doi:10.3772/j.issn.2095-915x.2020.02.004

多国议会数据集及平台建设研究

开放科学（资源服务）标识码（OSID）

黄佳妮[1] 王君领[1] 沈嘉裕[1] 王伊杨[1] 张约翰[1] 王佳敏[1,2] 胡吉明[1,2] 陆伟[1,2]

1. 武汉大学信息管理学院 武汉 430072;
2. 武汉大学信息检索与知识挖掘研究所 武汉 430072

摘要：议会文本是国家就国内和国际问题讨论所形成的记录。为解决议会数据获取和分析的问题，本文深入调研当前议会数据开放情况及其用于科学研究的现状，提出一种议会数据平台架构实现方案和议会数据集建设方案。方案综合使用了大数据、自然语言处理等技术构建了多国议会数据存储、展示、计算的基础框架。议会大数据平台和开放数据集可为科研人员和政府相关部门提供研究资料和参考。

关键词：多国议会；议会大数据；议会平台；议会数据集

中图分类号：G350

Data set and Platform Construction of National Parliaments

HUANG Jiani[1] WANG Junling[1] SHEN Jiayu[1] WANG Yiyang[1] ZHANG Yuehan[1] WANG Jiamin[1,2]
HU Jiming[1,2] LU Wei[1,2]

1. School of Information Management, Wuhan University, Wuhan 430072, China
2. Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China

Abstract: Parliamentary texts are records of discussions of domestic and international affairs. To solve the problem of parliamentary data acquisition and analysis, this paper investigated the current status and progress of parliamentary data opening and its application in scientific research. Based on this, a plan for constructing the data sets and platform of parliaments are proposed and implemented. Various techniques, such as big data analysis and natural language processing, has been used to implement the basic framework of storage, display and calculation of the parliamentary data in five countries. The parliamentary data platform and open data sets proposed in this paper can contribute to provide research materials and references for the government and researchers.

Keywords: Multinational parliament; parliamentary big data; parliament platform; parliament data set

基金项目：国家级大学生创新创业训练计划项目“国外议会数据平台建设与深度挖掘研究”（201910486015）。

作者简介：黄佳妮（1999-），本科生，研究方向：文本挖掘；王君领（1999-），本科生，研究方向：数据科学；沈嘉裕（1999-），本科生，研究方向：数据挖掘；王伊杨（1999-），本科生，研究方向：数据科学；张约翰（1998-），本科生，研究方向：数据挖掘；王佳敏（1992-），博士研究生，研究方向：文本挖掘、知识网络，E-mail: wangjm@whu.edu.cn；胡吉明（1985-），博士，副教授，研究方向：电子政务与政府信息资源管理；陆伟（1974-），博士，教授，博导，研究方向：信息检索、知识挖掘。

引言

当前国际局势正在发生深刻的变化，世界格局变动转型，国际形势中的不稳定、不确定因素明显增加。在这种背景下，掌握其他国家的社会、经济、政治等变化所带来的国家发展态势，对我国政府积极应对外交形势变化，制定外交政策具有重要意义[1]。近年来，多个国家或组织开放了大量政府数据，不同学者从多个角度对外交关系进行了研究，如外交态度分析，外交分歧，国家利益及国家关系分析等[2-6]。其中，议会辩论文本（Parliamentary debates）是对议会活动中政客发言内容的高度结构化笔录，其内容包含了本国与其他国家相关问题的言论，能够实时、直观地反映对外关系[7]。多数国家将议会辩论文本内容进行结构化处理，甚至建立了检索系统，以便于议会辩论文本的深层次挖掘和利用。对议会辩论文本的解读和分析，挖掘一个国家在对外关系中的规律和态势，成为近年来国际关系量化研究的一个主要趋势。

然而，当前各个国家政府网站提供的议会数据结构化程度低，为议会数据分析带来一定困难。且各国议会数据分散在多个网站上，加大研究人员数据获取的成本。另一方面，各国的议会数据开放网站不提供高级的检索功能和有效的分析工具，因此难以供科研人员直接获取需要的数据并加以利用。

为解决议会数据的获取和分析问题，本研究利用爬虫技术和自然语言处理技术，获取了英国、美国、俄罗斯、加拿大及欧盟 5 个国家或组织的议会数据，进行结构化处理后得到了相应的 5 个数据集。进一步基于这些数据集建立了议会大数据平台网站①，为国内外学者提供数据的免费开放获取以及查询、分析服务，为科研人员和政府机构开展相关科研和决策工作提供支持。

