中国历代人物传记资料库（CBDB）对历史网络的结构化处理、记录与分析①

傅君劢 / 加利福尼亚大学尔湾分校东亚语言与文学系

王宏苏 / 中国历代人物传记资料库项目经理

摘　要：文章对CBDB的发展概况和数据结构予以简要说明，并着重介绍CBDB在历史网络研究方面的价值。通过从历史文献中提取越来越多的数据，CBDB对网络研究的价值正在得到进一步增强。中国历代人物传记资料库（CBDB）是关系型数据库。截至2020年5月，CBDB共收录了47万名中国古代人物的生平信息。CBDB的独特之处在于，用户可从数据库中的任何人物或群体出发，生成这些人物的亲属关系网络和社会关系网络。CBDB建立之初，我们曾尝试借鉴群体传记学领域中其他数据库项目的开发模式。在这个过程中，我们发现我们正在开发的数据模型与其他数字化群体传记学项目截然不同：CBDB侧重数据分析，致力于从大量中国历史文献中批量提取传记数据并对其进行规范化处理。文章将介绍CBDB的发展概况和数据结构，并着重介绍CBDB在探索亲属关系和社会关系网络方面的价值。通过从历史文献中提取越来越多的数据，CBDB对网络研究的价值正日益增强。

关键词：中国历代人物传记资料库　社会网络分析　群体传记学　关系型数据库

①本文英文版刊载于《历史网络研究》学刊2021年第5期，见Michael Fuller, Hongsu Wang, "Structuring, Recording, and Analyzing Historical Networks in the China Biographical Database," *Journal of Historical Network Research*, vol. 5, 2021, pp. 248-270. 我们特别感谢中国历代人物传记资料库项目指导委员会委员陈松教授在中英文版撰写中的深入讨论、修改与编辑。考虑到中文的表述习惯，本文以意译为主。文中第四节提到的Gephi与NetworkX数据与代码均公开于https://github.com/cbdb-project/Developing-Historical-Networks-in-CBDB-JHNR/tree/master/section%204。

引　言

郝若贝教授（Robert M. Hartwell，1932—1996）的中国历史研究数据库（Chinese Historical Studies，缩写为CHS）是中国历代人物传记资料库（下文简称CBDB）的前身。当时郝若贝使用中国历史研究数据库记录宋代官僚的亲属关系和社会关系，但他并没有对这些数据背后的社会网络进一步探索。当傅君劢重制郝若贝数据库时，他设计了递归检索。这一功能首先应用于亲属关系查询，随后推广到社会关系查询中，用于生成可进行量化分析的社会关系和亲属关系网络。

我们认识到递归检索的力量后，也意识到了对亲属关系和社会关系数据进行标准化的重要性。一开始，数据收集者希望尽可能保留文本的原始面貌，但这种做法很容易让数据库中对社会关系的描述一再重复、纷繁芜杂。因此，我们决定建立一套系统性规范来管理从文献中提取的种种社会关系。同样，文献中对亲属关系的描述也给我们带来类似问题，让亲属关系分析变得非常困难。

对于这些问题，我们的解决方案是，对具体的社会关系进行多层分类。学者在探索社会关系网络时，可灵活决定使用笼统或具体的社会关系和亲属关系类别。另一方面，我们设计了一个对检索结果中的社会网络大小进行限制的条件，学者可自行调整从检索起点开始到生成社会关系网络中的其他节点的最远距离。

在亲属关系查询中，我们从世代（祖先与后代）、同辈（兄弟姐妹）、婚姻三个维度对亲属关系进行定义（例如，父亲的兄弟的妻子与兄弟的妻子的父亲在这三个维度上的参数值完全相同）。各种亲属关系都可以通过这些参数的组合进行描述。学者可通过指定这些参数的上限来设置生成亲属关系网络的大小。

生成社会关系或亲属关系网络后，学者便可将相关数据导入常用的社会网络分析软件中做进一步研究。更方便的是，CBDB是关系型数据库，关系网络中的人物在CBDB中可以关联到大量的属性数据。将人物属性数据与社会网络分析相结合，学者可以挖掘到更深的问题。比如，最常见的属性数据是与人物生平活动有关的地址信息（如籍贯、任官地、葬地等）。在社会网络分析软件中，可以利用这些地址信息把表达人物之间关系的网络转换为表达地址之间关系的网络。此外，CBDB也可以直接将关系网络数据输出到常用的地理信息系统（GIS）分析软件中。

CBDB是关系型数据库，通过设置各种条件，学者可以很方便地选取自己感兴趣的群体，进而探索他们的亲属关系和社会交往关系，如，宋代、明代、清

代医生的关系网，或是四川出身的进士登科人物的关系网等。经过长期的发展，CBDB为历史网络研究提供了强大的工具。通过灵活的参数设置，学者可从47万名中国古代人物的传记数据中生成符合自己研究需求的关系网络，并对其进行多维度探索。

郝若贝在1991年《基于计算机的中古中国社会经济史综合分析》一文中介绍了他的中国历史研究数据库项目，阐述了这一雄心勃勃的项目期待达成的目标：

> 在过去的20年里，尽管对中国中古时期的研究大量涌现，但这些研究对于社会、经济、政治和思想等多种变量之间的复杂关系仍缺乏全面的认识。然而恰是在这些变量的共同作用下，才造就了中国社会的结构和特质。在上述研究中，学者们对于从汉代到明初的诸多课题提出了自己的观点，其中不乏构思精密的专著。但它们往往各擅胜场，互不相顾，有的研究特定地域、特定的家族世系，有的研究任职于某个政府机构的官员、某个政府机构的职能，有的研究某个行业的发展、某些财政政策的性质。然而，以上这些研究中（既包括中国和日本学者的研究，也包括西方学者的研究）并没有（也不可能）提供对各个地方、各个政府机构之间的异同之处及其在特定历史时期的嬗变情况加以分析所必须的定量数据，这些研究中所提出的观点因而难以检验。本文所描述的项目旨在为满足这一研究需求建立必要的数据库，并在此基础上为日后的中国古代史研究构建一个可与欧洲史、美国史相媲美的研究框架。①

1996年郝若贝去世后，他将中国历史研究数据库捐赠给哈佛燕京学社。这个数据库经过结构上的调整后，迁移到新平台，成为了中国历代人物传记资料库。在当时的项目经理陈松及其继任者们的协助下，包弼德（Peter K. Bol）与来自中国和日本的学术同行共同合作，对中国历代人物传记资料库进行了新一轮开发。

在数据库的建设过程中，我们给CBDB添加了根据查询条件生成学者所需历史关系网络的功能。CBDB的这一特色在很大程度上得益于它所使用的计算机语言。具体说来，郝若贝使用dBase建立了中国历史研究数据库，我们将数据库移

①Robert M. Hartwell, “A Computer-Based Comprehensive Analysis of Medieval Chinese Social and Economic History,” *Characters and Computers*, eds. Victor H. Mair, Yongquan Liu, Amsterdam: IOS Press, 1991, p. 89.

