数字人文与中国研究的网络基础设施建设

包弼德（哈佛大学）
夏翠娟翻译整理（上海图书馆）
王宏甦审校（哈佛大学）

摘　要　哈佛大学中国史教授包弼德在“第九届上海国际图书馆论坛（SILF2018）”的主旨报告，讨论了“网络基础设施”作为联接中国研究领域众多独立的数据库的桥梁，对于数字人文的发展来说，有着至关重要的意义。他认为数字人文不仅仅是工具，而是在知识进步的过程中引起范式和理论变革的方法和技术体系。他以中国历代人物传记资料库（CBDB）为例，解释了“关系数据库”与常用的文本数据库的本质区别，并以众多的案例来说明地理信息系统（GIS）、社会网络分析，以及文本分析工具和平台，是如何帮助知识的进步，并引起范式的变革和新理论的诞生。最后提出了联合各研究型图书馆，构建“全球智慧数据平台”的愿景。

关键词　数字人文　网络基础设施　社会网络分析　空间分析　文本分析

DOI: 10.13663/j.cnki.lj.2018.11.003

The Digital Humanities and a Cyberinfrastructure for China Studies

Peter K. Bol (Harvard University)

Abstract　Harvard University Chinese History Professor Peter K. Bol's keynote report at the 9th Shanghai International Library Forum (SILF2018) discussed “Network Infrastructure” as a bridge to connect many independent databases in Chinese research has a vital meaning for the development of digital humanities. He believes that digital humanities is not only a tool, but a method and technical system that causes paradigm and theoretical change in the process of knowledge advancement. Taking the Chinese historical biographies database (CBDB) as an example, he explained the essential difference between the “relational database” and the commonly used text database in a simple way, and explained how the geographic information system (GIS), social network analysis, text analysis tools and platforms help to promote knowledge advances and lead to paradigm shifts and the birth of new theories. Finally, it is proposed to combine the research libraries to build the vision of the “Global Smart Data Platform”.

Keywords　Digital humanities, Cyberinfrastructure, Social Network Analysis, Spacial analysis, Text analysis

上海图书馆是一个我曾作为用户来过的图书馆，它是一个在中国的图书馆资源开放和获取上贡献过众多标准规范的图书馆。当我们谈到“包容”时，上海图书馆的确是一个在这方面起着引领作用的图书馆，其领导者们为此做出了卓越的贡献。

1　“网络基础设施”的定义和意义

所以，今天我非常荣幸地站在这里。我想从“网络基础设施”的定义开始，我将用这样的方式定义它：它是联接两个方面的系统，一是计算、存储、交流的基础技术，二是软件、服务、平台和各种工具。现在，基础技术和软件平台需要独立于特定的项目和学科，但同时又必须可应用于特定的语言、项目和学科。“网络基础设施”需要那些理解技术并认同分享理念的人。这就是为什么与图书馆员谈这个话题非常重要的原因之一。因为与其他人相比，他

们有着较强的分享标准和规范的意愿，我们也希望他们能在分享技术上起到引领作用。

我们为什么需要一个中国研究的“网络基础设施”？今年3月，我们在上海开了一个会，全世界的主要中文图书馆，包括上海图书馆、北京大学图书馆、国家图书馆、浙江大学图书馆和台湾、香港地区的中文图书馆，还有日本、欧洲、加拿大、美国的中文图书馆，相聚在一起，讨论如何构建“网络基础设施”。为什么？因为我们看到了资源的蓬勃增长，这是“数字人文”发展带来的结果。有太多的不同数据库，相互独立，没有连接。我们看到图书馆员和学者重复相同的工作，数字化同样的资源，一遍又一遍，造成巨大的浪费。我们看到了一个不断扩张的数字化生态系统。但是，我们需要决定什么时候要自建，什么时候需要购买。

2 “数字人文”与知识进步的三种方式

什么是“数字人文”？现在让我们先讨论一下“数字人文”，最后再回到“网络基础设施”。我认为一是信息的发现、分析和可视化的技术，二是数字出版物（新的出版物不仅仅是印刷版本的复制品，而且能提供新的分析途径和多媒体展示方式），第三点，也是最具挑战性的一点，是数字人文研究和整个数字生态系统。