1 相关研究

1.1 各国议会数据开放现状

随着信息技术的发展和开放数据运动的推动，议会辩论数据、会议记录电子化成为新浪潮，多个国家、组织都在加速议会数据电子化进程。2012 年 9 月 15 日在罗马举行的世界电子议会会议（World e-Parliament Conference）上公布的《议会开放宣言》（Declaration on Parliamentary Openness），得到了来自加拿大、俄罗斯、英国、美国等 53 个国家的 76 个组织的支持，旨在呼吁议会组织、机构提高公开度及公民参与度。OpeningParliament 论坛②在帮助世界上从事监测、支持和开放本国议会和立法机构的民间组织建立联系方面所扮演的重要角色越发得到各界肯定，目前已有 75 个国家的 140 多个组织加入其中。

多数致力于议会数据开放的国家都建立了本国的议会数据网站，用以开放决策文件和议会会议记录。Berntzen 等人[9]对挪威议会数据开放情况调研后得出，挪威议会一直在积极利用信息和通信技术提高透明度，并在议会网站

① http://parliament.whu.edu.cn
② http://www.openingparliament.org

doi:10.3772/j.issn.2095-915x.2020.02.004

提供了决策文件、会议记录的访问渠道。Berntzen 等人[10]还调查了斯堪的纳维亚议会（挪威、瑞典和丹麦）的数据开放情况，发现三个国家都已开放议会投票、辩论和委员会会议记录等数据。Seaton[11]对苏格兰议会和电子民主进行了调研，并指出苏格兰议会一直将互联网视为让苏格兰公民了解议会内容、参与议会事务的主要机制之一，电子请愿系统、网上会议记录等措施已有成效。

部分国家或组织的议会还提供精确的议员发言及议员详细政治背景信息。如，安达卢西亚将议会记录进行了电子化并在网络上发布，其中包含了议员的所有发言记录[12]。欧洲议会目前也已经开放了各个委员会的会议记录和议员信息，发言内容可以与发言人的个人信息、国家、政党、委员会等信息进行关联[13]。

随着议会数据的不断开放，议会数据开放政策和活动也得到越来越多的关注。Faria 等人[14]研究了巴西众议院的议会数据开放政策，将巴西众议院的举措总结为建立电子民主门户和开放数据、创建议会数据实验室两大类。Beelen 等人[15]在加拿大议会的数字化进程研究中提到，Dilipad 计划将为加拿大、英国和荷兰的议会创建统一的、可拓展的数字化记录，并促进研究人员对这些数据进行分析，利用这些数据解决有关意识形态、移民等实质性研究问题。

1.2 议会数据用于科学研究的现状

议会数据文本中蕴含了丰富的信息，对议会文本的分析和研究已经取得了一定的成果。Rohit 等人[16]对印度议会辩论中提供的发言信息进行了立场分类，采用人工分析的方式将其分为赞成、呼吁行动、发现问题、责备四类，借此分析议员们对特定法案的立场及其意图。De Campos 等人[17]采用无标注的正例文档学习方法，基于西班牙安达卢西亚地区议会数据，对议员们的政治偏好进行了分析，并开发了“推荐-过滤”系统，应用于对议会接收到的文件（如新闻稿或技术报告）进行过滤并推荐给具有相应偏好的议员。Zhang 等人[18]采用无监督方法对英国议会中议题的修辞作用进行了分析，以此对议会议题进行分组，揭示了议题中提问者的意图和背景，以及政府和反对党之间的分歧与质疑。Rheault 等人[19]对英国议会下议院1909-2013年的辩论内容开展了自动文本分析，描绘了政府官员情绪极性总体水平的变化，发现与其对经济衰退的反应相关。Przybyla 等人[20]基于波兰议会的文本数据，应用特征选择、回归和分类算法，提取了发言中的文本特征，对议员们的性别、教育程度、所属党派等背景信息进行预测。Kaptein 等人[21]对荷兰议会的半结构化会议记录数据进行重点检索和结果汇总，并以此开发了一个搜索引擎，支持分面搜索和检索结果分组。Jungermann 等人[22]以德国议会的会议记录和公开请愿记录为语料，采用信息检索、机器学习和预处理相结合的框架，进行主题提取、数据挖掘，并以此提供信息检索服务。

