近代报刊新闻报道的时间线抽取*

——基于《东方杂志》“大事记”专栏的文献整理试验

□李惠　夏翠娟　侯君明　朱庆华　刘炜

摘要　近代报刊的新闻报道记录并见证了这一特定历史时期的社会世相和众生百态，具有重要的史料价值。然而，从时间的维度，追踪和剖析海量新闻中重要历史事件的发展脉络，相关研究尚为数不多。该文由此构建针对近代报刊的时间信息标注工具，从批量新闻报道中抽取时间信息并自动转换为标准格式；在此基础上提出新闻网络模型，计算报道之间的时间关联性和内容相似性，利用有向加权网络的特有属性，追溯并探索特定历史事件的来龙去脉，不仅可以方便读者按照时间进程浏览新闻的前情回顾和后续发展；而且可以帮助研究者高效获取事件的背景知识和演变态势。该文使用《东方杂志》1911—1921年的“大事记”专栏的新闻报道作为实验数据，构建新闻网络，智能抽取历史事件的时间线，并结合特定史料加以分析佐证，旨在为近代史的知识发现提供新的研究视角。

关键词　近代报刊　大事记　新闻网络　时间线
分类号　G25　G210.7
DOI　10.16603/j.issn1002－1027.2021.03.017

1　引言

中国的近代报刊记录了朝代更迭和社会百态，是研究我国近代各领域的珍贵史料[1]。清末民初是一个风云变幻的时代，在此期间，中国不仅发生了洋务运动、辛亥革命、帝制复辟等举足轻重的政治运动，还发生了赫赫有名的思想启蒙运动如五四运动等[2]；与此同时，海外也发生了一系列重大事件，如第一次世界大战、俄罗斯十月革命、奥匈帝国瓦解等，对世界历史的进程产生了深远影响。在东西文化的碰撞与融合之中[3]，在内忧外患的挤压之下，我国近代各大报刊竞相设立纪事、记载、大事记等栏目，报道并评论当时中外政局的风云变化[4]。

近代（特别是晚清时期）报纸和期刊的界限一直比较模糊[5]，从名称的角度观之，多称为“报”，少数称为“杂志”；从内容的角度观之，有的虽称为“报”，但论文偏多或侧重特定专业学科。因而，本文不将近代报纸和期刊严格区分，皆为研究对象。民国报刊的大事记，多采用编年体的体例[6]，如《东方杂志》的专栏“中国大事记”和“外国大事记”，新闻事件按照时序分别列出，时间在前、标题次之、报道内容紧随其后，用简明扼要的文字，及时地向受众传递中外新近发生的重要事件[7]，让人一目了然。基于报刊的时效性和信息性，可以计算挖掘新闻之间的内在关联，既有益于近代报刊数字化资源的深度标注，也为近代文献的知识发现提供新的线索和思路，方便相关领域的研究者和历史爱好者纵览全局。但是迄今为止，从时间的维度追踪近代报刊中历史事件的发展动态，这类研究尚为数不多。

本文构建了针对近代报刊新闻报道的时间信息标注工具，并基于标准化的时间信息提出新闻网络模型，自动提取历史事件的时间脉络，旨在解答下述问题：(1)近代报刊的新闻报道中包含了大量未标准化的时间表达短语，如农历纪年、不完整的日期等，如果完全依赖人工将这些短语转换为规范时间格式，时间代价巨大，是否可以开发智能工具自动抽取并标准化。(2)以往的研究多关注含有特定关键词的相关事

* 国家社会科学基金一般项目“支撑城市记忆项目的‘数据基础设施’理论建构与实践探索”（编号：19BTQ007）的研究成果之一。
通讯作者：李惠，ORCID：0000-0001-7050-1845，邮箱：lh9743@126.com。