植到MS Access中，并使用Visual Basic开发数据库的分析功能。在这些分析功能中最值得一提的是，我们将用户查询到的亲属或社会关系人作为下次查询的条件，不断自动循环，直到循环次数达到用户设置的上限为止。这便是我们生成亲属或社会关系网络的过程。与此同时，我们意识到将这些网络与CBDB中有关其他命名实体（named entities，如入仕途径、任官资料、籍贯地等）的数据相结合会给历史研究带来巨大帮助。于是，我们一方面致力于改进数据结构和查询算法来生成满足历史研究需求的亲属关系和社会关系网络，为更好地理解这些网络提供必要的历史语境。另一方面我们也持续不断地收集与整理生成这些网络所需的人际关系数据。

一、作为古代中国社会关系模型的中国历代人物传记资料库

从郝若贝开始，CBDB的建设者们不断对这个关系型数据库进行设计和重构，不断完善CBDB的数据模型，以便更准确地对中国古代社会中形塑个人生活的核心要素进行抽象和建模。虽然CBDB源自郝若贝的中国历史研究数据库，但我们将实体—关系模型（entity-relationship model）运用到对史料的处理中，以此扩展他最初的设计。我们定义了9种基本的数据类别，并围绕它们建立CBDB。因此，当前CBDB的实体—关系模型包含了九种基本实体及其相互关系：

（1）人物（数据库的核心实体，其他所有实体皆与之相连）

（2）地址（除了人物之外，其他很多实体也会具有地址属性）

（3）亲属关系

（4）社会关系

（5）职官

（6）入仕途径

（7）社会机构（包括书院、寺庙、道观等）

（8）社会身份

（9）文本（文本常常是社会关系的载体，譬如书信往来、诗文赠答、为特定的场合撰写的诗文碑记都体现了当事人之间的社会关系）

人物和其他实体之间的关系是一对多关系。

中国历代人物传记资料库使用代码表和数据表将图1中的数据模型转换为数据库格式。代码表（code tables，即图1中的蓝色方框）用于记录九个基本实体的属性数据，而数据表（data tables，即图1中的黄色方框）则用于记录实体间的关系，将不同实体连接在一起。

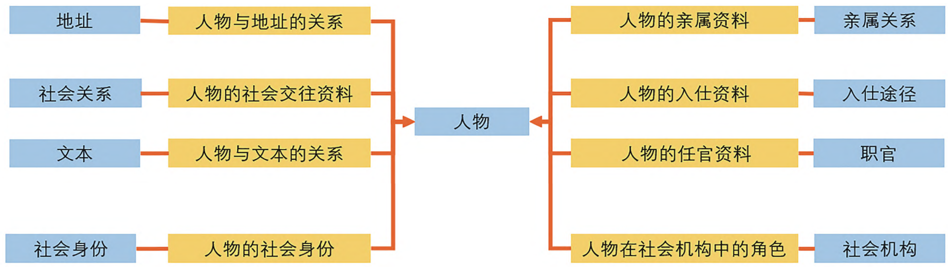

图 1　CBDB 中最基本的九种实体及其关系

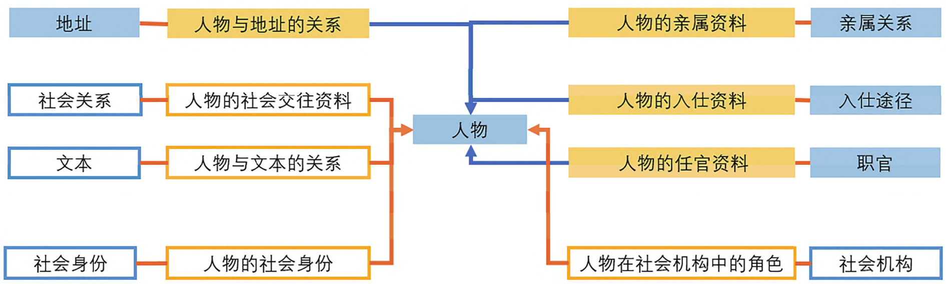

图 2　地址、入仕、任官以及关系网络四种实体之间的互动

古代中国的人际关系网络与形塑了当时社会生活的多种结构性因素相互作用，并受到这些因素的深刻影响。建设CBDB的目标之一便是方便研究者梳理人际关系网与这些因素之间的相互作用。

二、中国历代人物传记资料库的查询功能

中国历代人物传记资料库的优势既在于它的数据结构，也在于其不断增加的数据量。由于CBDB的数据结构非常复杂，傅君劢开发了一系列查询窗体，以方便用户使用。另一方面，当用户熟悉了CBDB的数据结构后，也可跳过这些查询窗体，直接使用MS Access软件的“查询设计”功能根据自己的需要进行更复杂的SQL查询，或使用任何程序语言从SQLite格式的CBDB数据库中直接读取和分析数据。

1．“按入仕途径查询”和“官职查询”窗体

通过“按入仕途径查询”和“官职查询”这两个窗体，学者可以查到符合某种入仕或职官条件的人物群体，并可将查询结果导出，再进一步生成此群体的亲属关系或社会关系网络。起家入仕是中国古代精英子弟的人生大事。从宋代开

始，科举成为最受推崇的入仕途径。举子往往要花费数年乃至数十年的时间进行准备，才能在科举考试中取得成功，不少举子更是在多次落第之后才得以登科。金榜题名不仅让举子获得了做官的资格，也把他们带入新的人际关系网中，使他们可以与同年登第的其他举子、当年的考官，以及家乡有功名的家庭建立联系。除了科举考试，还有其他一些入仕途径，比如荫补，即通过自己亲属的特权获得官职。这便引出了一个重要的历史学问题：以不同方法步入仕途的官员在婚姻模式和社会交往模式上是否有所不同呢？使用“按入仕途径查询”和“官职查询”这两个窗体，可以为探究入仕、任官、社会关系、亲属关系等诸因素的相关性快速搜集到所需的数据。

“按入仕途径查询”或许是CBDB中最简单的查询窗体。学者只需指定一个或一类入仕方式即可进行查询。学者也可以从时间（入仕年份或“指数年”）和空间（入仕地址或“指数地址”）两个方面对查询结果进行限制。

“指数年”和“指数地址”是CBDB特有的概念。CBDB是历史数据库，因此我们希望尽可能将每个人物与具体的时间（年份）关联起来。“指数年”就是用来建立这种关联的，这个想法始于郝若贝的中国历史研究数据库。郝若贝对宋代高层官员非常感兴趣，因此他特别留心宋代官员60岁前后的情况。我们继承了这一思路，并开发了指数年算法为史料中不曾记录生卒年份的人物计算指数年，并设置了各种推算方法的优先级，以求尽可能准确地进行估算。CBDB的早期版本沿用了郝若贝对指数年的定义，即以一个人物年届60的那一年为指数年，如其卒于60岁之前，则以卒年为指数年。但是在最新的版本中，我们修改了指数年的定义，统一以历史人物的出生年份为指数年。[①]除了时间信息，地理信息在中国群体传记学研究中也非常重要，所以我们用类似于“指数年”的概念设置了“指数地址”，利用与历史人物有关的一系列地址信息（如本贯、徙居地、祖籍、葬地等）按照一定优先级来指定“指数地址”。在最新版本的CBDB中，用户可以按照自己的需要对指数地算法所使用的各种地址信息的优先级加以调整。[②]

设置完入仕方式与其他参数（如上文提到的“指数年”“指数地址”）后，便可进行查询。图3显示了1130年到1160年间3,285位登科举子的信息。在查询结

① 本文中的截屏图片统一使用2022年初最新版的CBDB。因此，这些截屏和英文版中的截屏不同，反映了修改后的指数年定义、最新的用户界面和数据内容。

② 这里的“地址”指行政单位（譬如州或县），其名称、治所、辖区、统属关系因时而异。因此CBDB在涉及地址查询的窗体中提供了两种方式来处理上述问题。第一，窗体上有复选框，学者可以选择是否要将待查询地址的所有下级地址（如要查询的州级地址所辖的各个县级地址）包括在内进行查询。第二，根据待查询地址（及其下级地址）治所的地理坐标找出指定历史时段内所有具有相同坐标的地址（不论其地名如何变化），并将所有这些地址作为查询条件。

果中，CBDB列出了每位人物的各种属性数据，如指数地址、入仕年份等。理论上说，研究者可将这些数据作为双模网络数据加以分析。另外，学者亦可将此查询结果导出，再使用地理信息系统（GIS）软件在地图上分析这些人物的地理分布。如果希望探索与这些人物相关的关系网，学者可点击“存储人物代码”，然后在CBDB的各种人际关系查询窗体（如“查询亲属关系”和“查询社会关系网络”窗体）中点击“召回人物代码”，进而查询与这些人物相关的关系网络数据。