总的来说，议会数据可广泛地应用于科学研究，已有的研究包括主题分类、情感分析、信息过滤、政治文本分面检索等。随着议会数据开放程度和可获取性的增强，围绕议会数据的科学研究也会有所拓展。

1.3 议会数据集与平台建设现状

政府数据的开放已经成为世界性的趋势，各国政府已经把开放数据提到了前所未有的高度，部分国家和组织已建立并开放了议会数据集和平台。

在数据集建设方面，Shadbolt 等人[23]使用 EnAKTing 方法对英国政府数据开放平台“Data.gov.uk”进行研究，指出议会数据平台对于政府、技术社区和公民都有着极大的价值，并强调关联数据网络能挖掘出议会数据更大的价值。Bulut[24]构建了土耳其政府公开的议会法案数据集，指出该数据集可应用于法案修改趋势预测等研究。Galiotou 等人[25]认为加强其数据集与欧洲其他国家数据集的关联能够使希腊政府门户网站发挥更好的作用。

平台建设方面，目前各国议会数据联盟（Inter-Parliamentary Union, IPU）①是世界最大的全球性各国议会间合作组织。该组织网站提供了一个各国议会资料库，并且开放下载调研报告等文献和研究资料。Gielissen 等人[26]设计了有关荷兰议会的出版物的 Web 信息系统，包括功能设计、前后端设计等。英、美、德、加拿大、俄罗斯、欧盟已建成数据开放共享平台。mySociety 项目构建的议会监督网站 TheyWorkForYou②，旨在让英国公民更容易了解英国议会及苏格兰、威尔士和北爱尔兰议会的情况，该平台提供可追溯至 1935 年的议会文件的检索和下载，以及英国议会发言人的信息。美国国会网站③提供美国联邦立法相关信息，如法案信息、法案投票、议员信息与听证会视频。欧盟议会网站④支持欧盟议会记录文件的检索、下载以及议员的基本信息和发言记录等。

综合来看，全球很多国家都在积极进行议会数据开放并投入到议会数据平台建设中，但议会数据的开放程度和议会数据平台的功能还有待进一步提升。如何对这些数据进行组织、管理、分析、挖掘，构建数据集之间的联系，实现议会数据价值的最大化与开放数据利用程度的提升，是议会数据集开放过程中亟待解决的问题。

2 系统构建

2.1 问题发现

通过文献调研和实践调查，笔者发现英国、美国、加拿大、俄罗斯、欧盟在议会数据开放的实践方面在世界范围内处于领先地位，且在政治、经济等各个领域有着较高的影响力，因此本研究初步拟对上述国家、组织的议会（国会）文件进行收集和组织。调研发现，这些国家或组织网站提供原始的会议记录文件的下载和浏览，但其有效信息的分布较为零散，且部分网站开放的文件具有时效性，仅仅提供一段时期内的会议记录文件。对于用户来说有效信息的搜集难度较大，收集成本较高。且各国的议会网站各成系统，没能有效整合，不提供跨国家的检索和分析方法。

此外，各国议会联盟的网站虽然实现了对

① https://www.ipu.org
② https://www.theyworkforyou.com
③ https://www.congress.gov
④ https://www.europarl.europa.eu/portal/en

doi:10.3772/j.issn.2095-915x.2020.02.004

议会数据的跨国家分析，但依然存在着不提供议会文件资料、不提供议会文件整合数据集、文件未经结构化处理和标注、缺乏对于数据的深度加工和展示等问题。

2.2 数据集构建与发布

2.2.1 数据集构成

本研究构建的多国议会数据集由 5 个相对独立的数据库构成。其中，英国议会数据集主要包括英国上议院、下议院的议会文件记录；美国议会数据集包括美国国会会议文件；俄罗斯议会数据集包括俄罗斯上议院、下议院以及非定期会议文件；加拿大议会数据集包括加拿大议会会议记录、议员信息、议员发言记录和政党信息；欧盟议会数据集包括欧盟议会文件和议员信息。各个数据集具体的原始数据描述见表 1，总数据量约 47GB。

表 1 议会数据集描述

国别	年份	数据描述	数据格式	来源
英国	1999-2019	上议院会议	xml	https://www.theyworkforyou.com
	1935-2019	下议院会议		
美国	1995-2019	国会会议	pdf	https://www.congress.gov
俄罗斯	1997-2019	上议院会议	pdf	http://www.council.gov.ru
			doc	
		下议院会议	pdf	
		非定期会议	pdf	
加拿大	1901-2019	议会会议	xml	www.lipad.ca
		政党信息		
		议员发言		
		议员信息	csv	https://openparliament.ca
欧盟	1996-2019	议会会议	csv	https://europarl.europa.eu/portal/en
		议员信息	pdf	