件，但如何突破具体词汇的限制，在海量报道中自动抽取某一特定事件主题下的“前因”与“后果”，并按时序呈现与主题相关的新闻报道，这个问题并未解决。

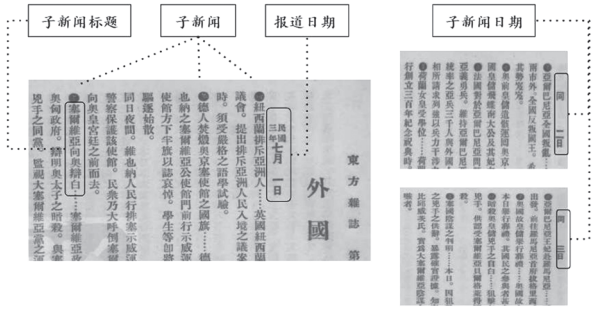
2 相关研究

目前，世界各地采用计算方法分析海量历史报刊的研究，大致可分为两类：一类侧重历史报刊的数字化建设，比如采用光学字符识别（Optical Character Recognition，OCR）技术识别扫描图片或者照片以及后续的校对及准确性评估。凯图宁（Kettunen）和帕科宁（Pääkkönen）通过计算词错误率、词形分析、词频分析等来评估 1771—1910 年芬兰报纸 OCR 之后的词汇识别质量[8]。苏拉德（Soullard）等提出一种基于卷积神经网络的算法，标注 1877—1944 年法语报纸图片的语义标签[9]。另一类旨在信息抽取和知识发现，侧重自然语言处理技术的运用，如词性标记、实体标注、话题分析等。佐根（Dzogan）等通过分析 1836—1922 年英国和美国报纸每一天的报道内容，观察词语使用的周期性[10]；阿维卡尼亚（Avikanien）提出一种基于小波变换的方法，检测芬兰报纸 1869—1918 年报道内容的历时变化，但并未涉及如何标准化报纸中的时间信息，以及如何追溯新闻报道中的事件主题[11]。史特根（Strötgen）等对八种语言的历史语料库抽取并标注日期信息，但并未涉及汉语历史语料[12]。现有公开的面向古代汉语的时间标注工具如古籍半自动标记平台（Markus）[13]，可抽取部分古籍中的日期表达，协助研究者人工标引时间信息，但标准化时间表达的功能尚未公开。

2 近代报刊的新闻时间线

2.1 日期标引

一篇近代报刊的新闻报道一般由多条子新闻组成，如图 1 所示，报道日期多与首条子新闻的日期一致，每条子新闻的日期相近。数字化的近代报刊中，新闻报道 R 一般由元数据和正文组成。元数据应包括报道日期 $t \in T$ 、摘要 $ab \in Ab$（多条子新闻的浓缩）、关键词 $kw \in Kw$ 等，正文由多条子新闻组成，而每条子新闻 $r \in R$ 包括子新闻标题 $tt \in Tt$ 和具体新闻内容 $c \in C$ 。报道的元数据和正文中存在着多种形式和类型的日期表达短语，本文主要处理两大类：(1)明确完整的日期表达（时间点），如“宣统三年三月初一日”“民国七年三月初八日”等，在对农历和公历表达有区分的前提下，可直接转换为标准日期格式。(2)部分完整的日期表达（时间点），如“前月二十七日”“同十三日”等，需参考上下文转换为标准日期格式。



注：选自《东方杂志》“外国大事记”专栏 1914 年第 11 卷第 3 期

图 1 近代报刊新闻报道示例

近代报刊的新闻报道中，大多数日期表达，都属于这两类，因而本文针对这两种类型的日期表达，初步构建了时间信息标注工具，工作流程如图 2 所示，具体应对策略如下：(1)农历和公历的转换。本次实验数据的时间周期从宣统元年到民国十一年，其中宣统元年到宣统三年以及短暂的宣统九年①的新闻报道，相关日期表达仍是农历制，需要转成公历。因此本文选取香港天文台的公历和农历日期对照表②，根据词表匹配，将农历年月日转换为公历，即年月日格式（YMD）。民国元年之后，月与日则同公历月与公历日。(2)部分完整的日期表达式，若前文有日期表达，缺失的信息以文中距离最近的标准化日期信息为准；若前文未有，以报道的首日期为参考日期，补充完整日期表达并标准化。