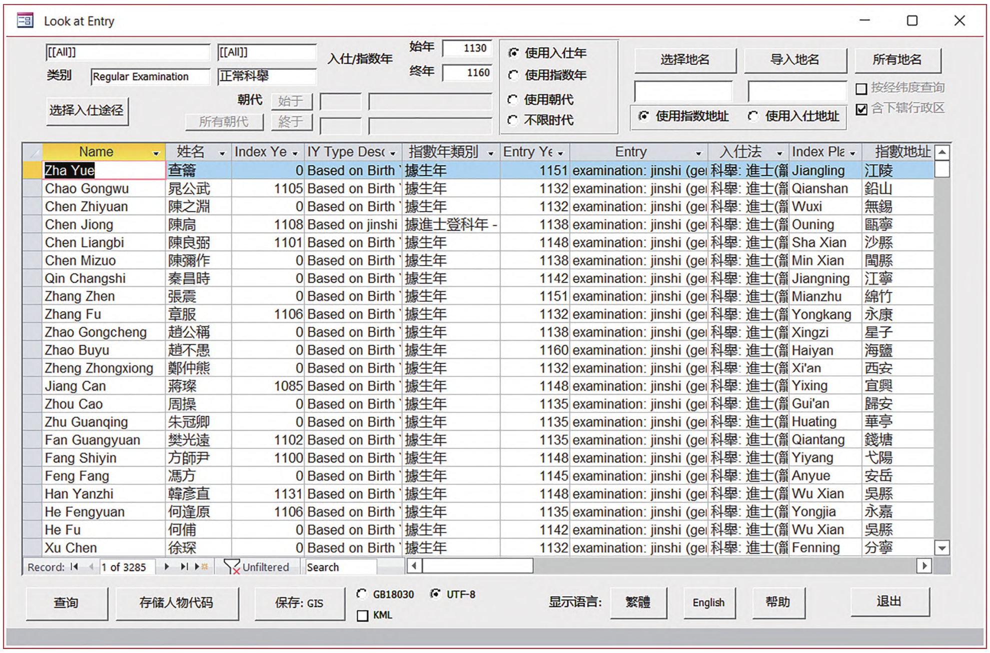

图3　使用“按入仕途径查询”查找在南宋早期登科的举子

“官职查询”与“按入仕途径查询”非常相似。中国古代职官制度非常复杂，因此CBDB为学者提供了灵活的查询方式。譬如，学者可以指定具体的官称进行查询，也可通过将职官系统中的某个级别或门类下的所有官职作为查询条件。此外，学者还可以通过设置指数年、朝代、指数地址和任官地址等条件对查询结果加以限制。

图4是对南宋县令的查询结果。此例中，县令的各种属性数据与“按入仕途径查询”的检索结果一样，可用于构建双模网络，亦可用于生成这些人物的亲属与社会关系网。例如，靖康之变后，北方士大夫纷纷逃往南方，他们将如何在南方立足呢？具体来说，南宋初年从北方避乱南迁的县令是否与其辖区内的南方精英家族建立了婚姻关系或其他社会关系呢？使用“官职查询”可以非常容易地获得一个北宋晚期地方官员的名单，学者可从GIS的角度分析其家乡和任官地的分

布情况，亦可通过“查询亲属关系”和“查询社会关系网络”等窗体来探讨他们的人际关系网。

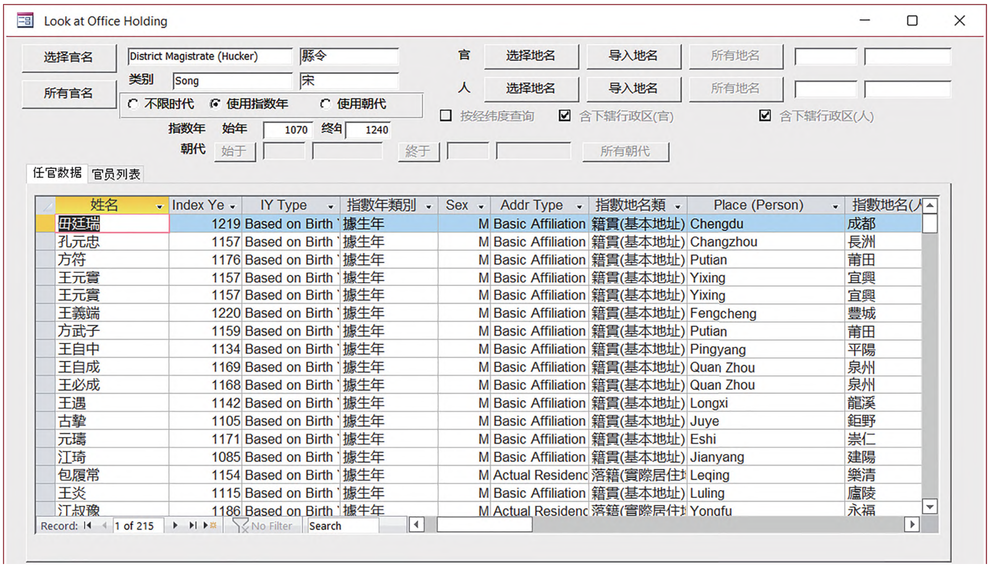

图4　使用“官职查询”窗体查找南宋时期的县令

2.“查询社会关系”窗体

截至2020年5月，CBDB收录了472,090位历史人物和149,611条社会关系记录。学者可以使用CBDB的“查询社会关系”窗体，按照不同性质的社会关系对历史人物进行简单的分组。

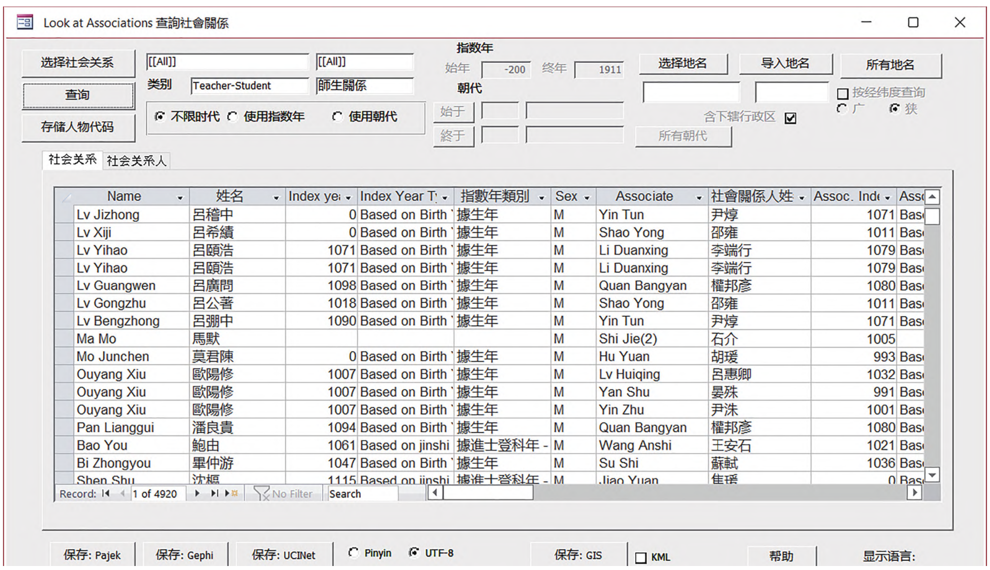

图5　CBDB“查询社会关系”窗体中所有师生关系

如图5所示，CBDB收录了大量师生关系数据。面对这些数据，我们可以提出许多问题。比如，这些师生关系是否构成了许多相对独立的区域性师生关系网，还是超越了地域限制形成了全国性的大网络？这些网络是从什么时候开始出现在史料中的？它们的规模和地理分布是否随时间而变化？在“查询社会关系”窗体中，学者既可以笼统地将一大类社会关系（如“学术关系”）设置为查询条件，也可以从较具体的社会关系分类（如“学术关系”中的“师生关系”）出发进行检索。学者还可通过指定地点或时间对查询范围加以限制。检索结果即可导出为数据表，也可导出为一些专业分析软件（如Gephi、UCINet、Pajek、ArcGIS、QGIS）可以识读的格式，用于社会网络分析和地理信息系统分析。此外，学者亦可在“查询社会关系”窗体中将查询结果存储下来，将其作为其他查询窗体中的查询条件进行进一步探索。例如，成为某人的学生或成为某些人的同门意味着什么？这些学生、老师及其家庭成员是否通过亲属关系或其他形式的社会关系结为一体？如需利用CBDB数据探讨这一问题，用户可以使用“存储人物代码”将师生关系检索结果中涉及到的人物代码（Person ID）保存起来，然后在“查询亲属关系”窗体中点击“召回人物代码”将这些人物代码设置为查询条件，便可找出这些人物的亲属关系数据。