2.2.2 数据处理

不同国家议会数据的语言类型、文件类型、包含内容等有差异，因此，我们对其分别做了以下处理：

（1）格式转化。对不能直接通过程序处理的、非结构化的数据转化为可供程序处理的格式，最后构建为数据库。例如，对美国、俄罗斯、欧盟等国家（组织）的 pdf 文件，采用 pdf 文本解析技术将其转化为可供计算机处理的 txt 格式。

（2）实体抽取。利用 XML 标签、机器学习等方法对议会文件中的重要实体，如主题、议员、发言、时间、地点（国家）、涉及的国家等进行抽取，并和原始文件进行关联，统一存储在数据库中。

2.2.3 数据存储

按照统一的信息资源目录体系，统一的面向对象数据组织的基本原则进行多国议会数据集数据库的设计，并遵循以下技术路线要求：

（1）以关系型数据库管理系统 MySQL 为

支撑，进行数据建模、组织和管理。

（2）以 Java、Python 为主要开发语言，以 Navicat Premium 为数据库管理工具，完成多国议会数据库的构建。

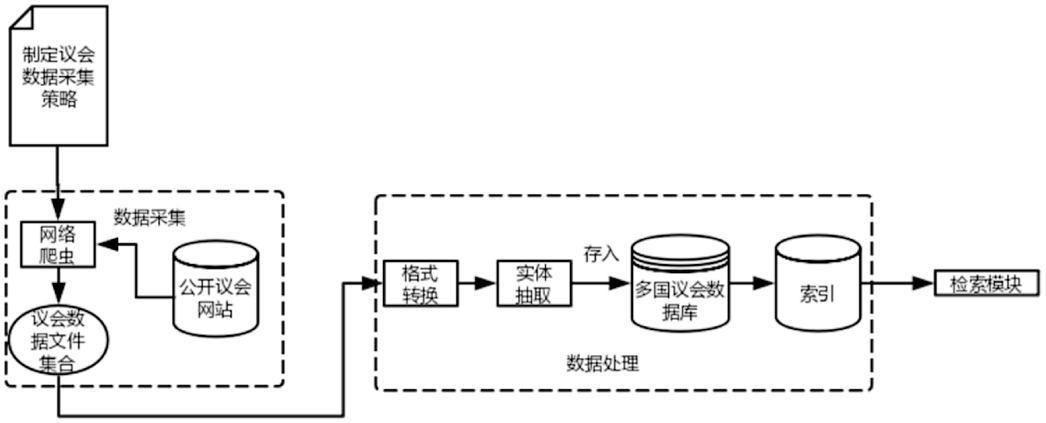

图 1 数据处理流程图

2.2.4 数据集发布方案

如图 2 所示，议会大数据平台系统由下至上分为六层，包括数据存储层、数据交换层、应用支持层、应用层、展现层、用户层。数据存储层包括各国议会文本数据、发言人数据、基本参数、网页结构数据、用户数据等。在数据交换层，数据处理类、文件上传类、下载类和权限管理类负责将数据通过数据总线传输到上层。应用支持层将数据交换层传送上来的数据运用到各个功能模块中，包括后台管理平台、Web 门户、文件传输模块等。应用层实现各功能模块，包括文件导入、权限设置、界面更新、浏览、查询、文件下载等。展现层则是根据用户层的不同角色，有选择地将部分界面和相应功能呈现给对应的用户，当前本网站针对未注册的用户，提供数据集的浏览、检索服务；针对已注册的用户，提供数据集的浏览、检索、批量下载以及相关数据分析结果的可视化展示等服务。

2.3 功能及实现

议会大数据平台的核心功能是对议会数据的检索、下载及在细粒度实体抽取基础上的关联网络构建与知识挖掘。检索、下载功能有助于用户方便、快捷的获取到所需的数据。关联网络构建与知识挖掘从实际应用角度展示了对议会数据的分析和利用，指明了其潜在的一个应用方向。

