CBDB清华小组

数字人文研究方法综述

潘长在 段楷文 王宏甦 徐力恒 邓柯

2018.8

# Table of Contents

[Table of Contents 2](#_Toc526712631)

[主题模型（Topic Model） 4](#_Toc526712632)

[概述 4](#_Toc526712633)

[在数字人文中的实际应用 4](#_Toc526712634)

[模型介绍 5](#_Toc526712635)

[模型的创造性实践 6](#_Toc526712636)

[序列联配 (Sequence Alignment) 7](#_Toc526712637)

[概述 7](#_Toc526712638)

[数字人文中的应用 7](#_Toc526712639)

[模型方法 7](#_Toc526712640)

[模型的创造性实践 9](#_Toc526712641)

[中文分词 (Chinese Word Segmentation) 10](#_Toc526712642)

[概述 10](#_Toc526712643)

[数字人文中的应用 10](#_Toc526712644)

[模型方法 10](#_Toc526712645)

[主成分分析（Principal Component Analysis） 15](#_Toc526712646)

[概述 15](#_Toc526712647)

[数字人文中的应用 15](#_Toc526712648)

[模型方法 15](#_Toc526712649)

[聚类算法（Cluster） 17](#_Toc526712650)

[概述 17](#_Toc526712651)

[数字人文中的应用 17](#_Toc526712652)

[模型方法 18](#_Toc526712653)

[支持向量机（Support Vector Machine） 22](#_Toc526712654)

[概述 22](#_Toc526712655)

[数字人文中的应用 22](#_Toc526712656)

[模型方法 22](#_Toc526712657)

[词夹子方法（Word-Clip） 25](#_Toc526712658)

[概述 25](#_Toc526712659)

[数字人文中的应用 25](#_Toc526712660)

[模型方法 25](#_Toc526712661)

[词嵌入方法（Word Embedding） 27](#_Toc526712662)

[概述 27](#_Toc526712663)

[模型方法 27](#_Toc526712664)

[社会网络分析 (Social Network Analysis) 29](#_Toc526712665)

[概述 29](#_Toc526712666)

[数字人文中的应用 29](#_Toc526712667)

[模型方法 30](#_Toc526712668)

[地理信息系统 (Geographic Information System) 34](#_Toc526712669)

[概述 34](#_Toc526712670)

[数字人文中的应用 34](#_Toc526712671)

[数据库提升 (Database Upgrade) 35](#_Toc526712672)

[概述 35](#_Toc526712673)

[数字人文中的应用 35](#_Toc526712674)

[参考文献 37](#_Toc526712675)

# 主题模型（Topic Model）

## 概述

主题模型是一种用于发现文档中抽象主题的统计模型。对于一篇文章，我们认为其由多种主题构成，主题由一系列关键词组成，每个词在不同主题下出现的概率是不同的。

例如，对于一则关于NBA新闻报道，主题是体育，而科比这个词出现在体育主题中出现的概率比在娱乐主题中出现的概率要大。这样，通过对文章词句的统计分析，我们可以判断这篇文章是由哪些主题组成的，以及这些主题常用的词汇有哪些。通过这种主题模型分析，我们可以从语义层面解读文章，文章与文章之间的关系也可以从所含主题相近性的角度来进行分析。

## 在数字人文中的实际应用

在Text Mining Deutsche Textarchiv Using Digital Tools and More中，利用主题模型分析德语文献档案，为研究者史无前例地从宏观的角度展现了18世纪德意志精神世界的时代主题（王涛，2016）。

对于《自由中国》杂志中的内容进行主题分析，分析杂志中法治、司法、军队和经济等几个主题中出现关键词和主题分布，进一步分析杂志立场 (陈奕安，江子扬，蔡铭峰，薛化元&刘吉轩，2016) 。

## 模型介绍

主题模型中最基础常用的模型是LDA（Latent Dirichlet Allocation）主题模型，在2003年由David Blei, Andrew Ng和 Michael I. Jordan提出。（D.M. Blei, A.Y. Ng & M.I. Jordan, 2003）它采用了词袋模型，不考虑词与词之间的顺序，将每篇文章看作是一个词频向量。

