arXiv:2302.07784v1 [cs.DL] 15 Feb 2023

# 原住民数据库历史档案中个人记录匹配案例研究

Matthew Currie 和 Ramon Lawrence

不列颠哥伦比亚大学基洛纳分校计算机科学系，加拿大不列颠哥伦比亚省基洛纳，V1V 2Z3 mfsc@student.ubc.ca，ramon.lawrence@ubc.ca

**摘要**。 历史记录的数字化产生了大量可供分析和解读的数据。一项关键挑战是如何关联不同档案库中的历史信息，以便将数据置于恰当的历史语境中。本文介绍了一项关于历史信息集成与记录匹配的真实案例研究，旨在提升包含1800年至1920年间数据的档案的历史价值。这些档案包含有关加拿大梅蒂人和原住民以及他们与欧洲定居者互动的独特信息。档案中有数千条记录，发现其中的关系和相互联系后，这些记录的价值将进一步提升。本文的贡献在于提出一种适用于历史档案的记录链接方法，并评估其有效性。实验结果表明，该方法有望以较高的精确率发现历史关联，从而带来新的历史发现。

**关键词：** 记录链接、记录匹配、历史、数据库、档案库、原住民

## 1 引言

数字化的历史数据使历史学家能够跨越各种数据源和时间段搜索、分析并关联数据，以用于多种应用。随着纸质历史记录被转换为电子格式并加载到数据档案库中，维护数据质量和完整性至关重要。在数据加载过程中，记录会被保存下来，而确保历史准确性并确定档案库中已有数据之间的相互关联是一项艰巨的任务。数据记录之间的相互关联具有重要价值，因为只有将个人记录联系起来，才可能更详细地了解个人随时间的变化。要关联不同档案中的记录，需要采用为数据库集成开发的记录匹配和链接技术 [3,6] 。

历史背景下的记录匹配同时面临数据记录和转录不精确，以及可用数据量有限这两方面的挑战。常见的记录匹配应用是，在一个或多个数据库中识别属于同一人或同一客户的记录 [3]。已发表的

方法和商业系统通过利用记录中的详细信息实现较高的匹配准确率，这些记录通常包含姓名、地址、电话号码、电子邮件以及用于身份识别的信息，例如国家身份标识符（如社会安全号码）等字段。在历史背景下，可用数据仅限于姓名和记录事件发生的日期。这限制了进行准确匹配的能力。另一个关键目标是在尽量减少假阳性的情况下实现较高的精确率，从而确保历史记录中的人员仅被正确匹配。在许多应用中，与其提出错误的关联，不如不进行匹配。

本文针对数字档案数据库项目（DADP）中的历史数据集开展记录匹配案例研究。该项目包含1800年至1920年间加拿大超过165,000条有关M´etis、原住民和欧洲定居者的记录。这个独特的数据集通过对众多独立数据源进行数字化和整合而建立。历史学家需要在各数据源之间进行记录链接，以确定某个人何时参与了多个历史事件。

本文的贡献包括：

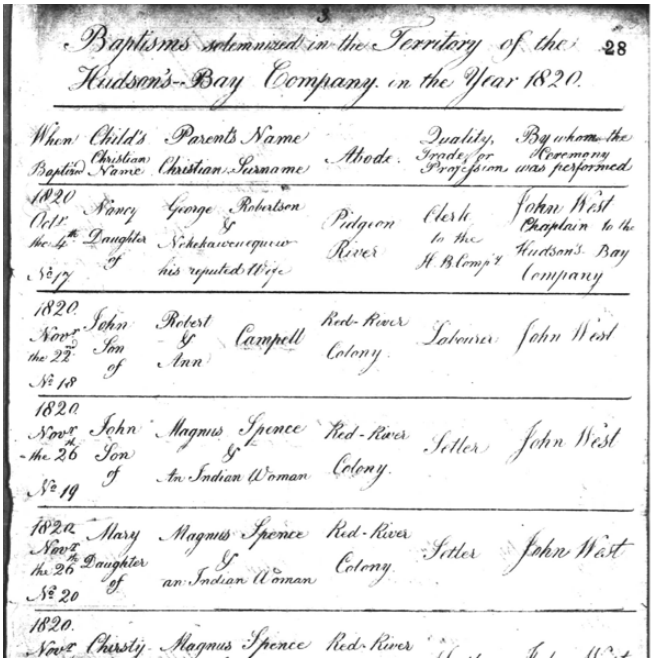
– 分析该语境下历史记录链接的具体要求。– 提出一种历史记录链接算法，即使数据记录仅包含姓名、日期和近似地点，也能实现高精确率。– 针对大型真实世界历史数据集，对该记录链接方法进行性能评估。