我们许多人把数字人文仅仅当成工具，但实际上，数字人文还有一个重要的领域需要我们去研究和了解，那就是数字化数据的膨胀、用户以及数字化数据生产者之间的协作。这不是我们可以扔给计算机科学家和社会学家的。让我们看看网站的增长量，5年前到现在，网络数据的增长量比全世界所有图书馆的馆藏加起来的还要多得多。我们必须要寻找一种方法来研究这些数据。

现在，让我们来谈谈知识的进步。自19世纪始，我们一般认为专业化就是知识的进步，这被证明是对的。知识的进步依赖于专业化，它开创了新的学科，可以集中精力只关注自己的领域而不关注其他学科，已经被证明是一种高效的知识进化方式。

知识进步的第二种方式是理论和范式的转型。学者、科学家们以某种特定的模式工作。在某种时候，这种模式开始发生变化，新的研究课题诞生，但这种变化不是每年都发生，是呈波浪状发生的。这在每个领域内都适用，包括文献学、历史学和信息技术科学。

知识进步的第三种方式是建立在工具的发展的基础上。这就是数字人文可以扮演某种角色的地方，也是为什么许多学者之所以看轻数字人文的地方。因为他们把工具看成纯粹的工具。但实际上，工具可以作为知识进步的仪器，就好比是显微镜和望远镜，可以帮助我们看到小到人眼看不清的和远到人眼望不到的东西。

3 “数字人文”所带来的智力和理论层面的五个飞跃

我的论点是，数字人文是人文研究的工具，为人文研究提供数据科学的研究方法。我想说的是这是一系列智力和理论层面的飞跃。现在我认为是图书馆员而不是用户已经创造了这样的飞跃。这也是我们需要让学生和研究人员接受教育和培训的领域。

第一个飞跃是从数据库到关系数据库，我将解释为什么这一点很重要；第二个是从地图到地理信息系统和空间分析的飞跃；第三是社会网络分析；第四是用于文本分析的工具和平台。最后是基于API的网络基础设施。我将按照顺序逐个进行解释。

3.1 从数据库到关系数据库

首先是从数据库到关系数据库。现在有许许多多不同类型的数据库，就像我们今年3月在上海举办的网络基础设施研讨会上看到的那样，每个人都有数据库。不仅是图书馆员自建的，还有图书馆付钱购买的商业公司生产的数据库产品，学术期刊也是这样的数据库。它们都是一些文本数据库，这是我们对数据库的通常认识。关于文本数据库的一个重点是它们可以被检索，这一点对学术研究非常有益。文本数据库的检索功能，尤其是那些可以让我们看到很多很多原始资料的数据库。以中文研究为

例，如果图书馆能付钱购买的话，几乎每位学者都会使用的爱如生数据库。台湾“中研院”的汉籍电子文献数据库则是另一个例子。

今天，我们也有一些图像数据库。有些数据库只有图片而没有文本，最著名的例子是敦煌的图像档案库。现在我们可以更进一步，我们有一些从文本和图像里抽取信息的数据库，如明清人物权威档，在该数据库中，包括大量从文本里抽取出来的信息，以及通过超链接来连接不同来源的文本（不是通过代码表创建的关联关系）。我认为这样的数据库构筑了我们的未来。

我现在要说的是关系数据库（Relational Databases）（译者注：包教授这里提到的“关系数据库”不仅仅是指技术层面上的“关系型数据库”，而是指存储了大量数据间标准化的关联关系的数据库），关系数据库更复杂，它们利用复杂的算法从文本中提取信息，经常需要写很复杂的正则表达式来读取数字化文本和数据点（Data Points）。它们使用大量的代码表、数据表，所以能支持多种多样的查询。我假设这里的每个人都或多或少地了解关系数据库，我不打算精确地解释它是怎么工作的，我只想说一点，那就是当关系数据库被很好地设计，存储着大量的信息，并被良好定义，就会支持以前无法实现的查询，让我们发现以前无法发现的知识。

中国历代人物传记资料库（CBDB）是我领导的与北京大学中古史研究中心，以及“中研院”史语所合作的一个项目，在中国也有很多合作伙伴，包括上海图书馆和北京大学图书馆。CBDB 现在有了 42 万余人，主要的时间覆盖范围是从 6 世纪到 20 世纪早期，在接下来的几年，有望再增加一到两百万数据。