LDA定义了每篇文章的生成方式如下：

1. 确定主题词汇概率分布及文章主题概率分布

2. 从文章主题分布中抽取一个主题

3. 从被抽到的主题中依照该主题的词汇分布抽取词汇

4 重复2、3步直到文章结束

语料库中的每一篇文档是K个主题构成的多项分布（记作），每个主题是词汇表中V个单词构成的多项分布（记作）。和的先验分布是 Dirichlet分布。

对于第m篇文章，令z代表一次抽样得到的主题随机变量

对于第k个主题，令w代表一次抽样得到的词随机变量

其中和的先验分布是 Dirichlet分布：

这样第m篇文章的生成概率分布为：

我们需要通过已有的大量语料库来推断文章主题分布和主题词汇分布，即推断分布中的参数。由于边缘分布过于复杂，依靠极大似然方法求解困难，可以用Gibbs抽样的方法和EM算法来近似求解。

根据得到的文章主题分布和主题词汇分布，我们可以分析出每篇文章描述的主题，以便进一步分析。

## 模型的创造性实践

在适用于中文史料文本之标记式主题模型分析方法研究中，作者开发了加入标记改进了LDA主题模型，并加入种子词提升主题的群聚结果 (陈奕安，2016) 。

# 序列联配 (Sequence Alignment)

## 概述

在中文文本分析中，有时需要比较多组文本之间的关系。序列联配是一种通过对比两个序列（通常是字符串），判断它们之间的相似度或差异性，在生物，信息学，计算机方面都有应用。

文本也是序列的一种，有了衡量序列间相似度的序列联配的工具，我们可以判断文本之间的关系，例如哪些文本之间较为相似或了解文本间不同的地方在于哪几个字等等，对于文本的比对也有重要帮助。

## 数字人文中的应用

Tariq Yousef（2016）等利用序列联配对古希腊和拉丁语同源史料进行了自动的比对和挖掘。

## 模型方法

最简单衡量两个字符串间距离的方法是Levenshtein距离。Levenshtein距离也称编辑距离，是衡量两个序列之间差异的一种度量。它是衡量一个字符串转换到另一个字符串所需要的最少的单个字符编辑（添加，删除，替换）次数。

Levenshtein距离是通过动态规划算法计算的。其数学定义如下：

设两个字符串a和b，它们的长度分别为。a和b的Levenshtein距离记作

是特征函数，如果则函数值为1，否则为0。

更复杂的序列联配的方法有Needleman-Wunsch算法。它实际上是对Levenshtein距离的一种推广。Needleman-Wunsch算法也是用动态规划算法，对字符串的匹配进行打分。首先，需要选择一个打分系统，即定义匹配，未匹配，插入和删除这四种情况的分数，该分数可以先验给定也可以通过训练给出。匹配即为该位置对应的两个字符相同；未匹配即为该位置对应的两个字符不同；插入和删除即为一个字符对应于一个空缺设匹配分数为a，未匹配分数为b，删除和添加分数为c。递归公式如下：

Needleman-Wunsch算法更广泛的推广有加入空缺惩罚，考虑其他特殊相似度矩阵等。用这些算法，可以在特定的情况下量化字符串与字符串之间的距离，并可以找到字符串之间最佳的对位匹配方式。

## 模型的创造性实践

Tariq Yousef（2016）等应用iAligner对文本进行分析，将Levenshtein距离和Needleman-Wunsch算法进行改进并应用到文本上来。Stefan Jänicke（2018）等人致力于将联配应用到文本上并对文本可视化进行探索。

# 中文分词 (Chinese Word Segmentation)

## 概述

文章有段落句子组成，而句子由词组成，词在语言学上是表达意思的最小单元。汉语和英语不同，英语词与词有空格隔开，而中文没有空格。分词是理解语义进行中文分析的基础，所以中文分词就成了必不可少的一步。

目前中文分词器有许多种，例如jieba分词、THULAC，哈工大LTP，斯坦福分词器和Topwords等，它们对于不同的语料库各有优劣。

## 数字人文中的应用

几乎所有数字人文研究都会使用分词。譬如让计算机理解“晋太原中武陵人捕鱼为业”由晋、太原、中、武陵人、捕鱼、为、业，这些词和短语组成，需要依赖分词算法的自动识别。

## 模型方法

TopWORDS

TopWORDS是由邓柯等（2016）提出的中文无监督分词方法，是一种自上而下的新词发现和分词方法。它的优点是不需要预先指定的词汇或大型相关训练语料库，并且擅长发现新的词汇，这一点对于文言文等古汉语语料显得十分重要。

TopWORDS基于词汇字典模型（WDM）为基础。设是研究语言的基本字符集合，在中文中就是在文本中出现所有不同的汉字。一个词w定义为A中元素组成的序列，即。令是研究文本的词汇表（字典）。 WDM认为每个句子S（以及整个文本）是从D对于词以采样概率随机抽取的单词的拼接。表示词的使用概率的向量，，从WDM模型生成一个K个词（已分词的）句子的概率如下：