本文首先介绍DADP历史档案库的背景、记录链接要求，并将其与先前的记录链接工作进行比较。第3节介绍记录链接方法。第4节讨论实验结果。最后，本文总结未来工作和结论。

## 2 背景

数字档案数据库项目（https://dadp.ok.ubc.ca）是一个历史档案库，汇集了加拿大多个不同来源、涵盖1800年至1920年时期的各类历史文献。各数据源包括来自多个传教站和教堂的圣事记录（出生、婚姻、死亡）、人口普查信息、毛皮贸易记录以及航海者合同。每个数据源都通过转录原始文献图像，并对姓氏（姓）和地点等信息进行标注与标准化，整合到档案库中。图1展示了一份历史记录的示例图像。

原始来源文档图像，以及转录信息和元数据，都存储在档案库中。历史学家分析每条记录，提取姓名、地点以及记录中个人之间的关系等字段。尽管标注过程能够保持准确性并执行标准化，但数据质量和信息缺失问题仍然存在。



3

Baptisms solemnized in the Territory of the Hudson's-Bay Company in the Year 1820. 28

| When Baptized | Child's Christian Name | Parents Name Christian | Surname | Abode | Quality, Trade, or Profession | By whom the Ceremony was performed |
|---|---|---|---|---|---|---|
| 1820 Octr the 4th No 17 | Nancy Daughter of | George & Nekekawenequew his reputed Wife | Robertson | Pidgeon River | Clerk to the H. B. Compy | John West Chaplain to the Hudson's Bay Company |
| 1820 Novr the 22d No 18 | John Son of | Robert & Ann | Campell | Red-River Colony | Labourer | John West |
| 1820 Novr the 26 No 19 | John Son of | Magnus & An Indian Woman | Spence | Red-River Colony | Setler | John West |
| 1820 Novr the 26 No 20 | Mary Daughter of | Magnus & an Indian Woman | Spence | Red-River Colony | Setler | John West |
| 1820 Novr | Chirsty | Magnus | Spence | Red-River | | |

**图 1.** 历史文献示例

将数据加载到数据库后，历史记录中的每个人都会被添加到档案库中。例如，在一份圣事婚姻记录中，许多个人都会按姓名记录，包括丈夫、妻子、新郎的父母、妻子父母和见证人。某个个人可能会在不同时间的多条记录中被提及，即使历史学家掌握了额外的上下文信息，判断两条或多条记录中的姓名是否指向同一个人仍然具有挑战性。

通过姓名和其他元数据匹配数据档案中的个人，可以获得更完整的历史记录，并能够追踪个人随时间的变化。这一过程能够提升数据质量和用户体验，并建立记录之间的关系。这些关联带来了新的历史知识和见解，但人工操作十分繁琐。在本案例研究之前，DADP 尚未采用过记录链接流程，而所使用的方法会对档案库质量产生重大影响。

记录匹配技术广为人知，常用于在不同数据库中匹配客户和个人。如果匹配过程是在同一数据库中识别并消除相似记录，则称为去重。关于记录匹配和去重的综述 [3] 详细介绍了多种方法 [4,6]。历史数据集的主要区别在于可用于匹配的信息有限，通常只有姓名和日期。这限制了匹配方法的适用性和准确性。

