数据匹配算法应用对比研究
——以期刊数据融合中作者和机构匹配为例

盛怡瑾，张学福，孙巍，郝心宁
（中国农业科学院农业信息研究所，北京 100081）

摘要：为了评价数据匹配算法中常用的四种字段匹配算法——Smith-Waterman算法、编辑距离（Edit Distance）、Q-gram算法和Jaro-Winkler算法的效果和表现，本文选取由水稻领域18个重点期刊集成得到的作者和机构数据设计实验，使用Febrl清洗工具包对相似重复记录进行匹配。结果表明，四种算法适用条件不同，Smith-Waterman算法运行时间特别长，但综合表现以及精度和召回率都不错；编辑距离（Edit Distance）性价比比较高；Q-gram算法运算快但召回率低；Jaro-Winkler算法在此例中表现比较差。

关键字：数据清洗；数据匹配；期刊；作者；机构
中图分类号：TP391
DOI：10.3772/j.issn.1673-2286.2015.10.003

1 引言

数据质量是进行一切数据分析的前提。要提高数据质量，就必须要进行数据清洗，其核心是数据匹配。近些年来，随着数据仓库和数据挖掘技术的兴起和完善，对多种来源的数据进行合并变得越来越频繁，这个过程会产生大量的相似重复，也就是我们所说的数据匹配问题。因此，解决好数据匹配的问题有着非常重要的意义。

在过去的很长时间，人们往往依赖人工对数据进行匹配，效率很低。到后来，基于规则的方法能减轻一部分人工负担，但这种方法需要规则的设计者既了解领域知识，又了解整个数据集特点，所以当数据量变大，数据类型多样的时候就不太适用。渐渐地，们希望能够借助如数据挖掘等算法自动地匹配数据，并在这方面进行了大量的研究。使用字段匹配算法对字段的相似性进行判断，可以在不需要领域知识和专家协助情况下，自动地识别出相似重复记录，极大地提高效率，因此是未来发展的方向。

2 简单介绍

2.1 基本概念及流程

数据清洗（Data Cleaning，也称 Data Cleansing 或Scrubbing）的目的是检测和消除数据中的错误和不一致，以提高数据质量[1]。

数据匹配是将客观上表示现实世界同一实体的，但是由于在格式、拼写上有些差异而导致数据库管理系统不能正确识别的记录（也叫相似重复记录）进行识别的过程。

数据匹配要处理的问题主要有三个：一是实体错误，如误植、拼写错误、格式错误、字段遗漏等；二是编目标准差异；三是缩写[2]。

目前数据匹配的流程是，先依照一定标准对数据进行分块，再对同一分块内的记录两两对比，在两条记录对比时，要判断二者是否指向同一实体，则一般需要进行字段匹配和记录检测，一般先进行字段匹配，再综合字段匹配的结果来进行记录检测（见图1）。

