Computer Science and Application 计算机科学与应用, 2023, 13(4), 745-753
Published Online April 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.134073

Hans汉斯

基于马尔可夫动态编码的谷歌图书语料库质量方法

宋玉玲

温州大学计算机与人工智能学院，浙江 温州

收稿日期：2023年3月15日；录用日期：2023年4月13日；发布日期：2023年4月20日

摘 要

语料库是自然语言处理任务的关键，谷歌图书语料库是迄今为止最大的历时语料库，被广泛应用于从时间、空间维度上评估学科、语言甚至是文化等领域在社会发展中的现象和规律，但因其构建过程中的识别问题、元数据问题等原因被很多学者质疑。目前常见的处理方法主要是从语料库中提取所有可能的数据和从原数据进行预处理，这些方法耗时且费力。本文提出将语料库噪声问题转化为时间序列异常检测问题，使用传统的时间序列模型和马尔可夫动态编码去实现时间序列异常检测。实验结果表明，马尔可夫不仅可以保存时间相关性和频率结构，而且提供了一种自然的反向操作——将图形映射回时间序列，克服了传统时间序列模型的缺点，最终有效地解决了语料库的局部质量对齐问题。

关键词

谷歌图书语料库，马尔可夫模型，时间序列异常检测

Google Books Corpus Quality Method Based on Markov Dynamic Coding

Yuling Song

College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang

Received: Mar. 15th, 2023; accepted: Apr. 13th, 2023; published: Apr. 20th, 2023

Abstract

The corpus is the key to natural language processing tasks. The Google Books corpus is by far the largest ephemeral corpus, which is widely used to evaluate the phenomena and patterns of disciplines, languages, and even cultures in social development from temporal and spatial dimensions, but

文章引用: 宋玉玲. 基于马尔可夫动态编码的谷歌图书语料库质量方法[J]. 计算机科学与应用, 2023, 13(4): 745-753.
DOI: 10.12677/csa.2023.134073

it has been questioned by many scholars due to the identification problem and metadata problem in its construction. The current common processing methods mainly extract all possible data from the corpus and pre-process from the original data, which are time-consuming and laborious. In this paper, we propose to transform the corpus noise problem into a time series anomaly detection problem by using the traditional time series model and Markov dynamic coding to achieve time series anomaly detection. Experimental results show that Markov not only preserves temporal correlation and frequency structure, but also provides a natural inverse operation—mapping graphs back to time series, which overcomes the shortcomings of the traditional time series model and finally effectively solves the local quality alignment problem of the corpus.

Keywords

Google Books Corpus, Markov Model, Time Series Anomaly Detection

1. 引言

大规模的数据是可靠分析的基础。谷歌制作的最大的历时语料库促进文化组学这一新型领域的发展，它是一种定量分析长期文化变化的新工具，通过对关键词在语料库中的使用频率变化，展示五百年来人类文化发展史思想和文化中鲜为人知的趋势和现象。然而，其前期数据质量较差，人们通常会选择 1800 年之后的数据进行研究，目前对于语料库的处理方法耗时且费力，因此本文提出将语料库噪声问题转化成时间序列异常检测问题来处理，进一步提高数据质量，从而使得五个世纪以来的数据得到更充分的使用。

2. 相关工作

近年来，谷歌公司与大学图书馆合作将世界各地的从古至今的出版物通过光学字符识别(Optical Character Recognition, OCR)技术进行数字化，建立了涵盖多种语言的谷歌图书语料库(Google Books Ngram Corpus, GBNC) [1]，并发布了三个版本的 n-gram 数据集，分别为 2009 年 7 月(第 1 版)、2012 年 7 月(第 2 版)和 2020 年 2 月(第 3 版)。第一版的语料库是谷歌公司根据 OCR 和书目元数据的质量扫描了超过 500 W 万册书籍(5000 亿词)，约占迄今为止出版的所有书籍的 4%；第二版的谷歌图书语料库在建立过程中采用了更先进的数字化技术，包含了更多的标题、改进的 OCR 和修正的元数据等，同时增加到超 800 万本书籍(8000 亿词)的数据，约占迄今为止出版的所有书籍数量的 6% [2]。本文主要研究第三版本的语料，它涵盖 8 种语言，分别是英语、西班牙语、法语、德语、俄语、意大利语、中文和希伯来语，其中英语占据主导地位，它是当今世界上最大的历时语料库。