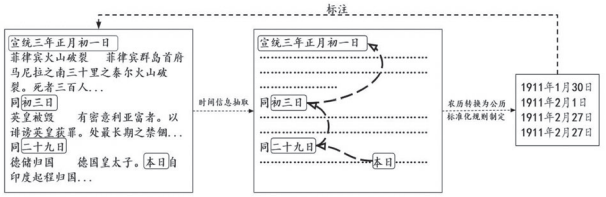


图 2 时间信息标注工具的工作流程

本文目前主要处理三类表达，简述如下：(1)例如“初三日”“二十二日”等：搜索前文中距离最近的标准化日期信息，确定年月信息；同时参考报道首日期创建日期，确定是农历还是公历纪年。若为农历，合并年月日表达之后，参考农历日期对照表转换；若

① 1917 年，张勋拥溥仪复辟，改民国六年为宣统九年。

② https://www.hko.gov.hk/tc/gts/time/conversion1_text.htm#。

为公历，合并年月日表达之后，可直接自动转换。(2)例如“本日”“本月”等：若为“本日”，以前文中距离最近的标准化日期信息为准；若为“本月”，参考前文日期信息，确定年月信息，标注为该月的第一天到最后一天的时间区间；“本年”同理。(3)例如“前月”“次年”等：搜索前文中距离最近的标准化日期信息，确定年月。若时间表达含“前”“上”等词，对应的年月日信息往前推一；若含“下”“后”等词，对应的年月日信息往后推一。注意：若前文日期表达出现“十二月”“一月”“三十日”“三十一日”“一日”等，标准化需另外写规则处理。

标引之后，以关系数据表的形式输出标引结果，共包含新闻报道ID、原始日期表达、原始表达所在文中的位置(词序)及标准化的日期表达这4个字段。

2.2 构建新闻网络

本文将近代报刊的新闻网络定义为一个有向图$G=(V,E)$，节点集合$V\subseteq R$代表新闻报道，报道日期作为属性赋予相应节点；边集合$E\subseteq 2^V\times 2^V$指代新闻之间的内容关联。如果两篇新闻i和j的相似性超过一定阈值，网络中节点i和j之间就存在一条有向边$<i_{t_i},j_{t_j}>$，有向边的方向遵循新闻报道时间的先后顺序。给定任意两篇新闻i和j，二者之间的相似性可计算如下：

$$sim(i,j)=\alpha_1\times sim(c_i,c_j)+\alpha_2\times sim(tt_i,tt_j) \quad (1)$$

其中$\alpha_1+\alpha_2=1$，$sim(c_i,c_j)$和$sim(tt_i,tt_j)$分别计算新闻内容和标题之间的相似性。本文将α_1、α_2均设定为0.5。考虑到选取的实验数据中新闻文本大多篇幅较短，主题提取效果并不良好，本文选择余弦相似度(Cosine Similarity)来开展文本相似度计算。余弦相似度是计算文本相似度的常用算法[14]，通过将文本映射到向量空间，计算两个向量的夹角余弦值，来衡量文本之间的差异性。

本文将一个新闻事件*event*定义为新闻网络中的一条有向路径$<v_1,v_2,\cdots v_k>$，即一组按时序排列的节点集合，其中被遍历的边$<v_x,v_{x+1}>\in E$，$1\leqslant x\leqslant k-1$。为了找出网络中的重要节点，本文采用中介中心性和PageRank这两种网络度量方法，从不同的角度综合判定网络中新闻的重要性；为了检测特定事件主题的潜在时间线，本文利用网络中的最短路径来提炼事件前情和后续发展。

中介中心性(Betweenness Centrality)。中介中心性衡量节点在网络中所起到的媒介或者桥梁作用[15]，是对该节点影响力的评价指标[16]。中介中心性越高，该节点对其他节点之间信息传递的控制力越强。