3.“查询亲属关系”窗体

截至2020年5月，CBDB收录了482,979条亲属关系。由于中国古代精英人物的亲属关系十分复杂，在使用CBDB数据动态地生成亲属关系网络时，如何对亲属关系网的规模加以限制便成了难点。

考虑到亲属关系的结构以及史料中各种亲属关系描述的详细程度，CBDB当前定义了479种亲属关系代码。在“查询亲属关系”窗体中，我们以这些代码为基础设计了四个简单明了的参数，用于限制检索过程中所生成的亲属网络的范围。这四个参数分别是先世值（F[父亲]、M[母亲]）、后世值（S[儿子]、D[女儿]）、旁系值（B[兄弟]、Z[姐妹]）、姻亲值（H[丈夫]、W[妻子]）。这四个参数值只在逻辑上对亲属关系网的范围进行界定，用作亲属关系检索算法的终止条件。由于这些参数值是抽离了特定历史语境的，因此学者往往要根据自己的研究需要使用更严格的条件对检索结果做进一步筛选。举个例子，史料中许多表达亲属关系的词汇具有多重含义，如表弟既可以是父亲或母亲的姊妹的儿子，也可以是母亲的兄弟的儿子。在亲属关系代码中，CBDB会将这三种“表弟”分别定义为FZS-、MZS-和MBS-。在CBDB的亲属关系算法中，这三种关系和堂兄弟关系（即父亲的兄弟的儿子）具有完全相同的参数值（先世值、后世值、旁系值皆为1，而姻亲值为0），但是对身处不同社会文化环境中的历史当事人而言，这些关系的亲

疏远近则未必完全相同。换言之，CBDB的四个参数值是偏向客观中立的，是独立于特定历史时期的社会文化价值的，因此学者或有必要根据自己的研究课题对亲属关系的检索结果加以调整。再如，一个人的兄弟姐妹可能是父母的养子女或私生子女，可能跟自己同父异母或者同母异父。尽管这些细微差别在CBDB的亲属关系代码中都得到了保留，但是在查询亲属关系时，我们的算法对这些关系的亲疏远近则不做区分，因此学者有必要按照自己的研究需要对亲属关系的查询结果进行二次筛选。综上所述，在建设亲属关系数据时，我们利用英文字母（如F指父亲，M指母亲）描述基本的亲属关系，再用这些英文字母的组合表示常用的亲属关系（如FF指祖父，MM指外祖母）。使用这种亲属关系表达法，CBDB不但在数据记录的过程中对自然语言中的亲属关系用语进行消歧，而且在递归检索时，也可以通过将这些表示亲属关系的英文字母连缀起来进而表达更加复杂的亲属关系，再据此计算这些关系在四个维度上的参数值，以此衡量这些关系的亲疏远近，对亲属关系的查询范围予以限定，从而生成符合学者需要的亲属关系网络。

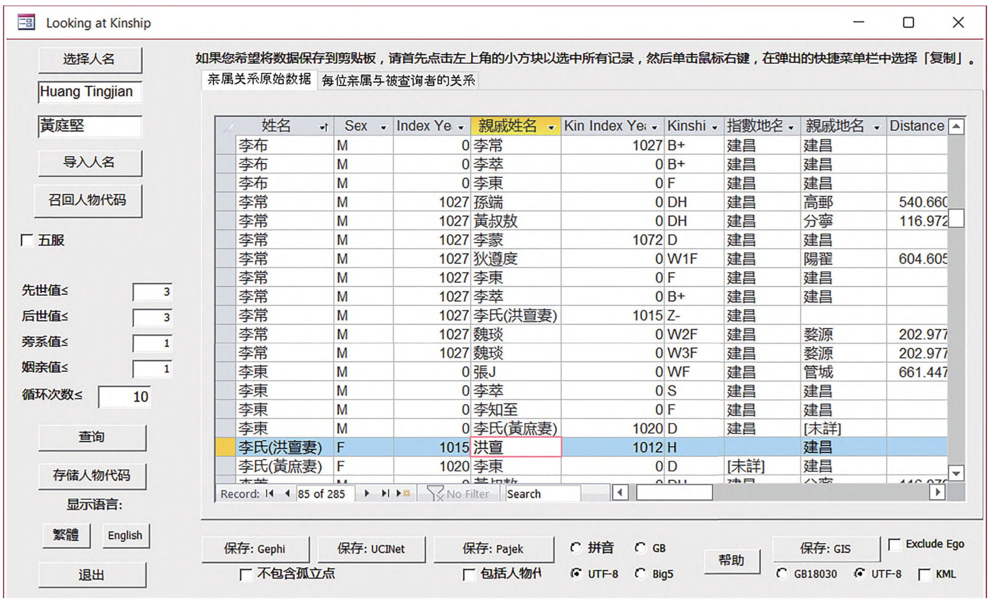

图6　使用“查询亲属关系”窗体查询黄庭坚的亲属（输出结果为CBDB中的原始数据）

图6显示了对黄庭坚亲属关系的查询结果，查询条件为先世值=3，后世值=3，旁系值=1，姻亲值=1。如果在“按人查询”窗体中检索黄庭坚，会发现只有32条亲属关系直接登记在黄庭坚的资料项下，而使用“查询亲属关系”窗体则可以找到294条符合查询条件的结果。这是因为后者使用了递归算法，即在找

到直接登记在查询对象资料项下的亲属后，计算机将继续搜索这些亲属的亲属，如此数次。CBDB的这种递归检索功能有助于学者发现史料中未曾直接记录的亲属关系，而且学者也可以通过改变查询条件来探索符合自己研究需求的亲属网络。在查询亲属关系时，我们建议用户首先生成一张庞大的亲属关系网，再根据需要删除其中的冗余信息，这样做比一开始就设置严苛的查询条件因而遗漏掉重要的亲属关系更加保险。