基于图书版权等原因，谷歌公司制作出谷歌图书词频统计器(Google Books Ngram Viewer, GBNV)来方便人们进行研究。它可用来分析 5 个世纪以来单词或词组的使用频率，是一个交互式地量化分析语言变化趋势的便捷工具。当用户进行搜索时，可以从多种语言中选择，对于英语，他们可以区分英语、美式英语、英式英语和英语小说。通常，X 轴代表语料库中作品的发表年份，Y 轴代表 n-gram 在整个语料库中出现的频率，用户输入 n-gram，然后可以选择区分大小写、日期范围、语料库语言和平滑等，即可得到对应的词组的语用频率。

Michel 等[1]首次将 GBNC 研究成果发表在世界顶级期刊《科学》杂志之上，其借助谷歌图书的海量

数字化资料，通过对关键词在语料库中的使用频率变化，分析了公元 1500 年到 2000 年间 500 多万本书籍语料库，展示了五百年来人类文化发展史中鲜为人知的趋势和现象，集中展示了其对于社会人文学科而言巨大的研究价值。这种通过电子化文本的量化分析研究人类行为与文化趋势的计算词典学方法的研究领域，被称为“文化组学(Culturomics)”，其迅速成为学者关注的焦点。

研究者对海量数字档案进行数据挖掘以研究人们使用的语言与词汇，进而揭示其中反应出来的文化现象。Twenge 等[3] [4]报道了美国书籍中个人主义词汇和短语出现频率的增加。Kesebir 等[5]记录了道德用语使用频率的下降，而 Twenge 等[6]则强调了脏话使用频率的增加。此外，Greenfield [7]的发现表明，在美国和英国，个人主义的增加和生态变化之间存在关系，这源于美国和英国英语语料库中的词频变化。在国内，文化组学的研究也如雨后春笋般涌现，如在中国这样的集体主义社会，通过追踪民间信仰或人称代词的概念，个人主义也在上升[8] [9]。邵斌[10]以浙江文化关键词为研究对象所展现的词频变动轨迹去探究浙江文化在英语世界中的影响力。曾凡斌等[11]通过传播学学科的关键词首次利用 GBNC 这一大数据分析工具分析传播学学科的发展。陈云松[12]通过社会学的关键词的词频研究 19 世纪以来的社会学发展趋势，并对社会组学建立展望。

尽管 GBNC 基于数据量之大、时间跨度范围之广、搜索功能之强大的特色被越来越多的研究人员接受，但也有学者质疑其结果的可靠性[13]-[20]，主要的批评包括 OCR 不足、缺乏代表性及错误的元数据。首先，此项工作是由不同的大学合作完成的，OCR 的差异导致了数据质量不佳，其中大部分是由于古籍印刷质量不佳所致，例如，字母 s 在古代英语书籍中经常被认作 f，在 17 世纪所扫描的书籍中，大约有一半被错误地识别为 beft，随着谷歌对于设备的升级，在 21 世纪的当代书籍中，错误率为 0.02%。其次，整个语料库文献的比例及单个作者可能用特定的词和短语严重影响数据集的准确性引起学者对于代表性的质疑。最后，书籍的规格、出版时间等元数据问题也引发争议。

如果依靠人力去阅读这些数据去分析文化组学这个问题可能需要一个人几千年才能完成，谷歌所完成的相对词频的分析暗示文化变化的各个方面，语言中特定单词和短语的流行度可能会因多种原因而发生变化，包括技术背景(如，计算机)，并且某些单词的含义会随着时间的推移而发生深刻变化(如，gay)。尽管如此，在大量单词中，频率的变化模式可能在一定程度上反映了人们感受和看待世界的方式的变化，因此我们认为 GBNC 仍是一种对专家和非专家都有利的语言趋势检测的工具。目前，较火的两种处理 GBNC 方案：第一种是从语料库中提取所有可能的支持数据，如不仅考虑单个单词，还要研究其他形式甚至是同义词的行为[21] [22]，研究不同语料库(普通英语、英式英语、美式英语)中含义相同的术语[3]；第二种是对原数据进行预处理，如删除所有非单词的标记(字符串) [23]等。上述两种处理方式耗时且费力，且纠正所有的错误是一件不可能的事情，因此本文提出将语料库噪声问题转化为时间序列异常检测问题来解决局部质量对齐问题。