对于未分词的中文文本T，我们让表示在字典D中对应于T的所有可能的分割句子的集合。那么我们有：

和条件概率，

它衡量在WDM下T可能被分割成S的可能性。T的最大似然（ML）分割因此定义如下：

TopWORDS采用“自上而下”策略进行词语发现。它由一个超完备的字典D开始，它由长度不大于并且在目标文本中的频率不小于（和是用户指定的阈值）的所有字符串组成。词典D中每个词分配一个使用频率参数。TopWORDS通过EM算法来得到的最大似然估计。另外，因为最初的大词典D中包含许多非词和复合词，所以许多估计的θ值是零或非常接近于零，故我们可以在每次迭代中删除使用频率足够接近于零的词，将D缩减为更小的字典。

在EM算法结束时可以对保留的词做进一步的排名。设是由TopWORDS基于未分割文本T得到的最大似然估计。对于每个，我们定义并且用模型和模型之间的似然比统计量的对数来计算重要性分数:

大的意味着在用WDM拟合目标文本时，有统计意义上的重要性。根据重要性分数，候选词中不具有统计显著性的可以被识别并剔除。另外，在一些文本分析工作中，K个不同目标文本的比较和对比是有意义的。为了突出每个目标文本的具体内容，我们还可以使用估计的“相对频率”

作为排名标准，即，与 “平均”背景文本相比，某个目标文本丰富词的程度。这里，是来自第k个目标文本的词的估计使用频率。我们发现的新词会按重要性分数或相对频率排列。

此外，TopWORDS还可以和其他方法相结合，作为一些文本分析手段的输入部分，例如主题模型、词嵌入方法等。有强大的先验知识也可以加入到TopWORDS的框架中来。

Jieba分词

Jieba分词是国内很常用的中文分词工具，该分词方法和TopWORDS分词都属于概率语言模型分词，即在切分所得的所有结果中求某个切分方案S，使得切分方案的概率P(S)最大。jieba自带了一个内有有2万多条词的词典, 包含了词条出现的次数和词性。它先由自带词典生成字典树，基于字典树结构实现高效的词图扫描，生成DAG。其中DAG记录了词的开始位置，值是可能的结束位置。进一步采用了动态规划查找反向最大概率路径, 找出基于词频的最大切分组合。对于词典中没有出现过的词，建立隐马尔科夫模型，四个状态分别为词的开始、词的中间、词的结尾和单个字成词。通过viterbi算法找到最大概率的状态路径，进而进行切词。另外，jieba分词还可以自定义词典以及根据TFIDF或TextRank算法进行关键词的抽取。

THULAC

李中国等（2009）提出了THULAC中文分词和标注算法。他们从标点符号得到启发，发明了一种在带有标点符号的大语料库中学习中文分词模型的方法，显著解决了未登录词的问题。

首先其定义了四个标签：L，中文词的第一个字；R，中文词的最后一个字；，中文词中除了第一个字剩下的字；，中文词中除了最后一个字剩下的字。那么，词的开始、中间、结尾、单个字构成的词分别定义如下：

对于一句话和b,m,e,s组成的标签序列有：

其中是周围的四个字。我们需要找到最大概率的合法分割。

用最大熵的对条件概率建模：

其中二值函数的十个特征如下：

在条件独立的假设下，条件概率为

类似地，可得m,e,s的条件概率。

一个字单独为一个词只有当如下情况发生时：

用100亿例子进行训练，后在MSP和PKU的数据上进行评估。该方法对于有很多新词的情况提升显著，而且倾向于多词的长表达式切分成组成它的词。

# 主成分分析（Principal Component Analysis）

## 概述

主成分分析方法是最重要的数据降维方法，在尽可能不损失信息的情况下，找出数据中的最主要方面来代替原来的数据进行分析，用少数几个综合指标来代替原来的大量指标。

## 数字人文中的应用

Táňa Dluhošová（2015）研究47件二战后的定期刊物，其中共有563位作者。对该数据进行主成分分析，分析可得第一主成分表现了作者的文学风格，而第二主成分基本反映了写作力量。根据前两主成分，进一步对作者群体进行基于文本的分类。