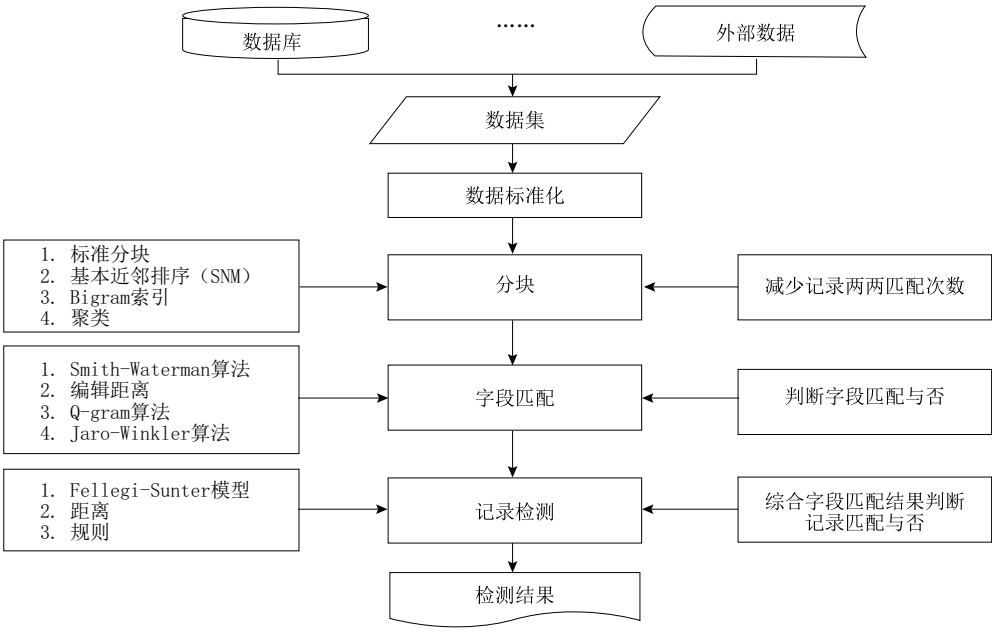

图1 数据清洗流程

分块的目的是为了改进传统方法中将每一条记录与数据库中所有其他记录进行比较时O（n^2）的时间复杂度。目前记录检测的主流方法是数据挖掘方法中的概率方法，也叫做Fellegi-Sunter模型。该概率模型的思想是，输入比较向量（Comparison Vector）x，通过一定的决策规则将x分配给不匹配集合U或者匹配集合M。对于这两类集合，x有不同的密度函数，如果密度函数已知，那么相似重复记录识别的问题就变成了贝叶斯推断问题。

2.2 字段匹配算法及分析对比

字段匹配主要使用字符串相似度算法，清洗中主要使用的有Smith-Waterman算法，编辑距离（Edit Distance），Q-gram算法和Jaro-Winkler算法，四种算法各有特点和适用情况（见表1）。

表1 四种字段匹配算法分析对比

	Jaro-Winkler	编辑距离	Smith-Waterman算法	Q-gram算法
时间复杂度	$O(\|s_1\|+\|s_2\|)$	$O(\|s_1\|*\|s_2\|)$	$O(\|s_1\|*\|s_2\|)$	$O(\max\{\|s_1\|,\|s_2\|\})$
特点	添加前缀范围因子，增加字符串起始部分的权重	考虑对单个字符的最小操作数量（插入，删除，替换）；复杂度高	对开头和结尾的不一致比较宽容；计算复杂；矩阵在计算其值时具有对称性，可减轻运算量	计算两字符串中任意连续q个字母的有序组合——q-gram向量间的距离来判断相似度；灵活；计算迅速
衡量标准	Jaro-Winkler距离的值越小越不相似	编辑距离越大，两字符串越不相似		q-gram向量间的距离越小，相似度越高
适用情况	短字符串（如姓的比较）	误植（键入/拼写）错误；对长度差异很大的两字符串进行快速过滤	长字符串；字符串有缩写、缺失和少量语法不同的情况	格式不一致；误植；常见拼写错误；比编辑距离更适合处理略长一点的字符串
不适用情况	长字符串；前缀相同但意义不同的两字符串	字符串与其缩写形式的匹配	字符串顺序颠倒的情况	

（1）Smith-Waterman算法

Smith-Waterman算法[3]最初是用来找到生物蛋白和DNA序列之间的演化关系的。算法如下：

m，n分别是两字符串a，b的长度；

建立矩阵H，H(i,0)=0，0≤i≤m；H(0,j)=0，0≤j≤n

$$H(i,j)=\max\begin{cases}0\\ (H(i-1,j-1)+s(a_i,b_j)，匹配/不匹配\\ \max_{k\geq 1}\{H(i-k,j)+W_k\}，删除\\ \max_{l\geq 1}\{H(i,j-l)+W_l\}，插入\end{cases}，1\leq i\leq m，1\leq j\leq n$$

s(a_i,b_i)表示字符串中字符的相似函数，一般是相同则加分，不同则减分；

H(i,j)表示两字符串后缀的最大相似系数；

W_i是gap得分系数。

该算法认为字符串开头和结尾的错配对于两字符串的相似度影响不大，因此对前缀和后缀的错配给予了比中间子串错配更低的权重。

Smith-Waterman算法时间复杂度是O（$|s_1|\cdot|s_2|$）。计算是比较复杂的。可以使用动态编程技术实现，矩阵在计算其值时具有对称性，可减轻运算量。