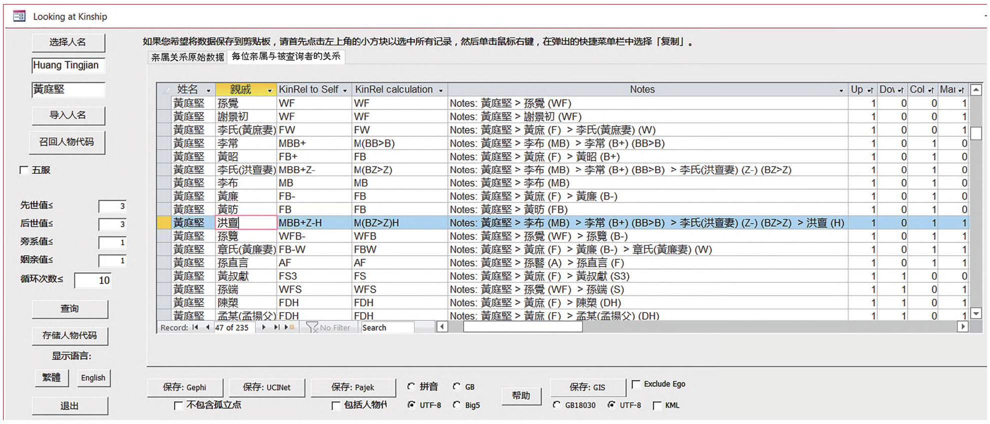

图7　使用“查询亲属关系”窗体查询黄庭坚的亲属

（通过将CBDB中的原始数据进行汇整，在输出结果中呈现每位亲属与黄庭坚的关系）

在查询指定人物的亲属关系时，CBDB会将查找到的原始数据进行汇整，将原始数据中表示亲属关系的英文字母连缀起来，用于表达检索结果中每位亲属与指定人物之间的具体关系。在此过程中，CBDB的亲属关系查询算法会对八种字母组合进行自动简化：BZ（兄弟的姐妹）简化为Z（姐妹），BB（兄弟的兄弟）简化为B（兄弟），ZZ（姐妹的姐妹）简化为Z（姐妹），ZB（姐妹的兄弟）简化为B（兄弟），SB（儿子的兄弟）简化为S（儿子），SZ（儿子的姐妹）简化为D（女儿），DB（女儿的兄弟）简化为S（儿子），DZ（女儿的姐妹）简化为D（女儿）。在图7所示的例子中，查询的第一步是根据直接记录在黄庭坚名下的亲属资料，发现李布是黄庭坚的母舅（MB，母亲的兄弟）。接下来，计算机进一步查找李布的亲属关系，并发现李常是李布的哥哥（B+）。之后，算法会将这两条关系缀合在一起，得出李常是黄庭坚的母亲的兄弟的哥哥（MBB+）并将这个关系简化为MB，从而得出李常是黄庭坚母亲的另一个兄弟（MB，母舅）。

在上例中可以发现，这样的简化会使亲属关系的旁系值减少1，这会增加满足当前查询条件的亲属关系人数量，于是更多的亲属关系人就会被添加到系统

生成的亲属关系网中。从下例中可以看出这种简化对查询结果的影响：洪亶之妻李氏是黄庭坚母亲的妹妹（MBB+Z- > MZ），但这条亲属关系并没有直接记录在资料库中，而是算法通过李布（MB）和李布的哥哥李常（MBB+ > MB）发现的。若不进行简化，李氏和黄庭坚的关系将表述为MBB+Z-，那么李氏就不符合查询条件中对旁系值为1的设定，也不会出现在检索结果中。当这条关系简化为MZ后，不但李氏符合设定的查询条件，而且从李氏出发又可以发现她的丈夫洪亶，进而可知洪亶为黄庭坚之姨父（MBB+Z-H > MZH，母亲的姐妹的丈夫）。我们相信，如果把亲属关系简化算法扩展到同辈之外的其他亲属关系上，将进一步增强“查询亲属关系”窗体的功能，是极有价值的探索方向。

4.“查询社会关系网络”窗体

“查询社会关系网络”窗体是使用CBDB数据探索历史中社会网络的主要界面。如图8所示，从界面上看，该窗体比“查询亲属关系”复杂得多。首先，它允许用户通过设置历史人物的地址与时间信息对查询范围加以限定。其次，它提供了各种各样的社会关系类型供学者选择。再次，该窗体还提供了将社会关系和亲属关系混合在一起进行查询的功能。尽管如此，控制社会关系网络的大小要比亲属关系网络简单得多：学者只需设定社会关系人与作为查询起点的指定人物之间的最大距离即可。

图8 “查询社会关系网络”窗体

“查询社会关系网络”向用户汇报三种数据（见图8）：即“社会关系”选项卡中的社会关系数据（边数据）、“社会关系人”选项卡中的社会关系人及其属性数据（节点及其属性数据），以及“合并多重社会关系”选项卡中汇整后的边数据（当两人之间存在多重关系时，将多重关系合并为一条记录汇报于这个选项卡下）。使用“保存：GIS”功能可将检索结果中所有人物的地理信息导出为ArcGIS与QGIS可识读的格式。使用“保存：UCINet”和“保存：Gephi”可导出“社会网络”选项卡中的边数据，而“保存：Pajek”导出的则是“综合多种社会关系”选项卡中汇整后的边数据。

为了让“查询社会关系网络”窗体更便于使用，我们还设计了一些特别的功能。首先，因为缺乏足够的信息，我们无法为许多历史人物推算指数年，但我们至少知道他们生活的朝代。因此，新版CBDB允许用户使用人物生活的“朝代”对查询范围加以限定。其次，一些使用者可能没有足够的历史地理学知识，不熟悉历史地名和管辖区域的嬗变情况，因此我们提供了一项新的功能：在用户设定一个或多个地址后，CBDB会找出历史上与指定地址具有相同经纬度的其他所有地址，并将这些地址一并纳入查询范围。再次，学者可以选择是否要将亲属关系混入社会关系中进行查询。在社会关系和亲属关系混合查询的设计中，用户可以自行选择是否使用“查询亲属关系”窗体中的四个参数（先世值、后世值、旁系值、姻亲值）对亲属关系的距离加以限定。最后，学者可将在其他查询窗体（社会区分、官职、入仕途径、地区关系等）中检索到的人物代码保存下来，然后在“查询社会关系网络”中载入这些人物，进一步对这些人物的社会关系网络进行查询。反过来，学者亦可将“查询社会关系网络”中的人物保存下来，在其他窗体中作为查询条件载入。

5.“查询两人之间社会关系”与“查询地区关系”窗体

为了便于进行群体传记学研究，CBDB还为社会网络分析设计了另外两个查询窗体。第一个是“查询两人之间社会关系”。在这个窗体中，学者可以输入两个人物或导入一个人物列表，然后查询这些人物相互之间直接或间接的社会联系。在图9的例子中可以看到，虽然苏轼与程颐相互攻击，但仍有一些扮演了桥梁角色的中间人将两人连接起来。本例使用指数年对人物生活的时段做了限制，从而查询结果中的人物大致生活在同一时代。检索结果显示，在这个将苏轼和程颐连接起来的社会网络中共涉及13人和107条社会关系。这些人物组成了一个规模可观的群体，该群体为苏轼和程颐这两个持不同文化理念的人物提供了沟通渠道。