针对历史数据库的记录匹配研究相对较少。概率方法 [1,2] 已用于匹配人口普查数据。这种方法利用字段相似度分数和期望最大化算法，来确定两条历史记录相匹配的可能性。对人口普查记录集合进行整体匹配时，该算法生成的匹配结果与人工匹配结果一致，适用于对趋势进行汇总性的历史分析。文献 [5] 也介绍了人口普查应用中的记录匹配。

这项研究的重点，是在各种历史档案中匹配个人，而不局限于人口普查数据。此外，研究旨在跨越时间与空间识别并追踪个人的历史事件和关系，而非进行大规模的汇总分析（例如人口和就业趋势如何随时间变化）。由于研究聚焦于个人及历史准确性，匹配时需要较高的精确率。将期望最大化算法应用于 DADP 中的人口普查数据集，成功生成了匹配结果，但精确率低于要求。圣礼数据集的结果则不尽如人意。概率记录匹配不适用于这些数据集，因为大多数记录只有姓名，其他信息也很有限。人口普查记录通常包含地点、年龄和出生日期，因此有更多匹配机会。研究尝试了通常用于匹配人口普查记录的统计方法，但认为这些方法效果不佳，主要原因是圣礼数据缺少更有力的身份识别数据，而这些数据在人口普查记录中较为常见，例如出生地和出生日期。

DADP 中的每条圣礼记录都记录了某一时间、某一地点发生的特定事件，以及参与其中的个人。例如，一场婚姻会涉及丈夫和妻子、家庭成员以及见证人。记录中会包含事件发生的日期和地点，以及相关个人的姓名。某个人可能仅凭姓名出现在多条记录中。考虑一位有几个女儿且女儿都已婚的父亲。这位父亲会作为“新娘的父亲”出现在多条婚姻记录中。除了姓名之外，还能利用关系角色，是匹配该档案库中数据所需功能的独特特征，也是本文所提出方法的关键特征。用于匹配个人的字段包括名字、姓氏、事件发生地点（例如举行婚礼或施洗的教堂）、事件的大致日期、个人在事件中的角色，以及事件中记录的其他人员。

总之，历史数据集中的记录匹配为增进历史知识和背景提供了机会，但也需要针对可用字段信息有限这一情况进行调整的特定方法。

## 3 记录匹配方法

记录匹配方法在主要由姓名、关系角色、事件日期和地点构成的数据集中，能够以较高的精确率生成匹配结果。使用关系角色（例如父亲、母亲、兄弟）可以提高准确性，因为两个家庭成员经常会一同出现在圣事记录中。

图2展示了婚姻记录中常见的关系角色示例。图中的节点代表婚姻记录中包含的一名人物，而边表示人物之间的联系。在这两条记录中，Adolphe Desroches 和 Elizabeth Langdon 均出现在婚姻记录中，并且两条记录中的关系角色都表明，他们是夫妻，也是新娘的父母。由于这类家庭二元关系经常出现在圣礼数据中，因此可以将这些信息用作判断两条记录是否匹配的有力指标。

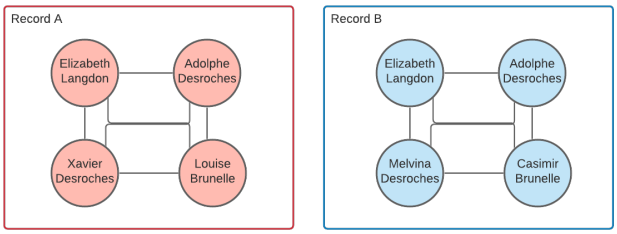



**图 2.** 利用关系进行记录匹配

记录匹配方法包括以下步骤:

– 清理与标准化 - 将姓名、地点和日期转换为标准化格式– 索引 - 根据关键属性（名和姓）对记录进行分组和索引– 匹配与分组 - 根据姓名、地点和关系角色匹配记录。关系角色用于将记录聚类成组。

图 3 概述了该算法。

### 3.1 记录索引

匹配算法的第一步是对记录进行索引和分组，以提高效率。面对数十万条记录，将每条记录与其他所有记录进行匹配并不高效。在匹配之前，会根据姓名相似度对记录进行分组和索引。只有姓名相似度足够高的记录才会进行比较。