基于记录有相同结构（对置换字母不敏感）和字母数字的假设，该法比较适合长字符串之间的比较。对于字符串中有缩写、缺失信息或者少量句法不同的情况比较适用[4]。

（2）编辑距离（Edit Distance）

由Levenshtein[5]提出，指的是从一个字符串转换到另一个字符串所需的对单个字符的最小操作数量（插入、删除、替换）。时间复杂度是O（$|s_1|\cdot|s_2|$）[6]。编辑距离越大，两字符串相似度越低。在实际的计算机操作过程中，编辑距离算法是一个动态编程算法，需要进行递归，复杂度高。

编辑距离算法对误植（Typographical Error），即键入和拼写错误非常有效。但对于字符串与其缩写形式的匹配效果并不理想[7]，比较适合短字符串。此外，编辑距离还可以用于对长度差异很大的两字符串进行快速过滤[8]。

（3）Q-gram算法

字符串的一个Q-gram 是指该字符串中任意连续q个字母的有序组合[9]。通过计算两字符串的Q-gram向量间的距离来判断相似度。

该算法认为，如果两个字符串相似，那么它们就有很多共同的Q-gram。其时间复杂度为O（max{$|s_1|$, $|s_2|$}）。

Q-gram算法对格式和误植不敏感，能适应常见的拼写错误。

（4）Jaro-Winkler算法

设两字符串的Jaro 距离是d_j，两字符串拥有的共同前缀的长度为L，前缀的范围因子是p，Jaro-Winkler 距离[10]的计算公式是：

$d_w=d_j+L\cdot p(1-d_j)$

L最大是4个字符，p不能超过0.25，一般给p赋值为0.1。d_w越大，说明两字符串相似度越大。

作为Jaro算法的变体，Jaro-Winkler算法也适用于短字符串。相比之下，Jaro-Winkler算法认为，起始部分相似的字符串，其相似的可能性更高，所以增加这种情况的权重。Jaro-Winkler算法速度是比较快的，曾用于美国人口统计中姓的比较。

3 对比研究实验设计

在图书情报领域的实际研究中，无论是开放获取期刊的集成还是领域知识库的构建，都涉及到将不同期刊甚至是网络主页上的文章信息、作者信息、机构信息等集合起来，对作者和机构的整合是必要的步骤。一个作者会在多个期刊上发表文章，不同期刊对所刊文章的作者和机构的格式、粒度和顺序要求都有差异，不同数据库的格式要求也不同，同一数据库不同时期的格式要求也可能不同。比如Web of Science上早期的作者名都使用缩写，后来才采用全名。作者也会在网络主页上发表作品，网络上的资源会存在一些拼写错误。那么，当我们把这些不同来源的期刊数据进行融合并建立作者机构表时，就会存在若干条不完全相同的记录指向同一个作者（如表2所示），因此必须进行数据匹配以保证唯一性。这就需要我们使用数据匹配算法对数据进行清洗。

3.1 数据来源

在进行数据匹配时，字段匹配算法的正确选择对最终的结果意义重大。本文选择了水稻领域的18个重点期刊：Plant and Soil， Plant Science，Rice Genetics Newsletter，International Rice Research Notes等，将这些期刊上的作者机构信息集成得到的12万条原始数据，对作者和机构进行数据匹配，并对比研究不同算法

表2 数据集中作者机构相似重复情况

Author	Affiliation
A.K. Singh	Narendra Deva University of Agriculture & Technology, Narendra Nagar, Kumarganj, Faizabad-224 22% U.P., India
A.K. Singh	N.D. University of Agriculture and Technology, Crop Research Station, Masodha, Faizabad-224229, U.P., India
A.K. Singh	Department of Genetics and Plant Breeding, N. D. U. A. T., Kumarganj, Faizabad-224 229 (U.P.) India
Abate Z.A.	Department of Extension Education, CCS Haryana Agricultural University Hisar 125004, India
Abate ZA	Department of Extension Education, CCS Haryana Agricultural University Hisar 125004, India