3. 时间序列分析方法

时间序列 (Time Series, TS)是一组按照时间发生先后顺序进行排列的数据点序列，时间序列分析是将数据文件中某一变量的观测值按时间顺序排列成一个数值序列，时间顺序可以是任何单位的，如时、日、周、月、年等，展示研究对象在一定时期内的变动过程，分析事物的变化特征。本文主要从以下两种方式来进行时间序列分析：一种是建立相应的数学模型，如指数平滑法、整合移动平均自回归模型(Autoregressive Integrated Moving Average Model, ARIMA) [24]等，描述系统的时序状态；另一种是可视化时间序列[25] [26]，重新构造数据，提取特征并转化为复杂的网络，在保持时间和频率动态的同时，将信息映射到原始时间序列不要使用空格、制表符设置段落缩进，不要通过连续的回车符(换行符)调整段间距。

3.1. 传统时间序列模型

简单指数平滑是在考虑所有的数据的时候给数据赋予不同的权重，对时间点 $t+1$ 的预测值 y_{t+1} 是 t 时的观测值 s_t 和 t 时预测值 y_t 的加权平均值，数学表达如公式(1)，其中 α 为平滑参数。

$$y_{t+1} = \alpha s_t + (1-\alpha) y_t \tag{1}$$

每个时序数据集都可以分解为相应的几个部分：趋势、季节性和残差。任何呈现某种趋势的数据集都可以用霍尔特线性趋势算法来进行预测，该算法由预测函数和两次平滑函数组成，一个是水平函数，一个是趋势函数 b_t，以及平滑参数 α 和 β。L_t 是观测值和样本内单步预测值的加权平均数，b_t 是根据 $L_t - L_{t-1}$ 和之前的预测趋势 b_{t-1} 在时间 t 处的预测趋势的加权平均值，预测函数就是将这两个方程相加得到的，数学表达公式如(2)、(3)和(4)：

$$y_{t+h|t} = l_t + hb_t \tag{2}$$

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{3}$$

$$b_t = \beta * (l_t - l_{t-1}) + (1-\beta) b_{t-1} \tag{4}$$

ARIMA (p, d, q)模型是将观察对象随时间拖移而形成的时间序列看作一个随机序列，用一定的数学模型去近似描述这个序列。它是自回归模型(Auto Regressive, AR)、Integrated 和移动平均模型(Moving Average, MA)的结合，将序列中的下一步预测结果为历史的观测值和残差的线性函数。AR 描述当前值与历史值间的关系，它适用于没有趋势和季节性的时间序列，预测与历时数据相关的时间序列即平稳的时间序列。一般 p 阶自回归模型 AR 数学公式定义如(5)：

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \varepsilon_t \tag{5}$$

Integrated 用来在预测非平稳时间序列时，对时序数据进行 d 阶差分化，使得时间序列变平稳，即数据值被相邻观测值的差值替代；MA 关注的是自回归模型中的误差项的累加，它将观测值与应用于滞后观测值的移动平均模型的残差之间的相关性合并，参数 q 表示预测模型种预测误差的滞后数。一般 q 阶移动平均模型 MA 数学公式定义如(6)：

$$y_t = \mu + \sum_{i=1}^{p} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{6}$$

自回归移动平均模型 ARMA 是自回归 AR 与移动平均 MA 的结合，将序列中的下一步预测结果为历史的观测值和残差的线性函数。一般自回归移动平均模型 ARMA 数学公式定义如(7)：

$$y_t = \mu + \sum_{i=1}^{p} \theta_i \varepsilon_{t-i} + \varepsilon_t + \sum_{i=1}^{p} \gamma_i y_{t-i} \tag{7}$$

ARIMA 是将非平稳的时间序列进行差分，计算相邻观测值间的差值，在将数据带入 ARMA 公式中。

3.2. 马尔可夫模型

马尔可夫转移场(Markov Transition Fields, MTF)是一种突出时间序列行为的可视化技术，将时间序列可视化为基于一阶马尔可夫过程和时间顺序的复杂网络，更好地保存了时间相关性和频率结构。一般来说，复杂的真实世界序列数据很难通过可视化来观察他们的特征，先要对嵌入在时间序列中的时间动态进行深层次的分析，一种可能的解决方案是重新构造数据，提取特征，以便更好地对时间依赖性进行视觉编码。在本文中，我们构造马尔可夫转移矩阵，将提取得到的特征转化为复杂的网络，并且保持时间

和频率的动态，同时存在一个反向操作将信息映射回原始时间序列。

首先，将时间序列量化。给定一个时间序 $X=\{x_1,x_2,\cdots,x_n\}$，将时间序列离散化，实行数据分箱 $Q\{q_1,q_2,\cdots,q_Q\}$。