DADP档案库的一项优势是，记录中的名和姓都经过标准化。这大大减少了姓名处理和索引方面的问题，使姓名索引具有很高的准确性。索引依据名和姓建立。随后，每个索引组都会按记录（事件）日期排序，以便高效分析。按日期排序有助于解决诸如比较同名父子记录之类的问题。

```
从数据库档案中检索记录 对姓名、日期和地点进行数据清
洗与标准化

按姓名建立索引并对记录分组

对于每个索引组：按日期对组内记录进行排序

    对于组内的每条记录：
        如果不在记录集中：
            使用该记录创建新的记录集

        对于组中的每条其他记录：如果姓名相似度高于阈值、日期
        记录在范围内、地点相似度匹配且关系一致，则将匹配记录
        添加到记录集
```

**图3**。 匹配实现伪代码

### 3.2 记录匹配

匹配过程会比较具有共同索引的记录，并将其分组为匹配记录集。该过程的输入是索引组集合。对于每个索引组，输出是若干匹配记录集。如果算法判定两条或更多记录相匹配，则它们属于同一集合。如果某个个体出现在多条记录中，那么所有这些记录都会被纳入该集合。采用这种基于集合的方法的一个优势是，随着更多包含额外信息的记录被添加到集合中，记录匹配的效果会得到改善。通过使用匹配记录集，可能会出现这样的情况：两条记录单独来看并不符合全部匹配标准，但结合集合中的其他记录后，却符合所有标准。这是因为其中一条记录可以充当另外两条记录之间的纽带，而后两条记录原本无法相匹配。例如，考虑三条记录 A、B 和 C：A 与 B 匹配，B 与 C 匹配，但 C 不与 A 匹配。由于 B 同时与 A 和 C 匹配，即使没有 B 时 A 和 C 无法匹配，A 和 C 仍会被纳入同一匹配记录集。

**姓名匹配** 姓名是记录匹配中最关键的要素，尤其是在圣礼数据中。名字和姓氏是最接近个人唯一标识符的属性。索引步骤按姓名对记录进行分组。匹配步骤则使用各种字符串比较算法进行相似度匹配，例如 Levenshtein 编辑距离或 Jaro-Winkler 距离。

**地点匹配** 地点和姓名一样，在记录方式上也可能存在差异。例如，一座教堂可能有不同的称呼，如“Rapids Church”也被称为“Grand Rapids Church”。地点比较可结合使用子串比较、编辑距离和 Damerau-Levenshtein 距离。

与包含个人出生地的人口普查数据不同，圣礼数据包含事件发生的地点。可以利用多个地点相似的记录，区分姓名相同但现居地点不同的人。结合日期，如果两个人在给定时间内无法在不同地点之间往返，就能准确判断他们是否为不同的人。

**时间匹配** 虽然圣礼数据中通常没有年龄或出生日期，但记录事件的时间是匹配时非常有价值的信息。例如，姓名“John Setter”在多个历史记录中多次出现，尤其是在 Rapids Church 的几份洗礼记录中。记录的日期可以作为判断两条记录是否匹配的依据。由于这些记录分别发生于 1847 年、1848 年和 1870 年，因此，相比相隔二十年的记录，时间间隔在一年以内的两条记录更有可能匹配。匹配算法可以配置记录之间的时间间隔，以确定哪些记录仍可视为匹配。在匹配 DADP 中的圣礼数据时，如果两条记录的时间间隔在五年以内，便符合匹配标准。这个数值是通过测试多个取值后选定的，以确定能够产生最高质量匹配结果的数值。请注意，对记录集进行匹配有助于满足时间匹配标准：两条记录单独来看可能不符合时间间隔匹配标准，但通过一条或多条关联记录，仍可归入同一匹配集。