网页等级(PageRank)。PageRank是谷歌搜索引擎用来排序网页搜索结果的算法。该算法认为，万维网中一个页面的重要性取决于指向它的其他页面的数量和质量[17]。换言之，如果一个网页被很多重要的网页指向，则这个网页自身也很重要。给定任意节点i，预设网络中有若干节点包含j和n有边指向i，则PageRank采用迭代算法计算如下：

$$PR(i)=(1-d)+d\left(\frac{PR(j)}{C(j)}+\cdots+\frac{PR(n)}{C(n)}\right) \quad (2)$$

d代表阻尼系数，取值在0和1之间，$C(j)$代表节点(j)的出度。本文利用这种算法，对网络中的节点按重要性进行排序。

新闻事件时间线。本文将特定新闻事件的时间线定义为新闻网络中相应有向路径的长度$|<v_1,v_2,\cdots v_k>|$，旨在定位事件的来龙去脉。路径的长度越长，新闻事件的时间线越长，时间跨度越长，影响越深远。

社群中的热点词汇。考虑到本文构建的新闻网络是有向加权图，笔者采用信息地图(Infomap)算法检测新闻网络中潜在的社群。Infomap源于信息论，旨在用编码的形式描述有向图中随机游走(Random Walk)的最短路径[18]，计算如下：

$$L(M):=q_{\curvearrowright}H(Q)+\sum_{i=1}^{m}P^{i}_{\circlearrowright}H(P_i) \quad (3)$$

M代表图中存在的m个社区，$q_{\curvearrowright}H(Q)$计算跳转社区产生的信息熵，$\sum_{i=1}^{m}P^{i}_{\circlearrowright}H(P_i)$计算社区内所有节点(包括离开节点)产生的信息熵。本文在实验中会进一步探索是否处在同一社群的新闻含有相同的话题。

3 实验分析

《东方杂志》是以时事政治为主的社科类综合性刊物，被誉为“中国近现代史的资料库”“杂志界的重镇”“杂志的杂志”。《东方杂志》所刊载的中外大事记、中外时事汇录和各类汇志，皆按月详尽辑录当月

中外重大政治、经济、文化事件和要闻，后人翻检，极为便利[19]。本文选取上海图书馆收藏的1911—1921年《东方杂志》的“中国大事记”和“外国大事记”两大专栏共290条新闻报道作为实验语料。考虑到每篇新闻报道包含若干条不同日期的各地新闻，本文通过正则表达式拆分每篇报道，预处理之后，所有语料中共含有10094条子新闻。需要注意的是，尽管每篇新闻报道都有首日期，但报道所拆分的子新闻都有各自的事件日期表述，如“同三日”“同十八日”等，并未和首日期保持一致。在此基础上，笔者使用构建的时间标注工具，从语料中抽取时间表达短语并标准化，共得到16118条日期表达。在此基础上，请相关领域的专家人工校验所有文本，查漏补缺，共得到16521条日期表达；并同时对本次自动抽取及标准化的实验结果进行评估，度量结果如表1所示。本文采用了准确率、召回率、F值和Value来评价工具的有效性，具体计算方法如下所述。

$$准确率(P)=\frac{正确抽取的日期表达数量}{抽取的所有日期表达数量} \quad (4)$$

$$召回率(R)=\frac{正确抽取的日期表达数量}{文本中实际存在的日期表达数量} \quad (5)$$

$$F1=\frac{2\times P\times R}{P+R} \quad (6)$$

$$Value=\frac{标准化正确的日期表达数量}{正确抽取的日期表达数量} \quad (7)$$

表1　日期表达抽取及标准化效果评测

指标	数据集
准确率（P）	90.49%
召回率（R）	88.28%
F1值	89.37%
Value	86.2%

从评估结果可看出，抽取的准确率可达90%以上，标准化的准确率可达86%以上。本次实验日期抽取的疏漏主要集中在农历纪年和时间词上，如“前清光绪二十四年”“翌日”等事先未考虑在规则内，以及时间区别的不当抽取，如“于十四日起”“至二十四日”等。标准化的错误主要是前文日期的干扰过滤不当，如一段中出现多个完整的时间表达，就近原则识别错误；以及月份、年份更替的不当处理，如“三十一日”后紧跟“一日”等。虽本次实验错例出现不多，但已引起重视，在后续工具的开发中，会将这些问题逐一解决。