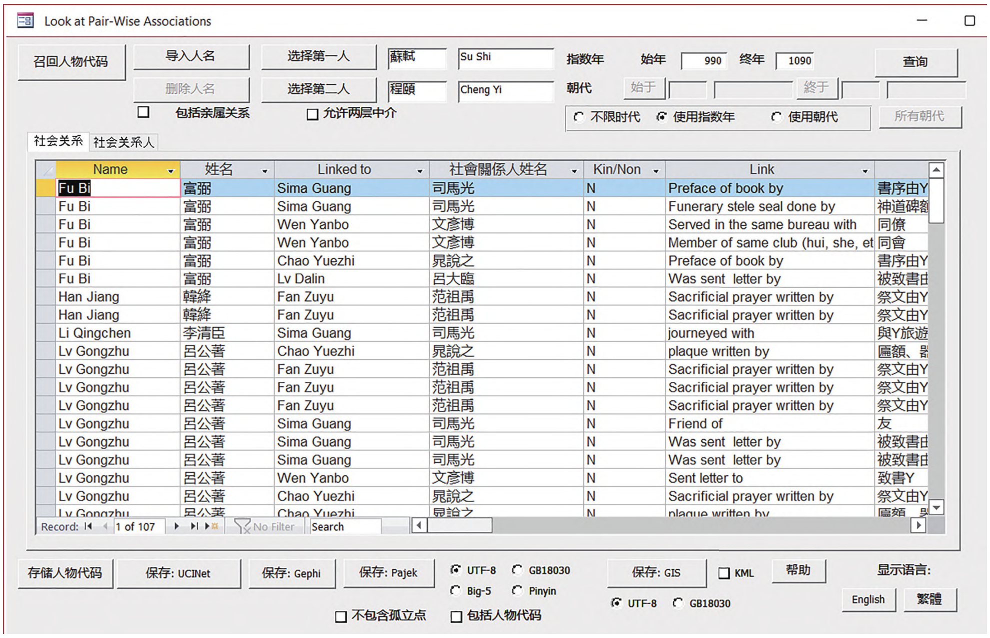

图 9 “查询两人之间社会关系”窗体

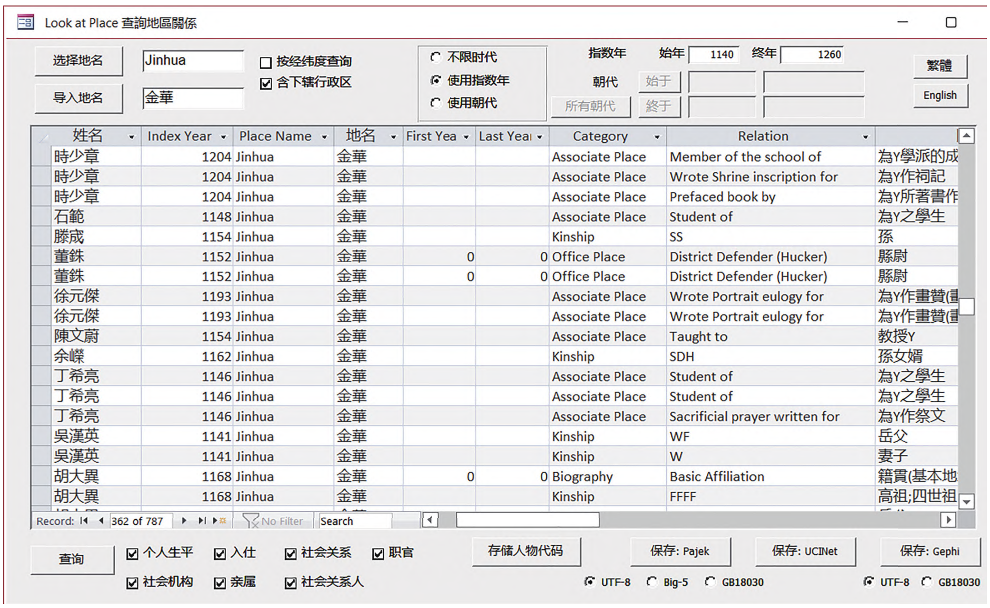

图 10 与金华县相关的南宋人物

“查询两人之间社会关系”的算法可以发现经过一层或两层中间人间接产生联系的情况。这对探索不同人群之间的互动关系有着广泛的应用前景。例如，人们一般认为南宋理学家排斥佛教。从这个想法出发，我们可以使用“查询两人之

间社会关系”窗体来探索这两个群体的社会关系网在多大程度上交叉重叠。

最后要介绍的窗体是“查询地区关系”。通过这个窗体，学者可以检索到与指定地址相关的各种历史人物，如在此地任职的人、在此地的书院讲学或肄业的人、在此地有亲属或社会关系的人、在此地参加了科举考试的人等。通过对这些人物及其人际关系的梳理，学者可以发现不同人物是如何借由相互重叠的不同类型的社会网络聚于一处的，而这些社会网络是很难通过阅读史料理出头绪的。

图10是使用“查询地区关系”找到的南宋末年与金华县相关的787条社会关系和亲属关系数据。“查询地区关系”窗体可以将这些分散的信息汇总起来，用于进一步分析。

三、中国历代人物传记资料库关系网络数据综述

1. 社会关系

截至2020年5月，CBDB共收集了482,979条亲属关系与149,611条社会关系（非亲属关系）数据。这些数据在时间上主要分布于公元600年至1912年之间（图11）。在数据收集和数据结构的设计上，我们始终贯彻着为群体传记学研究服务的目标。

从图11可见，CBDB中的社会关系数据集中分布在唐宋时期。这一现象与CBDB的开发历史有关。CBDB最早的开发者——包括郝若贝本人——都是研究唐宋史的学者。因此，直到最近，CBDB中有超过75%的社会关系数据都与这一时期有关。为了改变这一数据分布不平衡的现象，我们未来几年将把工作重心放在明清。

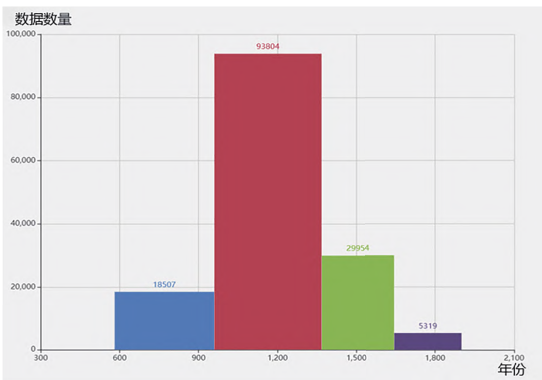

图 11　当前中国历代人物传记资料库中社会关系数据的时代分布

在收集社会关系数据的过程中，我们希望尽可能忠实地反映史料中对各种社会关系的描述，也希望数据库便于使用。为了平衡这两个目标，我们从两个层面来维护社会关系数据。在第一层面，我们定义了480种社会关系。这些社会关系通常是成对的，如“墓志铭由Y所作”和“为Y作墓志铭”“排挤”和“遭

Y排挤”等。在第二层面，我们将上述480种社会关系分为十个大类，其中著述关系、政治关系、学术关系三类的数据最多（图12）。这三类社会关系的数据量之所以较大，是由CBDB的史料来源决定的。与保存了大量教会记录和银行账簿的欧洲史料不同，在中国古代文献中，有关政治交往、学术交往和文学交往的信息非常丰富。这些信息往往以清晰的格式和程式化的语言记录在传记索引、文集和地方志等材料中，这种记录形式正适合我们进行半自动化的信息提取。

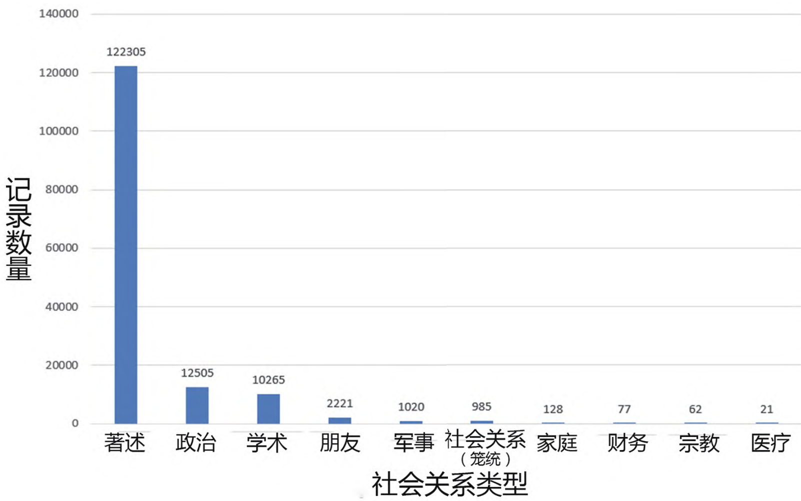

图12　中国历代人物传记资料库中各类社会关系的数据量

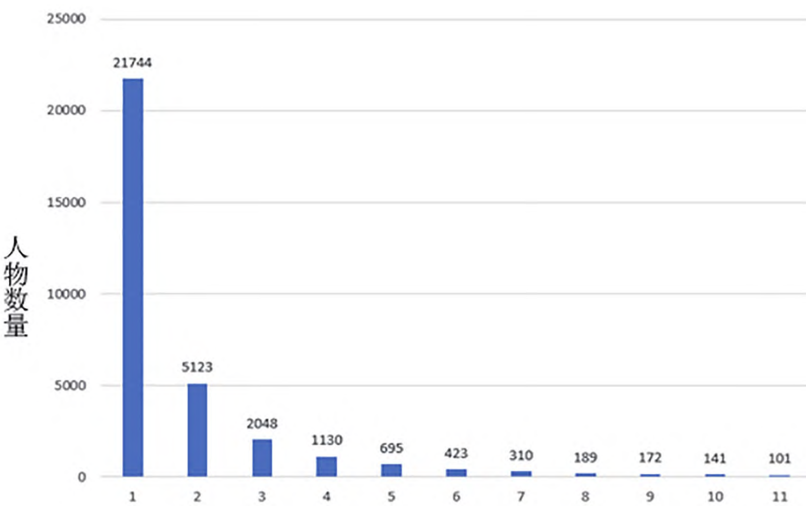

图13　中国历代人物传记资料库社会关系网络中不同加权度中心性值的人数分布
（本图仅显示人物数量≥100的加权度中心性数据）

将CBDB中所有社会关系数据绘制成图，可以得到一个包含33,433个人物（节点）与149,610条社会关系（边）的庞大网络。在这个网络中，若两人之间存在多重社会关系，我们就将这些关系合并，把被合并的边数算作边的权重。经过这样的处理，此网络中边的数量便从最初的149,610条减少到55,799条。将这个网络作为无向图（undirected graph）进行分析，可知该网络的平均度（average degree）为3.35，网络直径（network diameter）为19。其平均路径长度（average path length）为5.509，这说明此网络中任一节点到其他节点的平均距离为5至6步。