**关系连接** 使用关系连接是一项能够提高准确性的独特贡献。利用记录中可获得的任何关系数据，是提升圣礼数据匹配质量的关键。在 DADP 中，夫妻等家庭二元关系经常共同出现在多份历史记录中，这是区分不同个体的重要方式。比较关系最简单的方法，是将一条记录中出现的所有人员视为图中的节点。对每个图的节点进行集合交集运算，可以揭示两条记录是否包含多个可能的匹配。由于错误匹配通常不会在图中共享多个节点，因此，当集合交集产生两个或更多共同点时，可以放心地认为存在家庭关系。

图 4 展示了如何根据两个来源记录中的共同节点进行匹配。对记录 A 和记录 B 求集合交集后，可以看到 Adolphe Desroches 与他的妻子 Elizabeth Langdon 同时出现在这两条记录中。对记录索引求集合交集

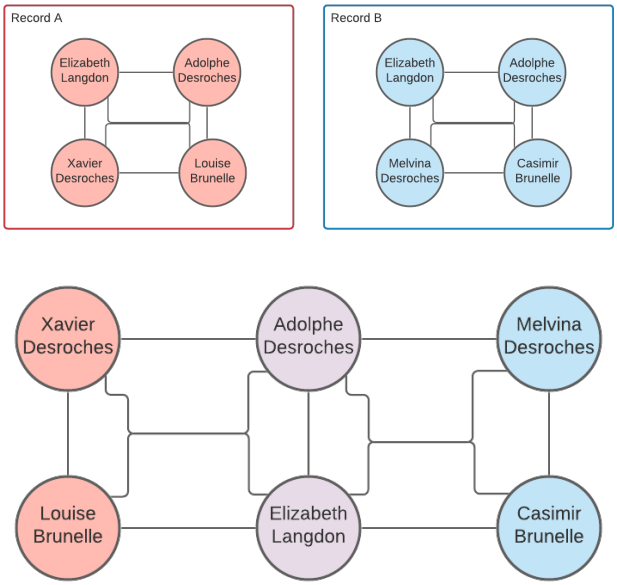



**图 4.** 利用关系进行记录匹配

通常会得到存在某种亲属关系的人物记录。在 Adolphe 和 Elizabeth 的示例中，这种关系是丈夫与妻子。

一旦确定一对记录中至少存在两个可能的匹配项，就必须确保其中的关系有效。在 DADP 的婚姻记录和洗礼记录中，常见的情况是，父亲和儿子因姓名相同且关系相似而被错误匹配。通过比较一个人在一条记录中的角色与其在另一条记录中的角色，可以避免这类错误匹配。例如，在一条记录中扮演丈夫角色的人，绝不应与几年后另一条记录中扮演父亲角色的人匹配。死亡记录也适用同样的原则：1870 年下葬的人绝不应与 1870 年之后的任何人匹配。每个数据库都有特定的验证规则，可根据关系背景和时间进行编码，以排除不可能的匹配。

## 4 案例研究与实验评估

该算法在 DADP 圣事数据库中的三个数据源上进行了测试。由于真实匹配和真实非匹配的集合未知，因此无法使用 F分数和召回率等常见指标来衡量准确性。因此，有必要采用其他方法来确定该算法的有效性。部分结果