CBDB 支持各种各样的查询，包括单个人物查询、入仕途径、官职查询，还可以查询社会关系网络，查询两个人之间的社会关系，甚至查询不同地区间人物的关系。这样的数据库为研究者提供了一种新的方式，基于大量数据来思考人类的过去和历史。因此理解像 CBDB 这样的数据库如何帮助用户看到大规模的数据是很重要的。它不是一个人物辞典，虽然人们经常像词典一样使用它。

当我们使用一个词典的时候，如果发现了一个错误、一个不准确的定义，如一个循环定义，我们会认为这个词典很糟糕。但是 CBDB 却不同。如果用户得到 1 000 个反映了一定趋势的例子，在这 1 000 个例子里面有 30 个错误，这些错误对研究者认知整个趋势的影响不大，对吗？这只是 3% 的错误率，这并不是什么大问题。真正重要的是学会如何对待大量的数据。

虽然我们已经花费了大量的时间和金钱来努力确保数据的正确性。但是我们也确实知道我们所使用的历史记录中有着大量的脏数据。所以我要说这是一件非常艰难的工作。你们看到这里有一个宣传册，介绍了 CBDB 和中文在线合作的一个项目，这个项目将帮助 CBDB 在中国得到更广泛的应用。明天下午（10 月 19 日），我们将与上海图书馆签订一个合作协议，感谢陈超馆长和刘炜副馆长，帮助我们把 CBDB 发布成了关联数据。

请记住 CBDB 将继续发展增长，请阅读这份宣传册并尝试理解，大量的数据意味着人们将通过人物传记来透视中国的历史。

中国最伟大的历史著作始于 2 100 年以前的历史学家司马迁的《史记》。很有意思的是，当司马迁开始延续他父亲的工作来撰写他所认知的那个世界的历史时，他本来想追溯到汉代历史的原始结构，但是最终，我认为他一定以失望告终。因为汉代历史并没有呈现出一个完美的结构。

这时，司马迁就把一半的《史记》转向了人物传记。因为传记向我们呈现了不同的人，有着不同人各自的观点和视角，它也告诉我们，历史是凌乱的，充满了冲突和异见、联盟和解体。但是从那时起，几乎所有的伟大中国历史著作，都是半部人物传记。在中国研究中，传记的重要性不应该被低估。

这里我想向大家介绍宣传册中所展示的中文在线的网站上有各种各样 CBDB 的可视化功能。我想说的另一件事是，大学不善于大众化推广，也不善于推销和赚钱，不过善于申请基金来花钱。当然对于研究来说这是好事。另一方面，但如果我们真的想向公众推广，必须让学术产品谋求商业化的道路。当我们在今年 3 月份的上海会议上提到这个问题时，一个参会

者说："商业化就是大众化"。我认为，尤其是在中国，这是正确的。所以我真的感到非常欣喜，可以和中文在线、上海图书馆、北京大学以及许多其他机构合作。我们希望他们的加入，能让 CBDB 成为中国研究的主要关系数据库。

CBDB 给我们多种多样的信息，大量的数据统计分析结果告诉我们，中国过去的平均死亡年龄是 61 到 63 之间，这让我们在学校里学到的过去的平均年龄是 35 变得没有意义。图 1 展示了男性的预期寿命。在我们的数据库里，有 10% 的女性，女性的预期寿命跟男性很不一样，参见图 2。我们看到了两个峰值，育龄女性的死亡率也达到了一个顶点，这在中国的过去是很合理的现象。CBDB 为中国人口的研究提供了完整的数据来源，尤其是精英人口。

3.2　从地图到地理信息系统（GIS）

接下来我想说一下从地图到地理信息系统（GIS）的转变。大家知道，许多人都认为 GIS 是关于地图的地理信息系统。地图存在了很长的时间，我认为早期的中国地图可以追溯到公元前 500 年到 600 年甚至更早。今天，大部分的历史地图已经被扫描并可访问，例如台湾"中央研究院"著名的"地图数位典藏整合查询系统"。