其次，建立马尔可夫转移矩阵。把一连串的数据 x_i 转换为不同等级的数据 $q_j\left(j\in[1,Q]\right)$，每个值 x_i 被映射到每个 q_j 中，构造一个 $Q*Q$ 的加权邻接矩阵 W，方法是沿着每个时间步，以一阶马尔可夫链的方式计算箱之间的转移概率。W 的数学公式如(8)：

$$W=\sum_{\forall x\in q_i,y\in q_j,x+1=y}\frac{1}{\sum_{j=1}^{Q}w_{ij}} \tag{8}$$

其中，w_{ij} 是 q_j 中一个点后面跟着 q_i 中一个点的频率，$\sum_j w_{ij}=1$。

然后，建立马尔科夫转移场。MTF 通过将每个概率沿时间顺序排列来扩展马尔可夫矩阵 M，即顺序表示马尔可夫转移概率，以保留时域中的信息。M 数学公式如(9)：

$$M=\begin{bmatrix} w_{ij}\mid x_1\in q_i,x_1\in q_j & \cdots & w_{ij}\mid x_1\in q_i,x_n\in q_j \\ w_{ij}\mid x_2\in q_i,x_1\in q_j & \cdots & w_{ij}\mid x_2\in q_i,x_n\in q_j \\ & \cdots & \\ w_{ij}\mid x_n\in q_i,x_1\in q_j & \cdots & w_{ij}\mid x_n\in q_i,x_n\in q_j \end{bmatrix} \tag{9}$$

为使矩阵 M 进行更高效的计算和可视化，对其采取两种常见的降维技术：分段聚合近似(Piecewise Aggregate Approximation, PAA)和模糊(Blurring)。

最后，提取 M 中的有用信息映射回原始时间序列，即将 MTF 对角线上的转移概率映射回原始时间序列。

4. 实验结果与分析

GBNC 获取的数据是关键词历年的使用频率，可以看作是数据文件中关键词的观测值按年排列成的一个时间序列，针对该语料库中的局部数据质量对齐问题，本文提出了将语料库噪声问题转化为时间序列检测问题。时间序列异常检测(TS Anomaly Detection)的主要目标是从时间序列中识别异常的事件或行为。异常的两个标准：1) 异常数据跟样本中的大多数数据不太一样。2) 异常数据在整体数据样本中占比比较小。目前，常见的无监督机器学习异常检测主要分为基于预测的方法和基于重构的方法。

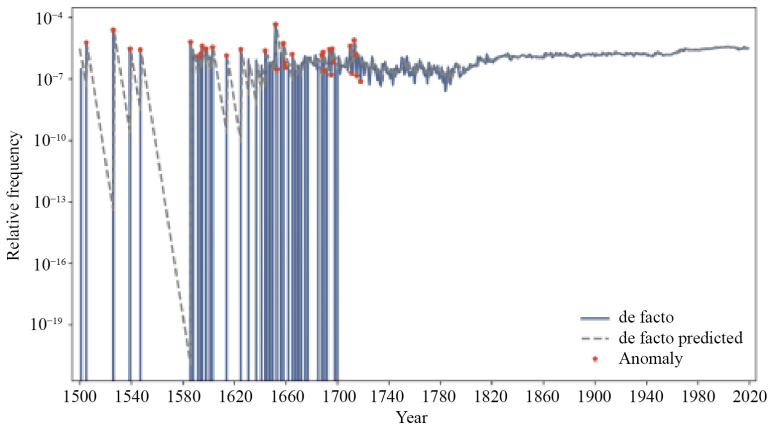
本文使用两种方法来将语料库噪声问题转化为时间序列的异常检测问题：一是使用传统的时间序列模型去预测时间序列，将预测值与实际值的差值作为特征放入到孤立森林里面去实现异常点检测[27]；另一种是可视化时间序列，从而找到偏离正常轨道的信号来达到进一步提高数据质量的目的[26]。