## 模型方法

主成分分析方法在空间上的解释为把样本点投影到一个超平面且保证这些样本点在这个超平面上尽可能的分开。

假设m个n维数据都进行了中心化。我们打算将数据从n维降到n‘维。设W是经过投影变换后的坐标。那么，样本点在新空间超平面的投影为。样本点的方差可以表达为。样本点在超平面上投影尽可能的分开即投影点方差尽可能的大，故优化目标为：

由拉格朗日乘子法得,故需要对进行奇异值分解，得到特征值为。其特征向量即为所求的

除了主成分分析，还有多种其他方法可以对数据进行降维，如因子分析等，这里不再一一赘述。

# 聚类算法（Cluster）

## 概述

聚类是研究分类的一种统计方法，在事先不知道具体类别特征的情况下，通过研究对象多方面属性后，将相近或相似的对象归为同一类，以便进一步对同一类的研究对象进行深入研究。

## 数字人文中的应用

在王平（2016）在对五代北宋山水画的研究中，作者将所有的诗词主题进行分层聚类，通过分层聚类树状图，可以清楚的展示出山水画不同主题之间的联系，以及进一步对主题进行分类。

Táňa Dluhošová（2015）从所有期刊的作者考虑写作力量和文学风格两个维度进行聚类。通过对人文数据进行聚类，Dluhošová 展现了二战后四个政治和文学主张不同的作者群体。

聚类同样可以用于抽象的的关系网络图：罗盘针（2016）在对台湾新闻的跨性别研究中，对于相关的网络词汇进行聚类。对于每个类别，研究其中的核心词，进而研究实体概念与抽象概念之间的关系。

## 模型方法

聚类算法有很多种，最经典的聚类算法是K-means算法，其他常用的算法还有层次聚类、GMM聚类和基于网络的图团体检测聚类等。

K-means算法：

K-means算法需要先给定一个合适的聚类的类别数K，定义数据点之间合适的度量。计算方法如下：

1. 随机选取K个中心点
2. 遍历所有数据点，将每个数据点归类到最近的中心点中
3. 计算每个聚成的类中点坐标的平均值，将该值作为新的中心点
4. 重复2、3步，不断迭代直到K个中心点收敛

它的优点是思路简单，缺点是需要给出合适的聚类数K，算法收敛较慢，初始中心点的选择可能会对聚类结果有很大的影响。

层次聚类算法：

层次聚类是分层次对数据点进行逐步聚类的算法，其分为自上而下和自下而上两种，最常见的是自下而上的合并算法。

首先要定义距离，定义单个数据点之间，数据点和类之间以及类与类之间的距离度量。例如，在欧几里得距离框架下，点a和点b之间的距离定义如下（i记录的是数据点的维度）：

对于点a和类B（有n个点）之间的距离可以定义如下：

对于类A（有n个点）和类B（有n个点）之间的距离定义如下：

算法步骤如下：

1. 计算已有数据点（或类）和数据点（或类）之间的距离
2. 找到最小的距离，将它们合并成一类
3. 重复1、2直到达到期望的类数或期望的类间距。

它的优点是可以根据实际情况适当的选择聚类类别，但是在数据点过多、数据点维度过大的时候计算效率较低。

高斯混合模型聚类

高斯混合模型聚类算法（Gaussian mixture model），简称GMM聚类算法，是一种基于模型的聚类算法。它最重要的优点是它是一种软分类，即不硬性指出每个点该归入哪一类，而是给出一个概率，即某数据点归入某类的概率。给出概率使得信息量更大，可以把概率接近的类来单独考虑，也可以根据概率来评价分类的好坏等等。

GMM聚类模型的假定是数据点服从混合高斯分布。高斯混合分布由K个成分构成，每个成分都服从高斯分布。也可以理解为先从K个成分中以一定的概率分布抽取某个成分，再根据该成分的分布抽取数据点。概率分布密度函数如下：

其中，都是需要估计的未知参数，K类别数是要事先给定的。其对数似然函数如下：

由于似然函数较为复杂，直接求出解析解是不可能的。我们可以用EM算法来数值求解。首设是一个K维的零一向量且各个维度和等于一，代表的是第i个样本是由混合高斯分布第几个成分生成的。那么，EM算法迭代公式如下：