经过人工核验，以确定每张表中的一组记录是真匹配还是误匹配。性能结果见图5。

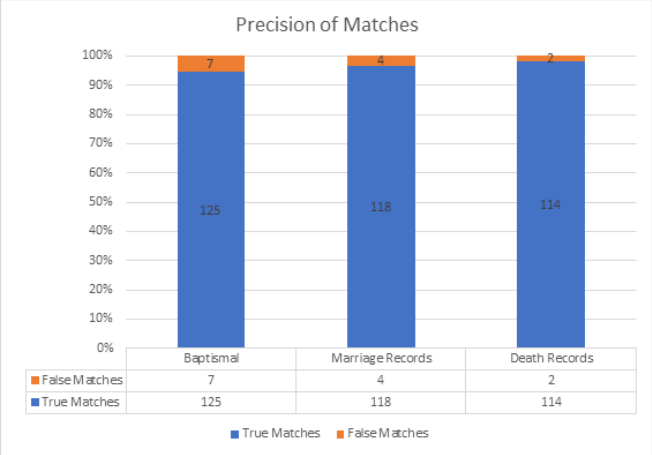



**图5**。 匹配结果

匹配精确率非常高（超过95%），适合在DADP中投入生产使用。算法最常见的错误是错误地匹配家庭成员，具体而言，就是将父亲及其儿子匹配在一起，因为他们拥有相同的姓名。使用关系角色可以显著减少这些错误，但错误仍然存在。即使是历史学家，如果没有额外的历史背景，仅凭记录中的信息，也很难消除这些指代的歧义。

这种记录匹配方法的优势在于高度灵活，并且可以适用于各种各样的数据。图5展示了将该算法应用于三种不同数据记录类型的结果：洗礼记录、婚姻记录和死亡记录。这些记录具有不同的格式和数据字段。使用关系角色可以提高准确率，并充分利用圣礼记录中可获得的独特信息。

一个有待改进的方面是处理许多个人共用同一姓名的档案。有关关系角色的更多信息和规则，或许能进一步帮助解决个人之间的差异。例如，三条关于 Joseph Allery 的记录都源自同一份婚姻记录，但各自具有不同的角色：第一见证人、丈夫的父亲和丈夫。必须采取某种措施，防止将父亲和见证人与正在结婚的 Joseph Allery 匹配起来，因为尽管他们拥有所有共同细节，但显然并非指同一个人。针对这种情况，一项建议是建立一个角色层级，按照这些角色在数据中产生真实匹配的可能性从高到低排列。此外，婚姻记录的匹配算法还应增加额外检查，确保正在结婚的人永远不会与父亲或母亲匹配。归根结底，这些问题将取决于

正在匹配的数据类型，并且应在适当情况下针对具体情形制定解决方案。总体而言，该案例研究展示了所生成的匹配结果，这些结果提升了 DADP 对历史学家的质量和历史价值。

## 5 结论与未来工作

与典型数据库中常见的数据集匹配相比，历史记录匹配和去重面临着独特的挑战。记录匹配算法必须处理数量有限的数据字段（通常只有姓名），同时仍要实现高精确率。本研究介绍了一个真实案例研究，将现有匹配方法扩展至纳入关系角色和家庭关系数据，以提高匹配准确率。通过运用基于这些关系、时间和地点的规则，生成的匹配结果中有 95% 为真实匹配。这些匹配结果用于提升 DADP 档案库的质量，并使历史学家能够跨越时间和空间，直观呈现个人的历史事件和关系。利用关系角色提高匹配准确率的方法和技术可推广至许多其他历史数据集和应用。未来的工作将进一步评估该算法在此数据集上的精确率和召回率，这需要历史学家为数据集标注；同时还将探索减少所需人工工作量的匹配技术。

## 6 致谢

作者谨此感谢加拿大原住民事务与北方发展部对本研究的支持。

## 参考文献

1. Abramitzky, R., Boustan, L.P., Eriksson, K., Feigenbaum, J.J., P´erez, S.：历史数据的自动链接。技术报告，美国国家经济研究局（2019） 2. Abramitzky, R., Mill, R., P´erez, S.：跨历史资料链接个人：一种全自动方法。《历史方法：定量与跨学科历史学期刊》**53**(2), 94–111（2020） 3. Christen, P.：面向可扩展记录链接与去重的索引技术综述。IEEE 知识与数据工程汇刊**24**(9), 1537–1555（2011） 4. Christen, P.：数据匹配——记录链接、实体消解与重复检测的概念和技术。《数据中心系统与应用》。施普林格（2012）。DOI 10.1007/978-3-642-31164-2。URL https://doi.org/10.1007/978-3-642-31164-2 5. Ruggles, S., Fitch, C.A., Roberts, E.：历史人口普查记录链接。《社会学年鉴》**44**, 19–37（2018） 6. Winkler, W.E.：匹配与记录链接。《威利跨学科评论：计算统计》**6**(5), 313–325（2014）