4.1. 传统时间序列

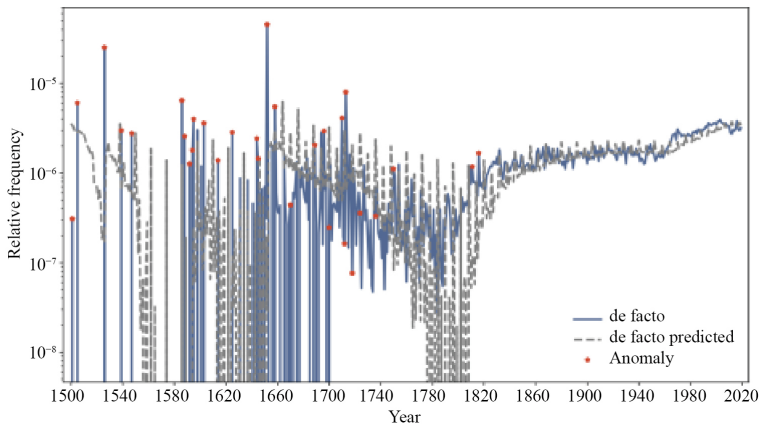
首先，获取数据。GBNC 被广泛应用于文化组学中去研究人类行为与文化趋势，此语料库包含了大量的关键词，如果下载整个语料库需要大量的人力与物力，因此使用爬虫去爬取 GBNC 中的部分关键字的词频。本文爬取 2135 组拉丁词组在英语语料库中的词频，其中 smoothing 设为 0，case_insensitive 设为 true，year 设为 1500~2019，corpus 设为 en-2019。

然后，对时间序列进行预处理，对时间序列进行平稳性检测，对于非平稳的序列通过差分、log 变换或平方根变换转化为平稳序列，使用简单指数平滑算法(图 1(a))、霍尔特线性趋势算法(图 1(b))、ARIMA (图 1(c))三种传统时间序列预测模型作为特征提取器。参数的选择是基于我们的经验和领域知识。这几种模型计算复杂度低，性能良好，已在异常检测文献中得到了验证，如果需要可以添加其他合适的预测模

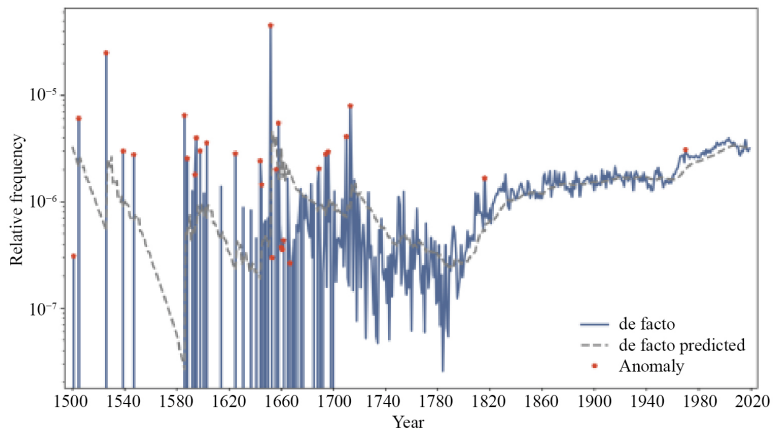
型。对于不同的时间序列预测工具，我们可以得到不同的预测值 p_i，然后预测值 p_i 减去实际值 X_i 并取绝对值，得到时间序列的误差序列 $|p_i - X_i|$，将这个作为数据 X_i 的特征，最后使用孤立森林(Isolation Forest)来对时间序列的特征来做无监督的异常检测。



(a)



(b)



(c)

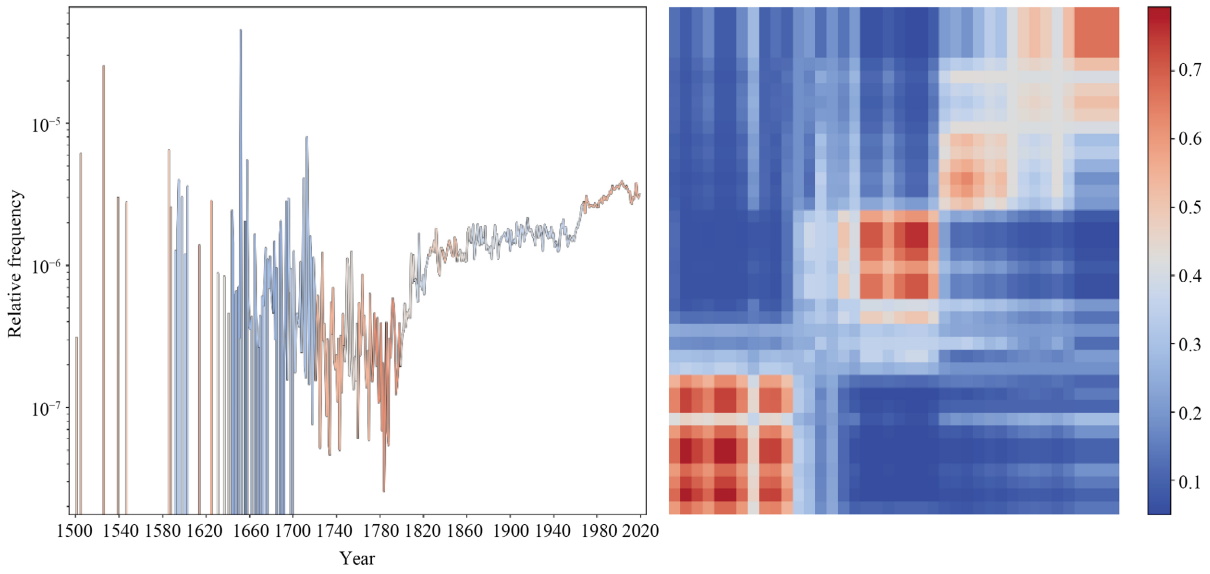
Figure 1. Diagram of “de facto” anomaly detection of traditional time series models
图 1. “de facto”的传统时间序列模型的异常检测图