的表现。原始数据存在的数据匹配问题有，同一作者有的记录用全称，有的用缩写；同一机构有的用全称，有的用缩写；机构描述顺序不同；地址描述粒度不同，拼写错误等。

3.2 数据分组

根据对比需要，从12万原始记录中随机抽取样本，构建小（275条记录）、中（550条记录）、大（1300条记录）三种数据集，并人工标记出若干对相似重复记录，即匹配对，便于对实验结果进行评价。每种数据集都按照低（25%）、中（50%）、高（70%）三种重复率构建，最后共生成9个数据集（如表3所示）。

表3 数据集构建表

数据集（条）	重复率	重复记录数（条）
275	低	69
	中	138
	高	206
550	低	137
	中	275
	高	413
1300	低	325
	中	650
	高	975

3.3 工具及评价指标

使用数据清洗工具包Febrl对数据进行处理。Febrl是开源的数据清洗工具包，既可以进行数据标准化，又可以进行相似重复检测。该系统整合了26种不同的字段匹配算法，同时也支持三种分块方法。为了检测上文提到的四种字段匹配算法的表现，本实验采用控制变量法（如表4所示），在分区和记录匹配阶段固定使用标准分块方法和Fellegi-Sunter模型，在字段匹配阶段对每一数据集分别使用Smith-Waterman算法、编辑距离（Edit Distance）、Q-gram算法和Jaro-Winkler算法进行操作，并对结果进行分析。

表4 控制变量法使用说明

阶段	分块	字段匹配	记录检测
算法	标准分块方法	Smith-Waterman算法 编辑距离 Q-gram算法 Jaro-Winkler算法	Fellegi-Sunter模型
控制变量法中的角色	常量	变量	常量

对结果的质量评价使用召回率（Recall），精度（Precise）和综合评价指标F值。

$$召回率Recall=\frac{TP}{(TP+FN)}$$

$$精度Precise=\frac{TP}{(TP+FP)}$$

$$F=2\frac{(Recall\times Precise)}{(Recall+Precise)}$$

其中，TP（True Positive）指的是真正匹配的记录对被识别为匹配对，FP（False Positive）指的是真正匹配的记录对被识别为不匹配，TN（True Negative）指的是真正不匹配的记录被识别为匹配，FN

（False Negative）指的是真正不匹配的记录被识别为不匹配。

召回率（Recall）衡量的是数据匹配的全面性，精度（Precise）衡量的是数据匹配的精确性，F值是对结果的综合衡量。

4 实验结果对比评价

对运行时间进行对比（如表5所示）表明，Smith-Waterman算法的运行时间最长，其时长比其他三种算法高出两个数量级，其次是Edit distance运行时间，比其他两种高出一个数量级，接下来是Jaro-Winkler，运行时间最短的是Q-gram，只用几秒就可完成。实验同时表明，数据集越大，耗费时间越长；重复率越低，耗费时间越长，但数据集过小时这种趋势不明显。

对F值进行对比发现（如图2—4所示），数据集大小对每种算法的F值得分影响不太大；Smith-Waterman,Edit distance,Q-gram效果差不多，但相对来说Smith-Waterman 和Q-gram 表现更好一点；Jaro-Winkler相比其他三种算法，效果比较差；重复率越高，F值越高。

再单纯考虑精度（Precise），也就是算法能够准确识别真正匹配对的能力（如表6所示），Q-gram>Smith-Waterman> Edit distance> Jaro-Winkler，说明Q-gram算法识别出的匹配对比较准确，Smith-Waterman次之，Jaro-Winkler识别出的相似重复记录最少。并且，还可以从表中看出，数据集的重复率越低，精度越低。

单纯考虑召回率（Recall），也就是算法能够更多地提供匹配对的能力（如表7所示），Jaro-Winkler> Edit distance>Smith-Waterman>Q-gram，说明Jaro-Winkler会将大量的非匹配记录对标记为匹配对，而Q-gram有可能识别出的匹配对比较少，会把一些真正匹配的记录没有识别进来。同时，表中也显示，数据集重复率越低，召回率越低。