我们有在线的老地图，包括国外的地图。但是地图的问题在于它们承载了大量不同类型的信息，却只能在一个图层上呈现：河流、村镇、山

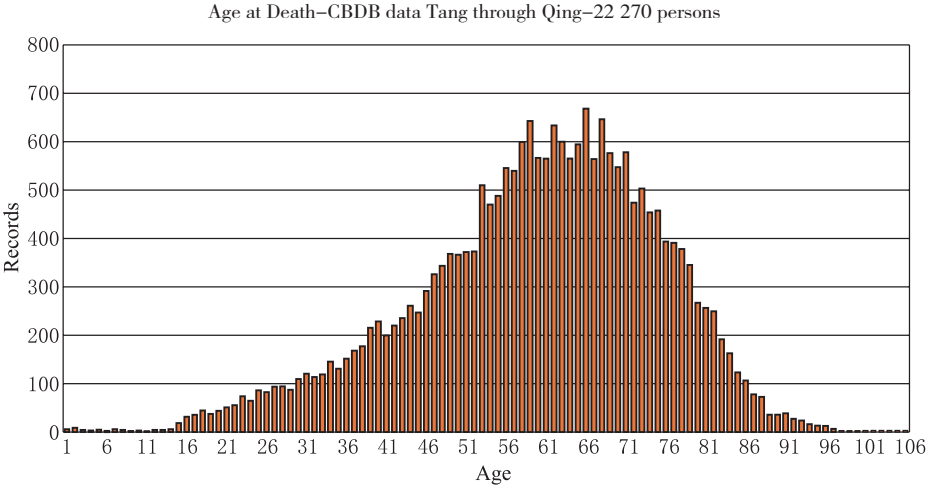

图 1　CBDB 中从唐代到清代 22 270 个男性的死亡年龄

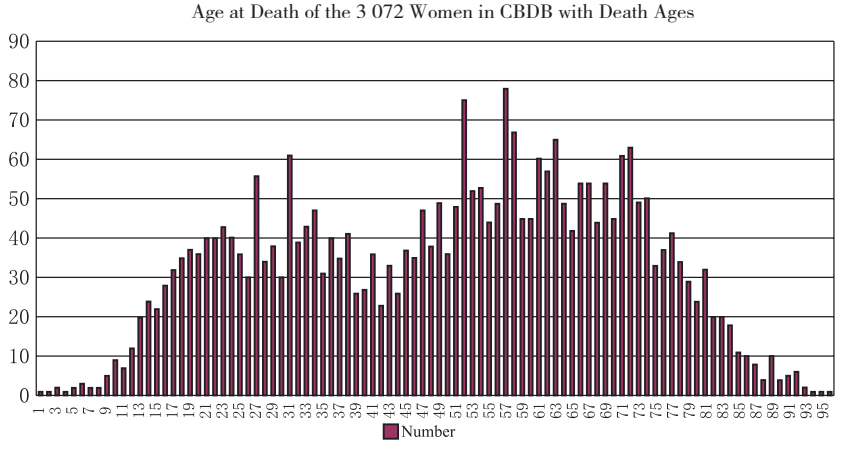

图 2　CBDB 中 3 072 个女性的死亡年龄

峦、道路、边界等全部印刷在一起，GIS 和空间分析的任务是将其分成不同的层，分别呈现。复旦大学的中国历史地理信息系统（CHGIS）就是这样一个支持空间分析的系统，它同时也是一个 GIS 数据集。该数据集包括从公元前 221 年到 1911 年的地名数据，现在也有了 20 世纪和 21 世纪的数据，从秦始皇到辛亥革命，这很重要，因为中国历史记录更强调地点而不是时间。

如果你看一个传记，它将告诉你一个人来自什么地点，而他生活的时间信息则常常是模糊的。所以在中国历史著作里，在汉代最伟大的历史著作两《汉书》中，我们可以知道历史是在哪里发生的。问题在于，我们是否想进行空间分析。那些地点在哪里？中国历史地理信息系统（CHGIS）可以回答这些问题。通过历史地理信息系统的分层地图能看到历史信息的空间关系。例如，可以看出，一个县的所在地，与周围河流湖泊之间是什么关系。我们可以基于矢量叠加源的数字高程模型，看到一个县的辖区是如何与当地的地理景观产生联系的。进一步，我们可以从 CBDB 中提取信息并映射到地图上去，例如生成明代（1368–1644）的进士地理分布图，参见图 3。进士是从中国古代的高等科举考试产生的，自 10 世纪开始，大量的官员来自进士群体。在这个例子中，如果我们仔细研究明代进士的地理位置，就会发现明代政治体系的空间特点。他们主要来自江南地区，中国的东南部以及江西，远远超过其他地区。