E步：

M步：

通过EM算法，可以估计出GMM模型的参数。进而，对于每一个样本点，我们可以给出它属于第k个部分的概率：

根据此概率，我们可以判断是否有足够的信心将样本点归入某一类中，也可以通过整体的概率大小评估某个分类结果的好坏。

图团体检测聚类

当样本之间的关系可以表示为一个网络结构时，我们希望找到网络中较为紧密的团体，这也是一种聚类。按照关系的亲密程度将样本划分成不同集合的聚类方法称为图团体聚类。

图团体的定义：一种顶点构成的子集，而且子集中每个点相对于子集外的其他点联系更为紧密。

模块性：衡量图团体划分质量的标准，值越大说明划分的越好。

其中，L表示网络中边的数量，是邻接矩阵的的第i行第j列，指第i个点的度数。

在这种情况下，该问题转化为了优化问题，即找到一种划分方案使得模块性值最大。遍历所有可能性使得时间花销太大，而快速贪婪模块性最大化算法可以解决这个问题。算法流程如下：

1. 每个点自己为一类，计算整体网络模块性
2. 选择任意两个模块聚合在一起造成的模块性改变
3. 取最大的进行融合，用此代替
4. 重复2、3步，记录所有的M直到最终成一类。找到过程中最大的M为需要的聚类模式。

# 支持向量机（Support Vector Machine）

## 概述

支持向量机是一种二分类模型。它的主要思想是将已有的数据看作是空间中的点，寻找一个超平面将空间切开，使得分开间隔最大。

## 数字人文中的应用

Biligsaikhan Batjargal （2015）希望自动识别出蒙古语史文献中的人名。这一需求可以由支持向量机的二分法进行判断：这个词是人名，或者不是人名。Batjargal 用来判断史料中词汇是否为人名而提取的特征包括：词的前后缀，是否在句子的开头，结尾的音节等。

## 模型方法

设有N个样本，，。一个超平面定义为

我们寻找一个超平面能将两部分点最大距离地分隔开。

假设超平面能能够正确分类，那么有

上述两个式子实际上可以合并为一个式子：

所以我们所寻求的被超平面隔开的两部分最大间隔D计算为：

由于寻求最大间隔，该问题可以转化为优化问题：

也可以写做为：

这样，支持向量机的求解问题实际上转化为一个凸二次规划问题，可以使用拉格朗日乘子法将其转化为等价的对偶问题进行求解。

在这里可以发现，优化问题只和向量x之间内积有关。这样，将替换为核空间，可以使得模型更加灵活，适应范围更广。核可以取多项式核

也可以取最常见的高斯核：

我们可以视不同的情况挑选不同的核，这也称作为核工程。

# 词夹子方法（Word-Clip）

## 概述

词夹子的方法是为了解决专有名词的分类问题。例如，如何判断史料中的名词是人名、官名、事件、地点哪一类。

词夹子方法的原理是：特定名词出现时，他的前或后总会有特别的词和它一起出现，例如：“派”某人“去”某地，或者“是”哪里“人”等。如果我们找到这样固定搭配的词，就可以将其作为词夹子，把我们需要的词夹取出来。

## 数字人文中的应用

张尚斌（1995）提出的词夹子的方法应用的实验数据是历史资料中的明清档案和古契书，通过词夹子方法找出史料中的人名和地名，该算法成功的自动识别了史料中75% （recall）的潜在人名和地名。

杜协昌（2015）对词夹子进行了简化，并用此算法对红楼梦中的人名和西游记中的兵器名称进行自动识别。

## 模型方法

词夹子的定义分为四类：前文、词首、词尾和后文。前文、后文指候选词之前的几个字，词首、词尾指候选词内的第一个和最后一个字。这四个字词组成的四元组称作中间候选词的词夹子。

词夹子演算法步骤：

1. 在文章里选定一些专有名词作为样本词。
2. 针对已有的专有名词在文章中出现的位置，前后找它们的词夹子。
3. 把已有的词夹子带入到文章，找这些词夹子出现位置夹到的词，把它们都当作候选词。
4. 给候选词的挑选打分，评分规则，例如，，R表示该词夹子所夹到的样本集的个数；T表示该词夹子一共夹到词的个数。若分数大于M，令M等于本回合的分数。
5. 把前n个候选词当作是正确的专有名词。
6. 重复2-6步，直到没有任何候选词高于M。