本文对上述词组的词频进行异常检测，基于版面原因仅展示“de facto”的异常结果，如上图所示，蓝色为实际的数据，灰色为时间序列的预测数据，红色的点为孤立森林算法得到的异常的数据点，异常的数据几乎都出现在前期，但基于以下两点原因传统时间序列模型不是本实验的最好方法：一是传统的时间序列模型在预测时间序列时会产生负值，这与实际意义不符(词频最低为零，不会出现负值)；二是传统的时间序列模型是从前往后去预测，而 GBNC 的数据前期的质量问题最严重，导致后续的研究会出现较大的偏差。

4.2. MTF 异常检测

本文采取 MTF 来进行异常检测，它解决上述实验的问题，重新构造数据，提取特征来进行时间序列分析。首先，使用 Python 包 Pyts 进行时间序列分类，将时间序列离散成不同的值，然后建立一个马尔可夫转移矩阵来计算转移概率。随后，本文将每个概率沿时间轴对齐，来构建“de facto”的 MTF，其中横坐标表示 1500~2019 年，纵坐标表示 1500~2019，颜色表示其自转移概率。

PAA 主要应用在分类和查询任务等展示主要趋势而忽略时间序列的详细结构，而本实验需要保留原始数据中的信息，通常选用对原始信号进行小的信息损失，因此本文使用模糊化来对 MTF 进行压缩从而获取聚合的 MTF (图 2(b))。最后，本文将自转移概率(MTF 对角线上显示的转移概率)映射回原始时间序列，以更好地理解时间序列的性质和行为，“de facto”结果如图 2(a)所示。我们发现，前期的数据其转移概率越高(蓝色偏深)，前期的数据质量较差；后期的数据其转移概率越高(红色偏深)，说明后面的数据质量比较高。



(a) (b)

Figure 2. Diagram of “de facto” anomaly detection of Markov Model
图 2. “de facto”的马尔可夫模型的异常检测图

通过观察整个时间序列的自转移概率，本文 n_bins 为 4，将低于 0.3 的时间点视为异常点，整合 2135 组拉丁词组的异常点，本文发现大部分的异常点都集中在 1500 年至 1800 年之间，其英语语料库中的结果如图 3 所示，这也与上文所提出的学者质疑相吻合。