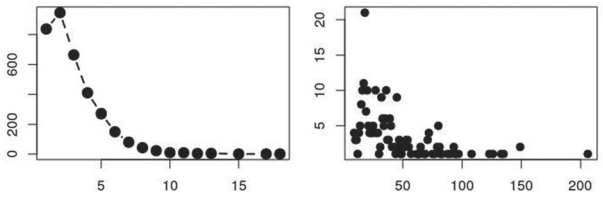


图3　《东方杂志》“大事记”专栏（1911—1921年）子新闻数目统计

本次实验对所有报道的首日期及子新闻数量做了统计，如图3(a)所示，横轴对应子新闻数量，纵轴代表具体数量在数据中出现的频次。实验数据中约90.5%的报道日期下，子新闻的数量都不超过5条；子新闻数量超过10条的报道在11年内仅有31篇；相比于图3(a)，图3(b)虽未呈现出较明显的规律，但也可看出，单篇新闻报道包含超过100条子新闻的，仅占所有新闻报道的3%左右，多数包含的子新闻数量小于30条。中外新闻频发于1912—1915年，且西方新闻数量最多的日期是1912年4月1日，共有14条子新闻；话题集中在“战争近况”“工人罢工”“军事建设”等，可见当时的世界政治局势并不稳定，为1914年的第一次世界大战埋下伏笔；东方新闻数量最多的是1915年9月20日，话题多聚焦在“国民会议”“法令颁布”等，应与签署“二十一条”有关。1912年第8卷第10期发表的西方新闻报道涵盖了1911年11月5日至1912年2月29日的206条国外子新闻事件，数量最多，均为国外大事，涉及“意土战争”“墨西哥叛乱”“名人逝世讣告”等话题；国内子新闻数量最多的也是这一期，包含1911年11月4日至1912年3月10日的104条子新闻，涉及“兵变”“总统就任”“临时政府成立”等话题。

热点新闻。基于上文所提出的新闻网络模型，本文将子新闻作为网络节点，基于内容和标题之间的相似性（阈值设为0.23），构建新闻网络。网络中共有3082个节点和4222条有向边，相应的中介中心性和PageRank计算结果分别如表2所示，中介中心性聚焦的是给定节点对网络中其余节点的控制性，换言之，中介中心性得分较高的节点（新闻）是网络中多数新闻的“前情”或者“后续”；PageRank聚焦

表 2 《东方杂志》新闻网络中的热点新闻

中介中心性		PageRank	
国内新闻	国际新闻	国内新闻	国际新闻
特任庄蕴宽为审计院院长	美参院通过对德战争终止案	派王宠惠为国际联合会全权代表	国际航务大会开幕
特派李经羲为约法会议议员	美国银行团借款与法国政府	特任刘承恩为湖北省长	墨西哥地震
公布修正行政执行法	土耳其议会解散	公布修正参议院议员选举法	国际联盟大会开幕
特任李经羲为审计院院长	俄国女子参政运动之失败	新任日斯巴尼亚驻京公使呈递国书	法政府批准法土条约
公布商会法	英国新设海军参谋部	特任潘龄皋为甘肃省长	瑞典新国会开幕
全国商会联合会在京开会	保国解散议会	公布文官高等考试法	万国女子参政大会开会
公布国民会议组织法	法国航空博览会开幕	新任墨国驻京公使呈递国书	日本预算案
任命于宝轩为内务次长	海参崴日俄谈判告竣	新任美国驻京公使呈递国书	英国煤矿工展缓罢工期
任命唐继尧兼署云南民政长	俄波休战条约签字	江苏省议会拟组织全国省议会联合会	希腊王病故
公布审计法会计法	万国女子参政大会开会	任命于宝轩为内务次长	英国煤矿罢工风潮扩大

的是给定节点的入度(影响力)，换言之，PageRank得分较高的节点(新闻)是网络中多数新闻的“后续”。因此，这两种方法计算的热点新闻并不完全相同，应综合观之。从表 2 中可以看出，无论是中介中心性还是 PageRank 的热点新闻，国内热点基本上是围绕官员任命和法令公布，这与当时改朝换代的政治氛围契合；国际热点虽有差异，但也有航务会议、女子参政、国会议会等共同话题，这可能也是当时编辑和民众的焦点所在。