在识别特殊名词的时候，例如人名，可以加入头衔、百家姓等先验信息，并过滤掉一些错误信息，可以提高算法的准确率。

# 词嵌入方法（Word Embedding）

## 概述

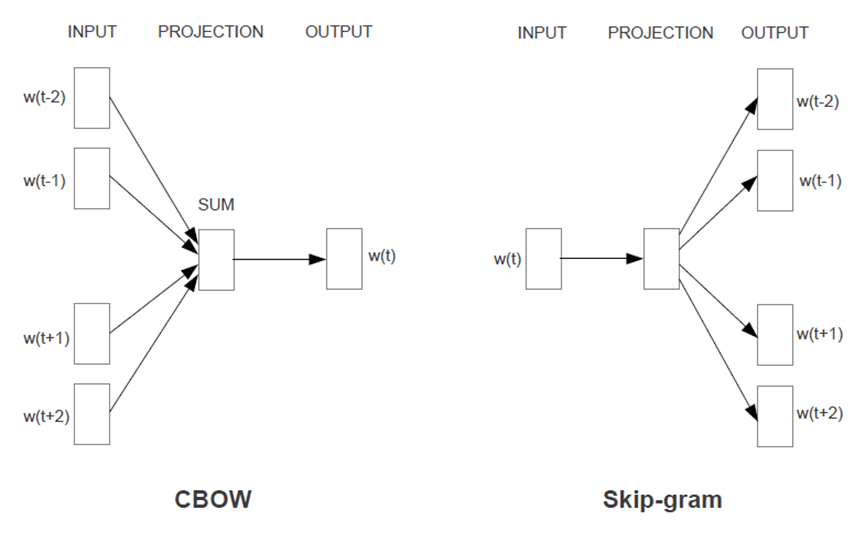
词嵌入是一种把词表示为词向量的方法，用向量中的数字化信息表达词汇中的信息。把文字词汇表示成这种数字向量的形式的一大优点方便后续的量化分析。比如词向量可以精确的计算出国王-男人+女人=女王。

这种把文字信息转化为数字信息的方法很大程度上提取了词汇之间的信息。和传统以孤立的词，或者有限的若干词汇为中心的分析方法相比，词嵌入在分析中可以引入丰富的语义场中的语意信息。

## 模型方法

词的向量表示分为独热表示和分布式表示，独热表示是最直观的表示方法。先构建文本的词库，表示向量的维度就是词库的大小。词向量中只有该词对应的维度上为1，其他维度都是0，例如我表示为，其中第四个维度代表的是我，所以我的词向量中第四维是1，其他都是0。这种表示简单直观，但是缺点是它忽略了词与词之间的关系，并且表示成的向量维数过大，会造成维数灾难。

另外一种方法就是词的分布式表示，将词表示为实数向量，在这方面最常用的方法是2013年T Mikolov等人提出的word2vec方法。该方法主要思路是根据大量的语料库，对于每个词结合上下文的信息训练神经网络进行词的向量表示。Word2vec方法主要分为两个模型：CBOW和skip-gram。CBOW即连续词袋模型，用目标词前后t个词的连续式分布性来表示该目标词，而skip-gram正好相反，是用当前词来预测它前后t个词的分布式表示。用语料库的数据对神经网络进行训练，得到神经网络的权重，而这个神经网络相当于一个编码器，把每一个词汇转化为一个实数向量。两种方法结构如下：

CBOW结构基于上下文预测当前词，而Skip-gram结构通过当前词预测其周围的词

这种方法既能够降低词向量维度，又可以捕捉到当前词的上下文关系，是最常用的词嵌入的方法。

# 社会网络分析 (Social Network Analysis)

## 概述

社会网络分析是综合运用图论和数学模型来研究社会行动者及其之间关系的一种研究方法。

社会网络分析重点关注人和人之间的联系、社会群体里的活跃人物、社会群体的分类（流派）等。研究者会从行为者之间互相影响的角度对网络结构进行分析。

## 数字人文中的应用

赵思渊（2016）构建了以契约权利受入人出人为核心的关系网络，画出并研究了歙县枧桥江氏地权交易网络，通过网络分析可以找到交易网络中的中心人物。

赵薇（2015）以《大波》三部曲为例，提取了其中人物的社会网络关系，并计算网络的关系强度和人物中心性，对人物的中心性进行排序，从人物关系的角度对李劼人从写人物到“写网络”的叙事学探索进行了宏观的研究。

李宗翰（2015）分析了范成大的《吴郡志》中的人物关系，研究其中九老会成员、无士官经历者与范成大社会关系，及其和龚颐正、龚明之家族之间关系。

祝平次（2015）以朱子语录的记录者为研究对象，对朱子语录的记录者和朱熹门人这两个群体进行了对比，并对语录中共同记录者的社交网络进行了研究。

郑文惠（2015）团队从新青年杂志中提取关键词建构词汇网络。在词汇网络中，比较“世界”和“今日”这两个观念在不同时期的区别。

Peter M. Broadwell（2015）对韩国流行音乐人构建网络，从中探索了1990到2015间韩国流行音乐团体的崛起过程，并与日本流行音乐团体进行了对比。