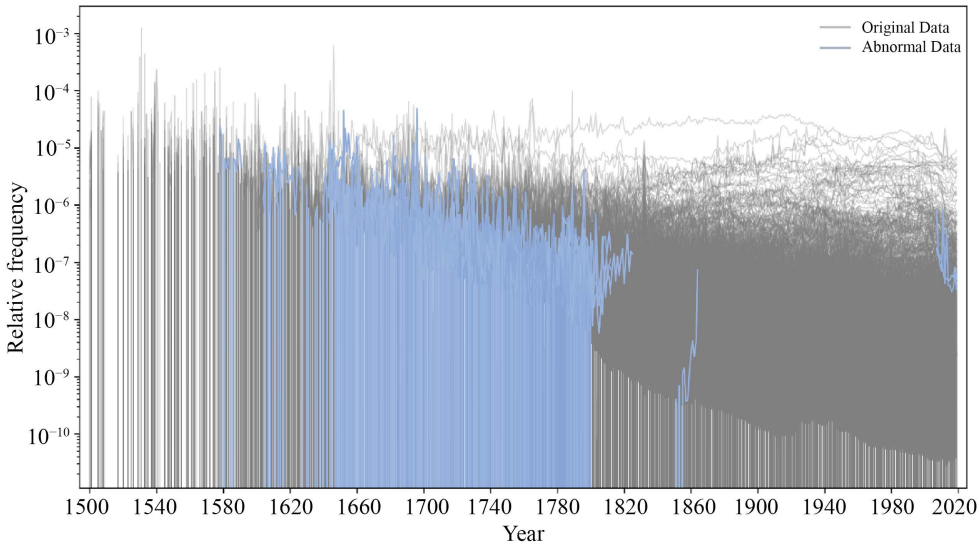


Figure 3. Diagram of all Latin phrases anomaly detection of Markov Model
图 3. 全部拉丁词组的马尔可夫异常检测图

5. 结束语

本文以 GBNC 的学者质疑为出发点，研究改进语料库局部质量对齐的方法，首次提出将语料库噪声问题转化为时间序列异常检测问题。本文通过传统的时间序列模型和时间可视化方法，结合数据本身特点来实现异常检测。最终，本文发现，GBNC 语料库的前期数据即 1500~1800 存在很大问题，异常数据较多，这也导致大多数学者选择 1800 之后的数据进行文化组学的研究，但前期 300 多年的历时数据也是具有很大研究意义的。最后，本文总结出现早期数据质量较差的原因如下：

- 元数据错误。这其中包括错误归属的作者、错误的出版日期、错误的主题分类、错误拼写的标题、作者和出版商等，甚至存在一本书的元数据附加到另一本完全不同的书。
- 扫描错误，包括书籍质量和 OCR 质量。在通过 OCR 转换成数字化文本的过程中，早期的书籍印刷质量较差，且实现较长不易保存，导致可能出现无法阅读、上下颠倒或顺序错误等问题，从而影响扫描质量；此项工程是由多个机构共同承担，它们的 OCR 技术不统一也可能会影响扫描质量。
- 书籍自身问题。书籍规格、题材等存在差异，存在单栏或双栏的格式，也会影响数据质量。
- 书籍出版有周期，时间跨度较大，数据存在噪声。

虽然 GBNC 的早期数据存在问题，但是对于研究历时变化(500 多年的文化变化)具有重大的意义，所以本文不建议直接将早期的数据直接忽略，后续的研究将针对如何处理异常的数据点，从而更好地完成文化组学的研究。在大数据时代，对海量数据的科学定量分析能提供给我们一些原先无法预测到的结果，这必将促进更多领域的发展。

参考文献

[1] Michel, J.B., Kui, S.Y., Presser, A.A., *et al.* (2011) Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, **331**, 176-182. https://doi.org/10.1126/science.1199644

[2] Lin, Y., Michel, J.B., Lieberman, A.E., *et al.* (2012) Syntactic Annotations for the Google Books Ngram Corpus. *Proceedings of the* 50*th Annual Meeting of the Association for Computational Linguistics*, Jeju, 8-14 July 2012, 169-174.

[3] Twenge, J.M., Campbell, W.K. and Gentile, B. (2012) Male and Female Pronoun Use in U.S. Books Reflects Women's Status, 1900-2008. *Sex Roles*, **67**, 488-493. https://doi.org/10.1007/s11199-012-0194-7

[4] Twenge, J.M., Campbell, W.K. and Gentile, B. (2013) Changes in Pronoun Use in American Books and the Rise of In-

dividualism, 1960-2008. *Journal of Cross-Cultural Psychology*, **44**, 406-415. https://doi.org/10.1177/0022022112455100

[5] Kesebir, P. and Kesebir, S. (2012) The Cultural Salience of Moral Character and Virtue Declined in Twentieth Century America. *Journal of Positive Psychology*, **7**, 471-480. https://doi.org/10.1080/17439760.2012.715182

[6] Twenge, J.M., Van Landingham, H. and Keith, C.W. (2017) The Seven Words You Can Never Say on Television: Increases in the Use of Swear Words in American Books, 1950-2008. *SAGE Open*, **7**, 1-8. https://doi.org/10.1177/2158244017723689

[7] Greenfield, P.M. (2013) The Changing Psychology of Culture from 1800 through 2000. *Psychological Science*, **24**, 1722-1731. https://doi.org/10.1177/0956797613479387

[8] Hamamura, T. and Xu, Y. (2015) Changes in Chinese Culture as Examined through Changes in Personal Pronoun Usage. *Journal of Cross-Cultural Psychology*, **46**, 930-941. https://doi.org/10.1177/0022022115592968

[9] Xu, Y. and Hamamura, T. (2014) Folk Beliefs of Cultural Changes in China. *Frontiers in Psychology*, **5**, Article 1066. https://doi.org/10.3389/fpsyg.2014.01066

[10] 邵斌. 浙江文化关键词在英语世界的影响力研究——基于文化组学的视角[J]. 浙江学刊, 2017(2): 201-207.