表5 运行时间对比

重复率	high			medium			low		
运行时间 / 数据集 / 算法	275	550	1300	275	550	1300	275	550	1300
Smith-Waterman	840	2400	3420	1380	2040	3120	1320	1680	2880
Edit distance	12	27	46	20	26	41	18	22	35
Q-gram	1.26	1.58	2.20	1.58	1.54	2.00	1.40	1.47	1.76
Jaro-Winkler	1.78	2.80	4.14	2.50	2.76	3.90	2.36	2.50	3.40

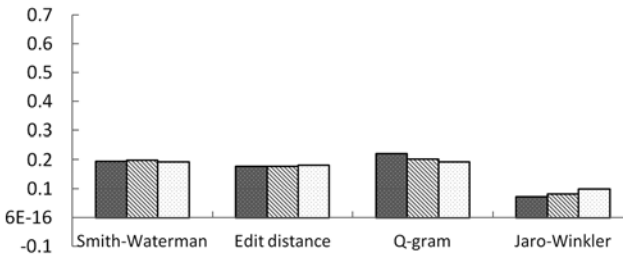

图2 25%重复率时F值

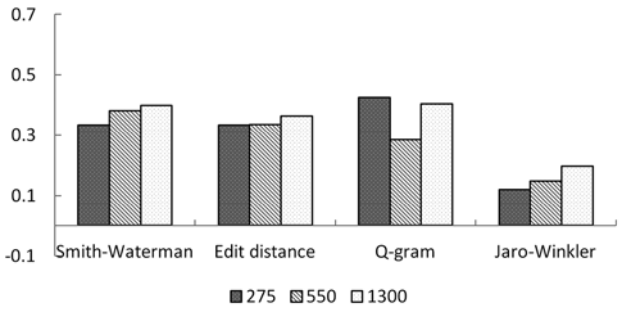

图3 50%重复率时F值

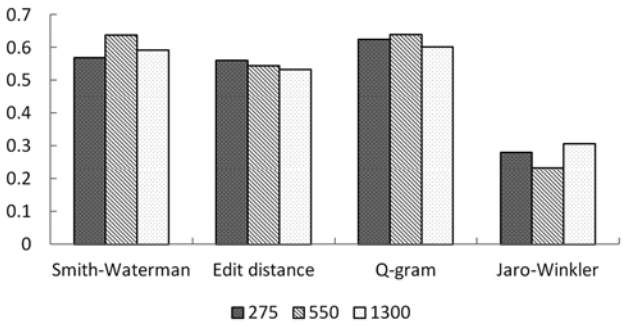

图4 75%重复率时F值

表6 精度对比

重复率	high			medium			low		
精度 数据集 / 算法	275	550	1300	275	550	1300	275	550	1300
Smith-Waterman	0.533	0.665	0.596	0.303	0.429	0.598	0.241	0.348	0.475
Edit distance	0.506	0.477	0.472	0.284	0.305	0.428	0.174	0.204	0.304
Q-gram	0.684	0.720	0.647	0.554	0.709	0.692	0.395	0.458	0.591
Jaro-Winkler	0.174	0.137	0.196	0.067	0.087	0.135	0.043	0.050	0.073

表7 召回率对比

重复率	high			medium			low		
召回率 数据集 / 算法	275	550	1300	275	550	1300	275	550	1300
Smith-Waterman	0.609	0.609	0.585	0.368	0.342	0.299	0.164	0.140	0.121
Edit distance	0.624	0.632	0.608	0.400	0.369	0.315	0.182	0.157	0.129
Q-gram	0.574	0.574	0.561	0.344	0.178	0.284	0.154	0.131	0.116
Jaro-Winkler	0.696	0.729	0.696	0.496	0.468	0.364	0.242	0.225	0.155