另外，此网络中节点的加权度中心性（degree centrality）遵循幂律分布：90.20%的人物加权度中心性小于等于4，只有极少数人（3.69%）在10以上（图13）。

同样，根据加权度中心性的分布情况计算出的网络基尼系数为0.636，这说明社会关系数据在CBDB所收录的历史人物中分布高度不均（图14）。

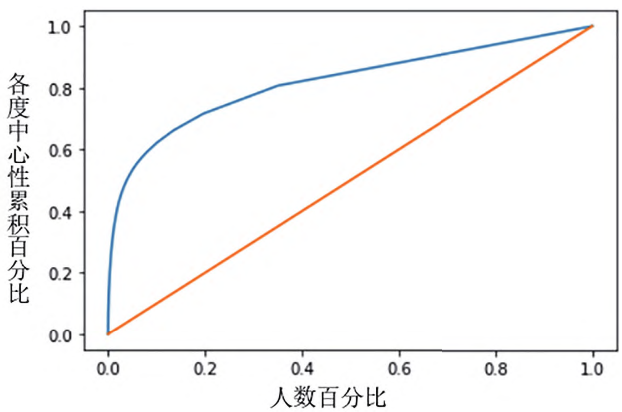

图14　中国历代人物传记资料库社会关系网络中加权度中心性的洛伦兹曲线

此外，还可以通过与Erdős-Rényi随机图进行对比来探索CBDB中社会关系网络数据的特质。Erdős-Rényi随机图是在给定节点数量和边数量的情况下，通过将边在节点之间进行随机分配而生成的随机网络模型。为与CBDB的社会关系网进行比较，我们设定一个Erdős-Rényi模型$G(n, M)$，其中节点数n=33,433，边数M=55,799。[①]接下来，我们从k-clique角度将CBDB社会关系网络与这个随机模型做一个对比。所谓k-clique是指由k个节点组成的、任意两个节点之间都有边相连的子网络。在上述Erdős-Rényi随机模型中，平均有119个3-cliques，但只有$1.91*10^{-5}$个4-cliques。也就是说，在随机模型中出现$k≥4$的k-clique的可能性极低。然而，在CBDB的实际数据建构的社会关系网络中，存在着313个3-cliques，154个4-cliques，54个5-cliques，27个6-cliques。这意味着CBDB的社会关系网络存在着强烈的局部聚集倾向。另外值得注意的是，尽管CBDB的社

① 在这个Erdős−Rényi模型中，任意两个节点之间连接成边的概率$p=m/\binom{n}{k}$=0.000267.

会关系网络中共含1,099个分量（components，即相互不连通的局部网络），但92%的人物都属于同一个巨型分量。

2. 亲属关系

将CBDB亲属关系数据中任意两人间的所有多重关系（如两人既为表亲亦为姻亲）进行合并后，我们可以得到一个包含244,658位历史人物与481,476条亲属关系的网络。此网络包含479种亲属关系，其中最主要的有13种。这13种亲属关系每种都拥有超过5,000条记录。它们分别是：儿子（S）、父亲（F）、哥哥（B+）、弟弟（B-）、丈夫（H）、祖父（FF）、妻子（W）、孙子（SS）、曾祖父（SSS）、母亲（M）、岳父（WF）以及女婿（DH）。（见图15）

这个亲属关系网是一个无向无权重网络，整个网络的平均度为1.97，平均路径长度为24.235，网络直径为79。[①]此网络和社会关系网络相比显得更加松散：它包含23,532个分量，规模最大的分量中只有51,738人（占总人数的21.15%），而规模次之的分量中更只有357人。

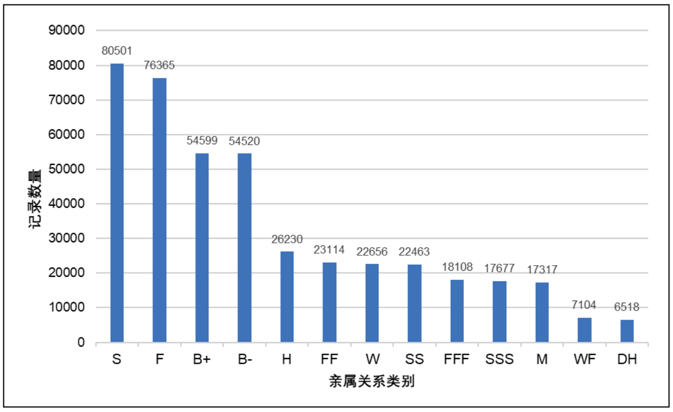

图15　中国历代人物传记资料库中不同亲属关系类别下的记录数量

（本图仅显示记录数量超过5,000的亲属关系）

CBDB亲属网络中有15,573人亦出现于上节所述的社会关系网络中。我们通过下列计算可评估两个网络的重合度：

$$\frac{\text{亲属关系网与社会关系网的重合人数}}{\text{社会关系网的总人数}} = \frac{15{,}573}{33{,}433} \approx 46.58\%$$

①CBDB只收集史料中明确提到的亲属关系。比如，在前文黄庭坚的例子中，史料明确提到了他的一个母舅（李布），而没有明确提及他的另一个母舅（李常）和姨母（洪亶妻李氏），尽管这两个人与黄庭坚的关系同样密切。因此，这里使用的网络分析指标（如平均度、平均路径长度、网络直径等）并不能十分准确地描述历史上亲属关系网的实际情况。这些指标所反映的主要是亲属关系在历史文献和CBDB中的记录形态。

反之，

$$\frac{\text{亲属关系网与社会关系网的重合人数}}{\text{亲属关系网的总人数}}$$
$$=\frac{15{,}573}{244{,}658}\approx 6.37\%$$

由这两个百分比可以推知，如果一个历史人物的社会关系被记录在史料中，那么他的家庭成员载入史籍的概率就会显著增加（由图15可知，这些更容易被录入史籍的亲属主要是他的儿子、父亲、兄弟等）。这意味着社会知名人士的家庭成员往往更有机会成为具有影响力的人。

四、从史料中挖掘数据

1. 选择合适的工作材料

CBDB优先处理那些适合使用半自动化方式、系统性地进行传记数据挖掘的语料。这些语料通常需要满足以下几个条件，即包含大量可信度高的传记资料、行文逻辑清晰一致、有文档可用。成系统地挖掘数据不仅能迅速扩大CBDB数据的覆盖面，更重要的是它保证了以CBDB数据为基础开展的群体传记学研究具有统计学意义。例如,《全宋文》对于CBDB项目来说就是理想的工作语料。作为有宋一代的文章总集,《全宋文》目录几乎巨细靡遗地罗列了宋代三百年间的17.8万篇文章，涵盖了书信、墓志、哀辞、颂赞、碑记等各种文体，反映了人与人之间多种多样的社会关系。