[11] 曾凡斌, 陈荷. 基于谷歌图书语料库大数据的百年传播学发展研究[J]. 现代传播: 中国传媒大学学报, 2018, 40(3): 135-145.

[12] 陈云松. 大数据中的百年社会学——基于百万书籍的文化影响力研究[J]. 社会学研究, 2015, 30(1): 23-48.

[13] Duguid, P. (2007) Inheritance and Loss? A Brief Survey of Google Books. *First Monday*, **12**. https://doi.org/10.5210/fm.v12i8.1972

[14] Solovyev, V. and Akhtyamova, S. (2019) Linguistic Big Data: Problem of Purity and Representativeness. *CEUR Workshop Proceedings*, Vol. 2523, 193-204.

[15] Solovyev, V.D., Bochkarev, V.V. and Akhtyamova, S.S. (2020) Google Books Ngram: Problems of Representativeness and Data Reliability. 21*st International Conference, DAMDID/RCDL* 2019, Kazan, 15-18 October 2019, 147-162. https://doi.org/10.1007/978-3-030-51913-1_10

[16] Pettit, M. (2016) Historical Time in the Age of Big Data Cultural Psychology, Historical Change, and the Google Books. *History of Psychology*, **19**, 141-153. https://doi.org/10.1037/hop0000023

[17] Pechenick, E.A., Danforth, C.M. and Dodds, P.S. (2015) Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLOS ONE*, **10**, e0137041. https://doi.org/10.1371/journal.pone.0137041

[18] Koplenig, A. (2017) The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets-Reconstructing the Composition of the German Corpus in Times of WWII. *Digital Scholarship in the Humanities*, **32**, 169-188.

[19] James, R. and Weiss, A. (2012) An Assessment of Google Books' Metadata. *Journal of Library Metadata*, **12**, 15-22. https://doi.org/10.1080/19386389.2012.652566

[20] Pechenick, E.A., Danforth, C.M. and Dodds, P.S. (2017) Is Language Evolution Grinding to a Halt? The Scaling of Lexical Turbulence in English Fiction Suggests It Is Not. *Journal of Computational Science*, **21**, 24-37. https://doi.org/10.1016/j.jocs.2017.04.020

[21] Younes, N. and Reips, U.D. (2018) The Changing Psychology of Culture in German-Speaking Countries: A Google Ngram Study. *International Journal of Psychology*, **53**, 53-62. https://doi.org/10.1002/ijop.12428

[22] Younes, N. and Reips, U.D. (2019) Guideline for Improving the Reliability of Google Ngram Studies: Evidence from Religious Terms. *PLOS ONE*, **14**, e0213554. https://doi.org/10.1371/journal.pone.0213554

[23] Bochkarev, V., Solovyev, V. and Wichmann, S. (2014) Universals versus Historical Contingencies in Lexical Evolution. *Journal of the Royal Society Interface*, **11**, Article ID: 20140841. https://doi.org/10.1098/rsif.2014.0841

[24] Ho, S.L. and Xie, M. (1998) The Use of ARIMA Models for Reliability Forecasting and Analysis. *Computers and Industrial Engineering*, **35**, 213-216. https://doi.org/10.1016/S0360-8352(98)00066-7

[25] Wang, Z. and Oates, T. (2015) Imaging Time-Series to Improve Classification and Imputation. *Proceedings of the* 24*th International Joint Conference on Artificial Intelligence*, Buenos Aires, 25-31 July 2015, 3939-3945.

[26] Liu, L. and Wang, Z. (2018) Encoding Temporal Markov Dynamics in Graph for Visualizing and Mining Time Series. *The Workshops of the* 32*nd AAAI Conference on Artificial Intelligence*, New Orleans, 2-7 February 2018, 178-184. http://arxiv.org/abs/1610.07273

[27] Zhao, N., Zhu, J., Liu, R., *et al*. (2019) Label-Less: A Semi-Automatic Labelling Tool for KPI Anomalies. *IEEE INFOCOM* 2019*—IEEE Conference on Computer Communications*, Paris, 29 April-2 May 2019, 1882-1890. https://doi.org/10.1109/INFOCOM.2019.8737429.