我们放眼整个中国的历史，都可以看到这种自然地理特征对整个国家发展的影响。现在浙江大学正在基于哈佛的世界地图平台建设学术地图发布平台，他们取得了卓越的进展。因为哈佛的这个平台是开放访问的，我们将代码给了他们。世界地图平台允许我们加入不同的信息，并向公众开放。它是免费的，任何人都可以共享。实际上，如果你想做一些类似的研究工作，你可以从 Github 中克隆代码，自己建立一个本地系统，现在有超过 150 万的用户在使用它。

3.3 社会网络关系分析

第三个重要的工具层面的飞跃是“社会网络关系分析”。当我们研究大规模传记信息时，这种方法尤为重要。我将用一个社会网络分析图作为例子来说明：这是蒙古统治时期的浙江南部的吴州府，或称金华（金华火腿的金华）。图 4 里有 900 个来自 CBDB 的人，如果要问他们的社会关系是什么？如果有这样一张社会关系图，我们就可以回答这样一些问题：这些人是哪些组成的，有哪些小团体，谁是最重要的人。像这样一张有大量节点的图看起来没什么用，但是能给人很深刻的印象，可以从中获得很多信息。进一步，我们可以简化这张不容易分析的社会网络关系图，比如我想让它只

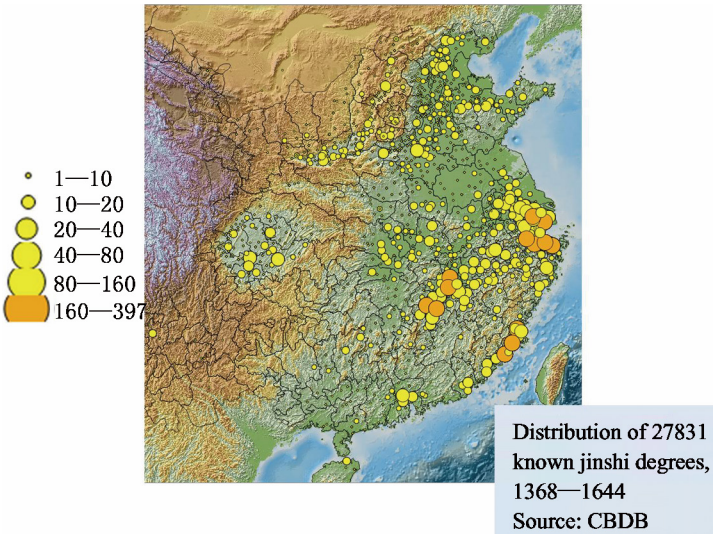

图 3 明代进士分布图

呈现与 4 个人有关系的人。从简化图（参见图 5）中，我们可以看到社会网络关系的最基本结构。

实际上，这就是概念飞跃产生的地方。我们要理解，一个地理信息系统其实是一个数据库，一个社会关系图实际是一系列可计算的关系的展示和呈现。这些可计算的数据非常重要，它告诉我们，在某一段时间内，谁的社会关系最多，谁参与的不同群体最多，谁有最高的关注度，或者谁关注别人最多。社会网络关系分析是基于数学计算的，它可以引导我们的学生以及图书馆员如何分析数学统计的结果，让他们不至于被误导，这也是我们教授数字人文的一项工作。