2. 为数据挖掘准备文档

CBDB优先选择有文档可用的史料进行数据挖掘。若文档有版权保护，我们会向版权方申请授权。有些史料尚未电子化，但是对扩大CBDB的数据覆盖面意义重大。这时我们会根据具体情况决定是否使用光学字符识别（OCR）和人工校对相结合的方式创建用于数据挖掘的电子文档，还是采用人工录入的方式将史料中的传记数据直接录入数据库中。

如果决定创建电子文档用于数据挖掘，我们会对文本进行高清扫描，再用OCR软件Abbyy将扫描件转化为文本。OCR的正确率取决于许多因素，包括原始文献的排版格式和扫描件的清晰度。我们通常会先选出若干页进行测评，检查OCR的正确率，只有当正确率超过95%时才会开始下一步的校对工作。

如果决定采用人工数据录入方案，我们会视录入材料的不同性质采取不同的工作流程。我们有时建议录入者使用CBDB开源在线录入系统直接录入数据，有时则建议他们将数据录入到我们预先设计好字段的Excel工作表中。在数据全部录入Excel表之后，我们会使用一些自动化方案对Excel表中的数据进行批量处理，将其转成符合CBDB数据模型的格式并上传到数据库中。在设计Excel表时，我们会尽量减少预设字段的数量。根据我们的经验，过多的字段会大大降低工作效率。

3. 数据挖掘

不同文本的排版格式与叙述方式各不相同。因此，我们针对不同文本采用不同的数据挖掘方法。最常用的是正则表达式，它的基本思想是利用文本中常见的词语组合和常见的句法规则来检索文本。它不需要任何训练集数据，它对于行文相对程式化的史料最为有效。以师生关系为例，史料中对师生关系的一个常用的表述方式是“从X游”（其中X是老师的名字）。据此，我们可以设计一个算法，将史料所有出现“从X游”的字符串都提取出来，并把每一个X识别为师生关系中的老师。

对于行文不够程式化、叙述语言多变的文本，我们会选择其他机器学习方案。一个方案是随机森林。这是一个监督性机器学习算法，我们曾经尝试过用这个方法从传记文本中提取社会关系。首先，我们从语料中收集所有包含两个人名的句子，并随机选取其中一小部分句子作为训练集。然后，我们根据训练集中每个字的字频建立一个高维向量空间，并把收集到的每个句子都表示为这个空间里的一个向量。之后，我们邀请学者使用CBDB代码表中的各种社会关系对训练集中的每个句子进行人工标注，再利用随机森林模型对这些句子所表达的社会关系与其表征向量的数学属性之间的相关性进行归纳，从而建立分类模型。最后，我们利用这个分类模型对所有包含两个人名的句子进行分类，也就是根据这些句子的表征向量来推测其所表达的社会关系。简而言之，我们将从史料中提取社会关系数据的问题转化为先找出所有存在人名共现的句子、再使用预定义的社会关系类别对这些句子进行分类的问题。

我们使用的另一个方案是双向长短期记忆网络加条件随机场（BiLSTM-CRF）。与随机森林将命名实体的识别问题转化为分类问题的做法不同，BiLSTM-CRF将其视为序列标注问题。与随机森林相比，我们发现BERT模型与BiLSTM-CRF组合非常有效。BERT是将文本转化为向量的无监督词嵌入语言模型。BERT超越了此前所有技术，因为它是充分的双向模型，并且考虑到了一个字每次出现

的上下文。用BERT将文本转换为向量后，我们再使用BiLSTM-CRF来推测文本中每一个字是否是作为人名、官职、地名、亲属关系等专有名词的一部分出现在文本中的。这种推测既考虑了当前字的上下文环境，也考虑了当前字在整个训练集中是否经常作为某类专有名词的一部分出现。我们运用这个方法，对地方志中的职官志部分进行了数据挖掘，结果非常成功。我们以很高的正确率从地方志中提取了大量传记信息，包括地方官员的入仕途径、所任官职及其亲属关系。

4. 数据标准化

数据挖掘完成之后，我们通常会得到一批带标注的XML文件或者分列清晰的Excel文件。这些文件中包含大量有关命名实体（如人物、官职、社会关系、亲属关系等）及其相互关系的数据。在CBDB的数据模型中，我们将每种实体及其属性数据存储在各自的代码表（code table）中，再将历史人物与不同实体之间的关系存储在各自的数据表（data table）中。因此，我们还要将通过数据挖掘得到的XML或Excel文件转换为一组符合上述数据模型的、相互关联的代码表和数据表。

在这一环节，我们面临的主要挑战来自于专名消歧（disambiguation）。CBDB中每一个人物拥有唯一的代码，其他实体（如地址、官职、各种社会关系与亲属关系等）亦是如此。然而，同一人物在历史文献中可能有不同的称呼方式，反之不同人物也可能出现同名同姓的情况。判断史料中出现的某个人名具体指代的是哪位人物，进而为它赋予正确的人物代码，这个步骤我们称之为“消歧”。在将挖掘到的新数据导入CBDB之前，我们会首先对这批新数据进行内部消歧，再将这些人物与CBDB中已经收录的人物进行比对和消歧。在这一过程中，我们有时会邀请相关领域的专家学者帮忙考证，但由于数据量庞大，我们必须根据史料本身的特征、人物的基本信息（生卒年、字号、籍贯、官职等）及其社会关系、亲属关系设计算法，尽可能地用自动化方法完成消歧。例如，如果同一姓名在同一史料中反复出现，那么它极有可能是指同一个人。反之，如果生活时代相差数百年、籍贯地和登科年各不相同，那么就算姓名完全相同，也不太可能是同一个人。麻烦的是，很多时候，史料中并未提供相关人物的生卒年、籍贯、登科年等信息。因此，近几年，我们也在使用社会关系和亲属关系数据开发消歧算法，并取得了很大进展。这一算法的基本思路是，如果不同史料中记载的“两个人物”不仅姓名相同，而且他们的社会关系网或亲属关系网高度重合（如他们的父母、师友姓名完全相同），那么即使没有其他信息，我们也能以较大把握认定他们是同一人物。

附录：现已发表的使用 CBDB 数据的研究成果

1. 刘飞燕、高剑波:《隋唐至宋时期精英社会网络动力学的演化研究》,《数字人文》2020年第1期。

2. 许雅惠:《北宋晚期金石收藏的社会网络分析》,《新史学》2018年第4期。

3. 严承希、王军:《数字人文视角：基于符号分析法的宋代政治网络可视化研究》,《中国图书馆学报》2018年第5期。

4. Peter K. Bol, “From Kinship to Collegiality: Changing Literati Networks, 1100-1400,” *Journal of Historical Network Research*, vol. 5, 2021, pp. 36-61.

5. Song Chen, “Governing a Multicentered Empire: Prefects and Their Networks in the 1040s and 1210s,” *State Power in China, 900-1325*, eds. Patricia Ebrey and Paul J. Smith, Seattle: University of Washington Press, 2016, pp. 101-152.

6. Hilde De Weerdt et al., “Is There a Faction in This List?,” *Journal of Chinese History*, vol. 4, no. 2, 2020, pp. 347-389.

7. Nicolas Tackett, “The Evolution of the Tang Political Elite and its Marriage Network,” *Journal of Chinese History*, vol. 4, no. 2, 2020, pp. 277-304.

Structuring, Recording, and Analyzing Historical Networks in the China Biographical Database(CBDB)

Michael Fuller, Hongsu Wang

Abstract: The China Biographical Database (CBDB) is a relational database of over 470,000 individuals from pre-modern Chinese history by 2020.5. CBDB is distinctive as a prosopographical database in that it allows users to generate kinship and social networks for individuals – and groups of individuals – in the database. At the beginning of the project, to develop CBDB, we sought models among other digital prosopography projects but realized that, with CBDB's focus on analytic procedures and extracting data from the vast resources of the historical Chinese textual archive, we were developing a highly different model for digital prosopography. This paper presents an overview of the China Biographical Database, its capacities for exploring networks, and how we are extending those capacities through the ever-broadening extraction of data from the corpora of historical sources.

Keywords: China Biographical Database Project; Social Network Analysis; Prosopography; Relational Database