## 模型方法

社会网络分析的两大要素是：行动者（在网络中用点表示）和关系，可以是友谊、书信交流、批判等等，单向或双向（在网络中用线表示）。社会网路分析的具体步骤如下：

1. 明确定义研究对象和网络边界、关系维度。
2. 按照需求收集整理数据
3. 建立关系矩阵，用数学模型或网络分析工具处理分析数据，可视化等。
4. 分析结果，解释结论。

社会网络分析的数学基础是图论，下面介绍一些图论中的定义。

图由点和线组成，在社会网络中指行动者和关系。

线分为有向线和无向线，对应于两种图：有向图和无向图。多重图指两个点之间可以有多个线相连。赋权图指的是每条边都有一个数值记录边的重要程度或关系的远近成程度。

点之间的距离定义为从一点到达另一点需要经过的线的个数。如果两个点之间有线相连接，那么称他们是邻接点，即距离为一的点。

点的度数指连接该点的线的个数，有向图点的度数分为出度数和入度数，分别指该点发出和进入的线的个数。

图的矩阵表示有很多种，最常见的是邻接矩阵。设图点数是p，邻接矩阵A是一个p阶方阵，其中

在社会网络中有一些分析方法，来判断图或网络的一些性质，如中心性分析，凝聚子群。

凝聚子群指当网络中一些行动者之间的联系足够紧密以至于结合成的次级团体。分析网络中子群类型和数量，子群内部和子群与子群之间的特点等的分析称作凝聚子群分析。凝聚子群中最常用的一个概念是派系。派系指至少包含三个点的最大可能的完备图，及派系中所有点都有边相连（完备），向派系中添加任意一点后构成的团体都不是完备的。通过分析派系的大小和数量可以看出研究人群范围中结成小团体的程度，深入研究每个小团体能发现一些特异性特征。

中心性探究网络中的个体或团体居于怎样的中心地位以及网络整体的凝聚程度。某点的中心度反映了该点在网络中的中心程度而中心势展现了网络的集中趋势。

中心度包括点度中心度、中间中心度和接近中心度，计算公式如下：

点度中心度表示与改点直接相连的点的个数：

中间中心性表示该点为中心的媒介程度：

其中代表了jk两点间不同最短路径的数量， 代表了jk两点间不同最短路径且经过i的数量。

接近中心度代表了该点处于整个网络中心的程度：

相应的，图中心势也包括点度中心势、中间中心势和接近中心势，计算公式如下：

图的点度中心势：

图的中间中心势：网络的中间中心势越高，表示该网络中的节点可能分为多个小团体而且过于依赖某一个节点传递关系，而该节点在网络中处于极其重要的地位。

图的接近中心势：对一个社会网络来说，接近中心势越高，表明网络中节点的差异性越大，反之，则表明网络中节点间的差异越小。

进行社会网络分析时，有多种社会网络分析时工具可以使用，例如 Gephi, UCINET等。

# 地理信息系统 (Geographic Information System)

## 概述

地理信息系统（GIS）是在计算机系统支持下对地球空间分布数据进行采集，管理和分析的系统，它将地理、地图、计算机等多个领域结合起来，对空间信息进行分析处理。

GIS方法可以将不同类型数据在同一张地图上叠加，通过跨时间、跨地域的对比为研究带来新的视野。

## 数字人文中的应用

廖泫铭（2015）提出人文信息要进行地理编码，整合各种平台信息包括环境、城市、经济、交通、社会、历史人口和水利为一起，进行时空对位分析。

祝平次（2015）在朱子语料的分析中对比了记录者和朱熹门人地理分布的不同。

# 数据库提升 (Database Upgrade)

## 概述

进行文献研究和文本分析非常重要的一点是资料库的构建，如何构建一个易于存储调用、查询、计算的数据库是进行后续分析的基础。计算机上有一些方法可以对数据库的结构进行提升。目前最流行的一种数据结构是关联型数据库。关联型数据库简单来说是把不同的数据集建立联系，将数据库中相同的对象对应起来，是一种结构化数据的方法。

## 数字人文中的应用

徐晨飞等人（2015）对东亚楚辞文献构建了数据库，将楚辞中出现的专有名词按类别进行分类，如人物名、地理名、植物名等，构建关系型数据库将专有名词与相应的专业知识相对应，各类专有名词又有关系相连接。例如，某人（人物名）的出生地属性和地理名相关联，该地（地理名）又盛产某种植物（植物名）。