3.4　文本分析的工具和平台

接下来，我要说的是文本分析的工具和平台。我的学生魏希德和其他人一起开发了

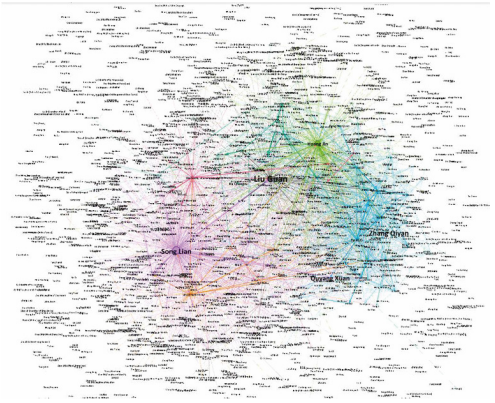

图 4　CBDB 中元代 900 个人物的社会网络关系图

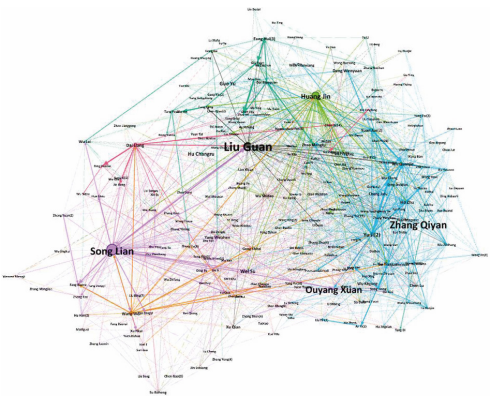

图 5　CBDB 中元代 900 个人物的社会网络关系简化图

一个文本标注工具，叫码库思（Markus，参见图 6），是免费开放的。你可以上传一些中文文本，选择时间范围，它将根据 CBDB 的代码表自动地标记文本中的内容（人名、地名、时间等），内容被标记后，可以导出下载。码库思是由 CBDB 和 CHGIS 的 API 支持的，这种使用 API 来整合不同在线数据的方式非常棒。

这里要介绍的一个著名的中文文本平台是 ctext.org，提供非常重要的文本分析工具，比如有一个工具可以帮助用户在不同的文本中找到相似的段落，深红色表示相似度较高，浅红色表示相似度较小。Ctext 上有成千上万的文本，你可以在这个平台上用你的文本与其他的文本进行对比分析，不需要自己开发代码，也不需要任何 IT 技能。这里有一个例子，一段出自《庄子》的文本和一段出自《吕氏春秋》的文本，我们可以看到在哪里有不同（参见图 7）。Ctext 甚至允许你在传统中文文本中查询插图。

3.5　用 API 构建“网络基础设施”

最后，我要谈一谈利用 API 构建“网络基础设施”。在使用电子资源的时候，我们可能会问如下的问题：我们想了解数字文本的什么？如果我们能看到电子化的资料，那么这些资料能不能检索、下载？当我们不能看也不能查某些资料的时候，无论它是免费的还是商业的，我们可能都会想知道谁拥有它，并寻找办法来访问它。我们能够用两个简单的 CSV 格式的文件来共享数据，一个是数据目录列表，一个是关于这个列表的元数据。

我们已经有了这样的一个平台，就是 textref.org。它集中了不同的文本库，并告诉你什么文本可获得，是否可以检索、浏览、下载。我们也为人物传记做了相同的工作，在网站 biogref.org 上，我们集中了 DDBC、CBDB、DNB，用一个非常简单的 Schema 来描述它们，这是图书馆员非常熟悉的方法。

4　联合建立“全球智慧数据平台”的愿景

但是最重要的进展是全球智慧数据平台，

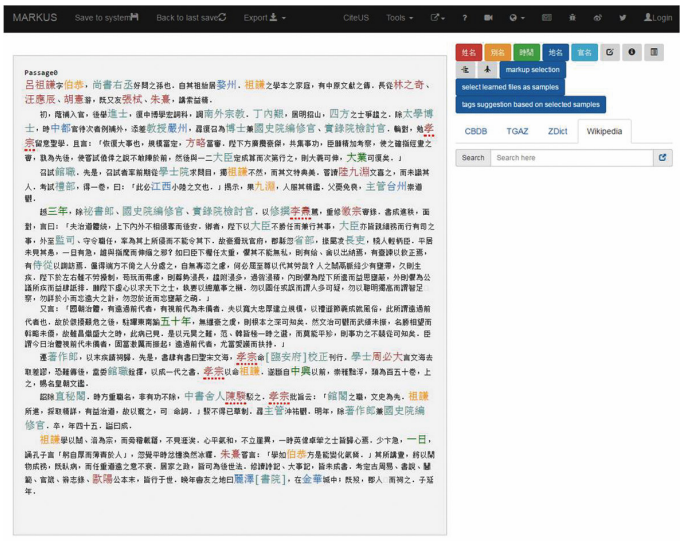

图 6 Markus 的标注界面

下周在北京，我将组织一个学术委员会，包括上海图书馆、浙江大学图书馆、国家图书馆、北京大学图书馆，还有日本、欧洲和美国的图书馆，一起讨论开发一个通用的平台，作为中国研究的网络基础设施。它将建立在超星发现平台上，除了拥有大家都熟知的功能外，我们还将建设第三方数据的提交和分享功能、中国数字人文研究维基平台（通过开源的维基接口自动更新内容），以及其他分析工具，为第三方数据、第三方工具、第三方图书馆定制免费公开的访问元数据和分享数据的规范和方案。此平台将支持用户浏览历史的记录和展示、用户自定义搜索域、第三方专业数据库的开放接口和配套方案、用户手动收集的数据库和 3D 数据过滤。我们的目标是为用户提供一键式的跨数据库学术资料检索服务，支持用户通过一次点击就能浏览、访问。该平台还将收集和分析用户交互数据，深入分析和展示用户的学术

