvirenti.An efficient duplicate

record detection using q-grams array inverted index [C].In: 2010 Proceedings of the 12th

international conference on Data warehousing and knowledge discovery,2010:309-323.

【21】郭志懋 周傲英

.数据质量和数据清洗研究综述

[J].软件学报

,2002 (11):2076-2082.

【22】邱越峰

,

田增平等

.

一种高效的检测相似重复记录的方法

[J].

计算机学报

,2001 (1):69-77.

【23】韩京宇

,

徐立臻

,

董逸生

.

一种大数据量的相似记录检测方法

[J].

计算机研究与发展，

2005,

(12):2206-2212.

【24】陈 伟

,

王 昊

,

朱 文 明

.

一 种 提 高 相 似 重 复 记 录 检 测 精 度 的 方 法

[J].

计 算 机 应 用 与 软

件

,2006,23(10):29-30.

【25】李 星 毅

,

包 从 剑

,

施 化 吉

.

数 据 仓 库 中 的 相 似 重 复 记 录 检 测 方 法

[J].

电 子 科 技 大 学 学

报

,2007,36(6):1273-1277.

【26】庞雄文

,

姚占林

,

李拥军

.

大数据量的高效重复记录检测方法

[J].

华中科技大学学报

(

自然科学

版

),2010,2(38):8-11.

【27】周丽娟

,

肖满生

.

基于数据分组匹配的相似重复记录检测

[J].

计算机工程

,2010:104-106.

【28】Raman, V., Hellerstein,J. Potter’s wheel:an interactive data cleaning system. Peoceedings of the 27th International Conference on Very Large Data Bases. Roma:Morgan Kaufmann, 2001:381-390.

【29】Lee, M.L., Ling, T.W., Low, W. L. IntelliClean: a knowledge-based intelligent data cleaner. Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston:ACM Press, 2000:290:294.

【30】陈伟 数据清理关键技术及其软件平台的研究与应用[D].南京:南京航空航天大学,2004:15-24.

【31】Monge A E, Elkan C P. The field matching problem:a lgorithms and

applications[C]. International Conference on Knowledge Discovery and Data

Mining, Portland,1996:267-270.

【32】 Elmagarmid A E, Ipeirotis P G, Verykios V S. Duplicate record detection: A

survey[J]. IEEE Transactions on Knowledge and Data Engineering. 2007,

19(1): 1-16.

【33】Alvarez V P, Vazquez J M. An evolutionary algorithm to discover

quantitative association rules from huge databases without the need for an a

priori discretization[J]. Export System with Applications, 2012,39(1):585-593.

【34】Levenshtein V. Binary codes capable of correcting spurious insertions and

deletions of ones[J]. Problems of Information Transmission, 1965,1(1):8-17.

【35】Liu Xuhui, Li Guoliang, Feng Jiahua, Zhou Lizhu. Effective indices for

efficient approximate string search and similarity join[C]. International

Conference on Web-Age Information Management, Zhangjiajie, 2008:127-134.

【36】赵作鹏,尹志民,王潜平. 一种改进的编辑距离算法及其在数据处理中的应用[J].计算机应用,2009,29(2):424-426.

【37】Wildani A, Miller E, Rodeh O. HANDS: A heuristically arranged non-backup

in-line deduplication system[C]. IEEE International Conference on Data

Engineering, Brisbane, 2013: 446-457.

【38】Hassanzadeh O, Miller R J.Creating probabilistic databases from duplicated

data[J].VLDB Journal.2009,18(5):1141-1166.

【39】Hy lton J A. Iden tify ing and M erging Related B ib liograph ic Records

[M ]. C amb ridge, MA, USA: M assachu setts Institu te ofT echnolo

gy, 1996: 46 - 50

【40】H ernandezM A. A Generalization of Band Jo ins and th e M erge/

Pu rge Problem [ D]. N ew Y ork: Colum b ia Un iversity, 1996.

【41】R ehm an M, E sichaiku l V. Dup licate Record Detection for Data

base C leansing[ C]. In: P roceedings of the 2nd International Con

ference onM achine Vision, Duba,i U n ited Arab Em irates. 2009:

333- 338.

【42】钟嘉庆, 张义芳, 卢志刚. 数据仓库中重复记录清理算法研究

[ J]. 微型机与应用, 2009, 28 ( 7): 4- 6.

【43】杨宏娜

.

基于数据仓库的数据清洗技术研究

[D].

天津

:

河北工业大学

,2006,24-26.

【44】Oktie Hassanzadeh,Renée J. Miller.Creating probabilistic databases from duplicated data[J]. The

VLDB Journal.2009,18(5):1141-1166.

【45】李坚, 郑宁. 对基于 MPN数据清洗算法的改进 [ J]. 计算机应

用与软件, 2008, 25( 2): 245- 247.

【46】M onge A, E lkan C. An E fficient Doma in Ind ependen t A lgorithm

for Detecting Approxi m ately Dup licate Database Record s[ C ]. In:

P roceedings of the SIGMOD W orkshop on DataM ining and K now l

edge D iscovery, Tu cson, Arizona . 1997: 23- 29.

【47】朱恒民, 王宁生. 一种改进的相似重复记录检测方法 [ J]. 控

制与决策, 2006, 21( 7): 805- 808, 813.

【48】Lee,M.L, Ling, T.W., Low, W.L. IntelliClean: a knowledge-based intelligent data cleaner. Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston:ACM Press,2000:290-294.

【49】<https://www.zhihu.com/question/60725976>

【50】https://zhuanlan.zhihu.com/p/20571505