本文对网络中的每个节点都计算了到达其他节点的最短路径，旨在追溯任意事件的“前因”以及搜寻引发关联的“后果”。以著名外交官顾维钧为例，在新闻网络中最早可追溯至 1915 年 7 月 11 日顾维钧被任命为驻墨西哥全权公使，最晚可延伸至 1921 年 11 月 15 日美国华盛顿召开太平洋会议，施肇基、顾维钧、王宠惠三人为全权代表，本文撷取 1919—1921 年顾维钧参与国际和谈的时间线片段，如图 4 所示。除了著名的 1919 年 6 月 28 日“中方在巴黎和谈中拒绝对德合约签字”事件，以及导火索中日之间的“鲁案”(日本一战期间侵占青岛及胶济铁路沿线车站)相关事宜辩论之外，同时穿插着 1920 年 5 月 31 日顾维钧等代表中国签署《万国航空专约》附件，使中国与其他国家具有同样的领空权；中方在签署对土耳其合约时，考虑到“凡此数端，为我国所欲设法解除而未能者，如竟行签字，仿佛赞成此项原则，似与我国外交政策不合，且增日后要求解除辩护之困难”[20]，故 1920 年 6 月 18 日未签合约。整体而论，中国当时的国际外交可圈可点，并不能算失败。

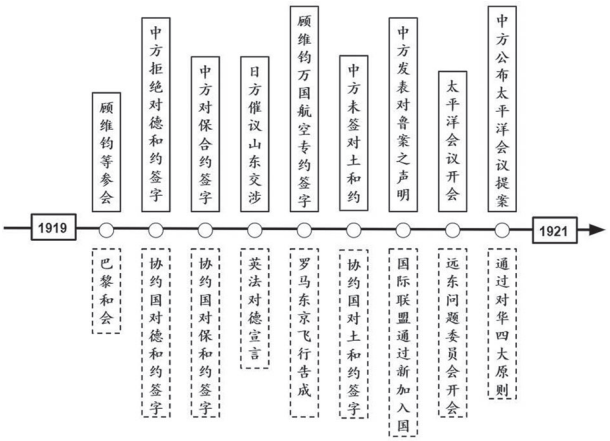


图 4 网络中的新闻时间线示例

这条时间线上的各条报道并未都含有相同的关键词，若仅通过关键词检索，未必能检索出全部事件，也未必能保证准确地按照时序排列这些事件。但通过我们构建的时间标注工具和新闻网络模型，这些隐含的主题相关联的事件可以在一条时间轴上清楚明确地展示出来，供学者开展全面细致的研究。如果后期可以扩大语料库的规模，中国百年来的外交政策变化和国际地位变迁皆可以利用笔者提出的新闻网络，直观而又客观地加以分析。

关键社群中的关键词。基于已构建的新闻网络，采用 Infomap 算法共检测到 410 个社群，考虑到本次实验的子新闻大多数篇幅较短，统计模型如话题模型(Topic Model)并不能取得较好的效果，因此，本文使用分词工具 Jieba、自定义的针对近代汉语的用户词典以及词频—逆向文档频率算法(Term-Frequency-Inverse Document Frequency,

TF-IDF)，计算社群中的高频词。图 5 展示了网络中最大的五个社群，每个社群中心圆圈内的数字代表社群内部节点的数量，围绕圆圈的是社群中新闻的关键词。可以看出，1911—1921 年，我国社会正处于一个内忧外患的状态，国内政治局势不稳定(“匪乱”“兵变”“革命军”)、自然灾害严重(“水灾”“旱灾”)、国际局势不明朗(“抗议”“袭击”)；从另一方面也说明《东方杂志》的“大事记”相关栏目对当时时政和时事的热点均有报道。通过对实验数据整体的词频统计所得到的高频词，皆为个体，未有聚类；如果单纯聚类词语的话，又不能看出新闻之间的关联。本文提出的新闻网络模型，在形成新闻社群的基础上提炼关键词，既保证了新闻间的关联性也提炼了网络中的关键词汇，一举两得。