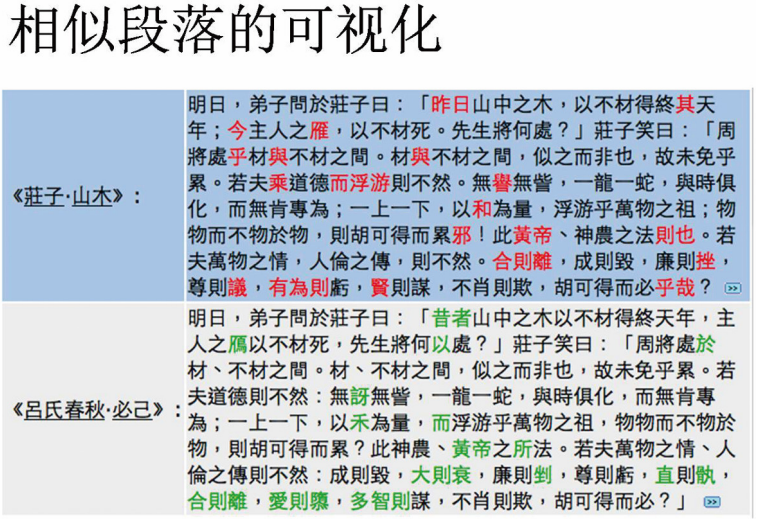

图 7 ctext 的相似段落可视化

动态。

网络基础设施的重要性在于，将独立的数据库、工具和平台连接起来成为可能。这就是我们的目标，这需要大家的共同参与。

到这里，我的报告已近尾声，非常感谢你们！

5　讨论与交流

吴建中（主持人）：包教授，您的演讲真是给了我一个惊喜。您和我们讨论了网络基础设施与数字人文，地理信息系统与社会网络分析，以及文本分析工具和平台，是如何帮助知识的进步，并引起范式的变革和新理论的诞生。所有这些，我认为都与IT技术人员有关，所以我感觉我老了。我有许多问题想要请教您，但还是先把机会留给其他人吧。

刘　炜：我有一个问题，您对图书馆这样的文化记忆机构有什么期望？关于图书馆如何与中文在线这样的商业机构、或Ctext这样的互联网开放服务提供者竞争，您有什么建议吗？

包弼德：这是一个很有趣的问题，刘馆长应该记得今年3月份在上海的会议，当时我们问了这样一个问题：如果Ctext的负责人德龙（Donald sturgeon）明天发生了意外，哪个图书馆将负责Ctext的运营？当时没有人愿意。

所以这就是数字人文面临的问题。数字人文的工具和平台大都是在图书馆外发展起来，图书馆会想要接管它们吗？如果图书馆想，我们就会给。但是这意味着图书馆需要投资，包括技术投资和财政预算。所以，陈馆长，你不得不给IT部门更多的钱。所以我认为这真是一个巨大的挑战。

还有另一个挑战，我们已经在图书馆外开发了许多依赖于软件产品的平台和工具。图书馆拒绝支持这些的一个原因在于，他们不想因为软件的更新换代而不断进行开发和维护。我可以理解这一点，但这不是一个解决问题的办法。如果图书馆想成为一个集成信息管理系统，而不只是一个仓库，就需要认真地考虑如何对待用户生成的内容，以及第三方创造的学术数据内容和数据库。

包弼德　（Peter K. Bol）哈佛大学中国史教授，专攻中国思想史和文化史。与复旦大学历史地理信息中心合作主持中国历史地理信息系统项目（CHGIS），与北京大学中古史研究中心，以及中研院史语所合作，主持中国历代人物传记资料库（CBDB）项目。

夏翠娟　女，上海图书馆（上海科学技术情报研究所）系统网络中心高级工程师，副研究员，上海市图书馆学会学术委员会数字人文专业委员会主任。研究方向：数字人文、知识组织。E-mail：cjxia@libnet.sh.cn　上海　200031

王宏甦　中国历代人物传记资料库项目经理，哈佛大学计量社会科学中心研究员，北京大学中国古代史中心访问学者、清华大学统计中心访问学者。研究方向：数字人文、数据库。　北京　100085