5 结论与展望

在期刊数据合并时的作者机构数据匹配中，我们发现，Smith-Waterman算法运行时间特别长，但综合表现以及精度和召回率都不错，如果不考虑时间因素，是一种效果比较好的算法。编辑距离（Edit distance）运行时间也不短，但比Smith-Waterman快一些，其表现比较中庸，属于性价比比较高的一种算法。Q-gram算法运算快，精度高，但召回率低，说明它在进行匹配的时候，精确性比较好，但全面性不好。Jaro-Winkler算法在此例中表现比较差，召回率高，精度低，它会把许多非匹配的记录认为是匹配的。

此外，实验表明，数据集越大，运行时间越长；数据集的重复率越低，精度、召回率和F值都越低，也就是说数据匹配的效果越不好。而现实世界的用于数据匹配的数据集，往往具有大量甚至海量且重复率低的特点，因此如何改进算法以克服这一点，将是未来研究的方向。

在未来的研究中，可以考虑结合四种算法的优点以提高效果，比如可以将运行时间最短的Q-gram用于快速的粗略筛选。将Edit distance用于大多数字段的匹配，将Smith-Waterman用于量少但重要的字段的匹配。同时，应用于算法的字段并不是同样重要的，有的字段在决策中更加重要，因此需要更大的权重。通过人工神经网络方法可以对字段赋予科学的权重，也可以考虑使用遗传算法对参与计算的属性进行筛选。

参考文献

[1] Rahm E, Do H H. Data cleaning:Problems and current approaches[J]. IEEE DATA ENG.BULL, 2000, 23(4):3-13.

[2] Hylton J A. Identifying and merging related bibliographic records[R]. MIT:MIT LABORATORY FOR COMPUTER SCIENCE TECHNICAL REPORT 678,1996.

[3] Smith T F, Waterman M S. Identification of common molecular subsequences[J]. JOURNAL OF MOLECULAR BIOLOGY,1981,147(1):195-197.

[4] Monge A E, Elkan C. The Field Matching Problem:Algorithms and Applications[C].KDD,1996:267-270.

[5]Levenshtein V I. Binary codes capable of correcting deletions,insertions and reversals[C]. SOVIET PHYSICS DOKLADY,1966:707-710.

[6] Monge A E. Matching algorithms within a duplicate detection system[J]. IEEE DATA ENG.BULL,2000,23(4):14-20.

[7] Elmagarmid 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.

[8] Nobarian M S, Derakhshi M R F. The Review of Fields Similarity Estimation Methods[J]. INTERNATIONAL JOURNAL OF MACHINE LEARNING AND COMPUTING ,2012,2(5): 614-617.

[9] 邱越峰,田增平,季文赟,等. 一种高效的检测相似重复记录的方法[J]. 计算机学报,2001,24(1):69-77.

[10] Winkler W E. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage[EB/OL].[2015-09-09]. http://www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf

作者简介

盛怡瑾，女，1988年生，中国农业科学院农业信息研究所硕士研究生，研究方向：信息资源管理，E-mail：shengyijin_syj@163.com。
张学福，男，1966年生，中国农业科学院农业信息研究所研究员，研究方向：农业知识组织与可视化分析，通讯作者，E-mail：zhangxuefu@caas.cn。

Comparative Study of Application for Data Matching Algorithms: Taking Author and Institution Matching in Journal Data Fusion as an Example

SHENG YiJin, ZHANG XueFu, SUN Wei, HAO XinNing
(Agricultural Information Institute of CAAS, Beijing 100081, China)

Abstract: To evaluate the effect and performance of four field matching algorithms commonly used in data matching——Smith-Waterman algorithm, Edit Distance, Q-gram algorithm and Jaro-Winkler algorithm, we chose authors and institutions information integrated from 18 key journals to design experiments, using Febrl to match approximate records. The results showed that the four algorithms have different applicable conditions. Smith-Waterman algorithm runs a particularly long time, but the overall performance, the precision and recall are good. Edit distance is relatively high cost-effective. Q-gram algorithm runs fast but has low recall. Jaro-Winkler algorithm doesn' t perform well in this case.

Keywords: Data Cleansing;Data Matching;Journals;Author;Institution

（收稿日期：2015-09-18；编辑：雷雪）