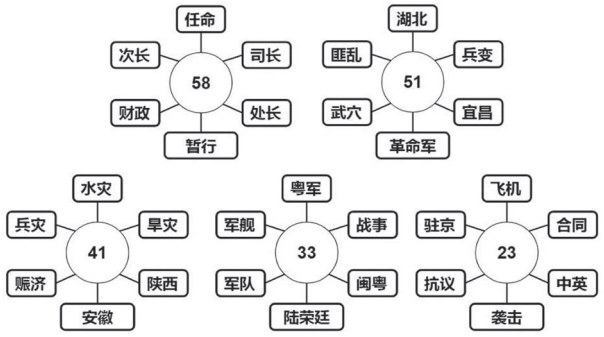


图 5　新闻网络中关键社群的词云

4　结论

近代报刊，属于近代大众传媒的重要组成部分，记录了近代发展的历史轨迹，是近代史的真实见证[21]；近代报刊设置新闻相关栏目报道国内外重大事件，向民众传递与之息息相关的政治消息，在当时的社会生活中，作用不容小觑[22]。文章以新闻报道的时效性为前提，构建针对清末民初近代报刊的时间信息标注工具，该工具可以被拓展到更早的年代，不局限于语料的规模。凭借该工具，自动抽取新闻报道中的时间信息，不仅是报道时间，而且是事件发生的具体时间，并智能标注为规范时间格式。在此基础上，整合新闻报道之间的时间关联性和内容相似性，提出针对近代报刊的新闻网络模型，旨在挖掘网络中潜在的热点新闻事件，抽取并分析重大历史事件的时间脉络。

文章的主要贡献如下：(1)文章构建的时间信息标注工具，可以帮助近代史的研究者抽取大规模报刊中的时间信息，并自动转换成标准时间格式，提高语料处理的效率，方便报刊信息的利用。(2)基于所提出的新闻网络抽取的新闻时间线，不仅可以直观地展示事件的发展态势，同时也提供同一时间点上发生的其他事件，使用者可以按照时间进程来阅读新闻，了解事件的来龙去脉，旨在帮助使用者从共时和历时的角度多维度地了解历史。

下一步研究计划包括：(1)时间信息标注工具完善。目前的时间信息标注工具尚处在初期阶段，针对的是文本中的日期信息，对文本中其他形式的时间表达，如“上午五点”“共三日”等尚未涉猎，下一步会扩大语料范围并扩展工具所处理的时间类型。(2)新闻自动摘要。目前针对近代报刊的新闻摘要主要是从单篇新闻报道中提取重要信息，尚未充分利用新闻隐含的时间信息。我们下一步的重点工作，即结合本文所抽取的新闻时间线，智能地提取息息相关的新闻事件中的重要信息，自动生成摘要，为相关学者和从业者提供便利。

参考文献

1　颜佳. OCR 在民国报刊数字化项目中的应用研究 [D]. 上海：华东师范大学，2008：30－31.
2　佟彧. 从《申报》看清末民初中国报纸通讯文体的发展(1896－1915)[D]. 沈阳：辽宁大学，2011：30－31.
3　张倩. 家国情怀与忧患意识[EB/OL]. [2020－08－06]. http://www.qstheory.cn/zhuanqu/bkjx/2018－09/08/c_1123399586.htm.
4　马少华.论孟森对期刊记事栏目的体例创新[J]. 国际新闻界，2011，33(9)：110－115.
5　刘英钦. 清末民初报纸新闻报道类文体的发展与变革[J]. 中州学刊，2013(6)：128－133.
6　刘英翠，聂进倩，孙美玲. 民国期刊述评专栏的属性变革与现象解读 (1918—1949 年)[J]. 出版发行研究，2020(7)：105－111，57.
7　翟小纯. 从中国新闻报刊史看其时代特征[EB/OL]. [2020－08－06]. http://theory.people.com.cn/n/2013/0516/c40537－21505493.html.
8　Kimmo K，Tuula P. Measuring lexical quality of a historical Finnish newspaper collection-analysis of garbled OCR data with basic language technology tools and means[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) . Portorož，2016：956－961.
9　Yann S，Pierrick T，Clément C，et al. Multi-scale gated fully convolutional DenseNets for semantic labeling of historical newspaper images[J]. Pattern Recognition Letters，2020，131：435－441.
10　Fabon D，Thomas L W，Nello C，et al. Discovering periodic patterns in historical news[J]. Plos One，2016，11(11)：e0165736.
11　Jari A. A method for wavelet-based time series analysis of historical newspapers [D].Helsinki：University of Helsinki，2019：4－5.