谢博宇等人（2016）基于DocuSky数据库进行研究。DocuSky是专门让数位人文学者进行文本研究的平台。DocuSky将文本与工具分离，减少系统维护且工具可以共享流通，还以浏览器作为分析工具的主要环境，易于散步和修改且便于开发。

资料来源：

6th International Conference of Digital Archives and Digital Humanities 2015

7th International Conference of Digital Archives and Digital Humanities 2016

Digital Humanities 2018

# 参考文献

[1] Biligsaikhan Batjargal ,Garmaabazar Khaltarkhuu,Akira MaedaPersonal.Name Extraction from Mongolian Historical Documents Using Machine Learning. 6th International Conference of Digital Archives and Digital Humanities 2015 [C];2015.

[2] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[3] Deng K, Bol P K, Li K J, et al. On the unsupervised analysis of domain-specific Chinese texts[J]. Proc Natl Acad Sci U S A, 2016, 113(22):6154-6159.

[4] 杜协昌. 半自动词汇撷取：简化的词夹子方法以及其JavaScript 元件的开发与应用[C].第六届数位典藏数位人文国际讨论会;2015.

[5] 李宗翰，郑莉. 范成大《吴郡志》中的社会关系网络-以CBDB于Pajek作为分析工具[C]. 第六届数位典藏数位人文国际讨论会;2015.

[6] 廖泫铭. 从历史GIS朝向空间人文学发展[C]. 第六届数位典藏数位人文国际讨论会;2015.

[7] 罗盘针,郑硕,江安琪,曾博扬.以语料库分析取径探究台湾新闻中的跨性别[C].第七届数位典藏数位人文国际讨论会;2016.

[8] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[9] Peter M. Broadwell, Timothy R. Tangherlini, Hyun Kyong Hannah Chang. From Trot to Cultural Technology: The Historical Development of Production Networks in Korean Popular Music and Beyond. 6th International Conference of Digital Archives and Digital Humanities 2015 [C];2015.

[10] Piing-tzu Chu. Text VS. Database: A Study of the Zhuzi yulei. 6th International Conference of Digital Archives and Digital Humanities 2015 [C];2015.

[11] Táňa Dluhošová, The Literary Landscape of Taiwan in the Early Postwar Period-A Quantitative Digital Humanities Perspective; 6th International Conference of Digital Archives and Digital Humanities 2015 [C];2015.

[12] Tao Wang, Text Mining Deutsche Textarchiv Using Digital Tools and More; 7th International Conference of Digital Archives and Digital Humanities 2016 [C];2016.

[13] Tariq Yousef, Intra-language Text Alignment using iAligner; 7th International Conference of Digital Archives and Digital Humanities 2016 [C];2016.

[14] Zhongguo Li, Maosong Sun. Punctuation as implicit annotations for chinese word segmentation[M]. MIT Press, 2009.

[15] 王平, 五代北宋山水画的数位人文研究——以「渔隐」主题为例; 第七届数位典藏数位人文国际讨论会[C];2016.

[16] 谢博宇、傅莉雯. 以DocuSky為核心的工具開發與配置[C]. 第七届数位典藏数位人文国际讨论会;2016.

[17] 徐晨飞，钱智勇，端木艺，周建忠. 东亚楚辞文献数据库建设及语义化应用研究进展综述[C]. 第六届数位典藏数位人文国际讨论会;2015.

[18] Yi an Chen,An enhanced topic model based on labeled LDA for chinese historical corpora. 7th International Conference of Digital Archives and Digital Humanities 2016 [C];2016.

[19] 张尚斌. 词夹子演算法在专有名词辨识上的应用, -以历史文件为例[J]. 台湾大学资讯工程学研究所学位论文, 2006.

[20] 赵思渊. 地方历史文献的数位化、数据化与文本挖掘： 以《中国地方历史文献数据库（上海交通大学馆藏地方历史文献数据库）》为例[C]. 第七届数位典藏数位人文国际讨论会;2016.

[21] 赵薇.“社会网路分析”（SNA）在现代汉语历史小说研究中的应用[C]. 第六届数位典藏数位人文国际讨论会;2015.

[22] 郑文惠，邱伟云，刘昭麟，林书佑.概念关系的数位人文研究：以《新青年》中的「世界」观念为考察核心[C]. 第六届数位典藏数位人文国际讨论会;2015.