12 Jannik S，Thomas B，Julian Z，et al. Extending HeidelTime for temporal expressions referring to historic dates [C]// Proceedings of the International Conference on Language Resources and Evaluation (LREC). Reykjavik：ELRA，2014：2390－2397.
13 Markus [EB/OL].[2020－08－06]. https://dh.chinese-empires.eu/markus/beta/index.html.
14 张振亚，王进，程红梅，等. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学，2005(9)：162－165.
15 袁康，汤超颖，李美智，等. 导师合著网络对博士生科研产出的影响[J]. 管理评论，2016，28(9)：228－237.
16 Mark N. Scientific collaboration networks. Ⅱ. shortest paths，weighted networks，and centrality[J]. Physical Review E，2001(64)：2－4.
17 Sergey B，Lawerence P. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems 30，1998：107－117.
18 Rosvall M，Bergstrom C T. Maps of information flow reveal community structure in complex networks[C]//Proceedings of the National Academy of Sciences of the United States of America (PNAS). 2008：105，1118.
19 上海图书馆.全国报刊索引[EB/OL]. [2020－08－06]. https://www.cnbksy.com/literature/literature/bbf23f666204789e30acb4e4fb7df771.
20 唐启华.顾维钧：舌战巴黎的青年外交家[EB/OL]. [2020－08－06]. http://www.bjnews.com.cn/culture/2019/04/26/572716.html.
21 张小虎. 近代时政刊物对英美宪政文化的介绍——以《东方杂志》"宪政研究专号"为考察视角[J]. 黑龙江省政法管理干部学院学报，2016(2)：1－4.
22 中国民办报纸的历史意义[EB/OL]. [2020－08－06]. https://www.163.com/dy/article/CRT8HO2G0525IT90.html.

作者单位：李惠，南京农业大学人文与社会发展学院，江苏南京，210095
夏翠娟、刘炜，上海图书馆/上海科学技术情报研究所，上海，200031
侯君明，上海古籍出版社，上海，200020
朱庆华，南京大学信息管理学院，江苏南京，210023
收稿日期：2020 年 11 月 22 日
修回日期：2021 年 2 月 28 日
（责任编辑：支娟）

Timeline Extraction of News in Historical Newspapers and Periodicals of Modern China

—An Experiment on the Basis of Chronicles in *Oriental Magazine*

Li Hui　Xia Cuijuan　Hou Junming　Zhu Qinghua　Liu Keven

Abstract：News in historical newspapers and periodicals of modern China recorded the political and social events that took place during this time period. However，there is a notable lack of research in tracing the development of significant events in extensive historical news. Therefore，in this paper we develop a temporal tagger especially for news in newspapers and periodicals of modern China，which can help users to extract temporal expressions from news and normalize them into standard time formats. On this basis，we propose a news network model and calculate the similarities between contents and temporal relations of different news. We exploit network properties of this directed weighted network to extract the timeline of historical events. This timeline，which not only facilitates users to read news which automatically following a time-order，but also assists researchers with background knowledge and dynamic evolution of important historical events. In this paper，we use a dataset of chronicles of events from 1911 to 1921 in *Oriental Magazine* for experiments and corresponding experimental results can provide a new perspective for the knowledge discovery in modern Chinese history.

Keywords：Newspapers and Periodicals of Modern China；Chronicle of Events；News Network；Timeline