2.3.1 检索功能

议会大数据平台的检索模块包括全文检索和高级检索两个子模块，可满足用户不同的检索需求。

（1）全文检索

全文检索模块支持对议会数据的全文检索，同时支持对检索式的自动纠正，以及检索式和数据集的模糊匹配。本系统使用 Lucene 项目工具包[①]构建全文检索的核心功能。全文检索系统的构建分为构建索引和构建对索引的搜索系统两部分。本系统首先使

① http://lucene.apache.org

doi:10.3772/j.issn.2095-915x.2020.02.004

用反向索引机制构建索引。使用不同的解析器对不同类型的文档进行解析，通过分词器提取索引，经过多次组合迭代最终组成索引列表。

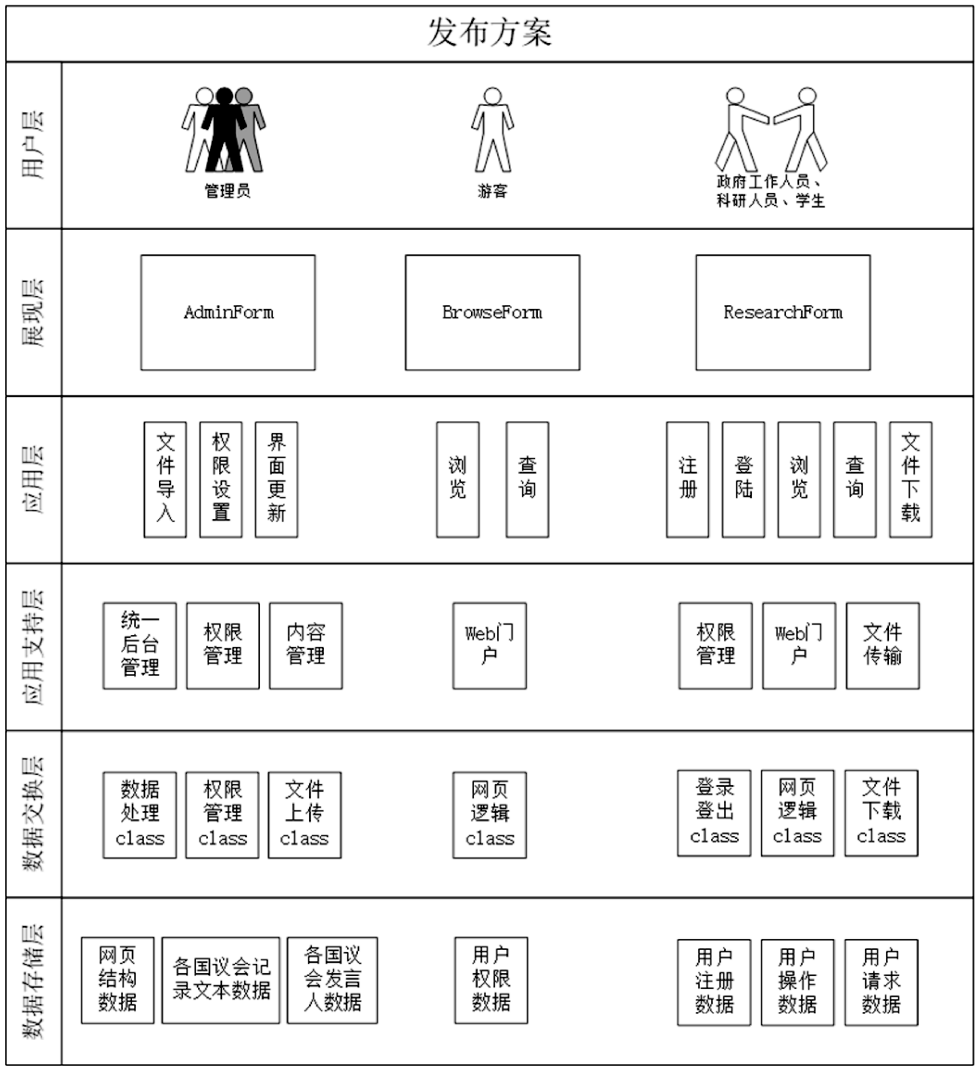

图 **2** 数据集发布方案示意图

在构建对索引的搜索系统方面，本平台使用 Lucene 工具包自带分词器对用户输入的查询语句分词，利用分词结果构成 query 检索式，然后使用模糊匹配算法将 query 和已经构建的索引列表进行匹配，返回匹配文件所在的路径。之后根据该路径查询数据库中对应的相关信息，将信息利用 servlet 技术返回给 jsp 结果界面，显示查询结果。

（2）高级检索

高级检索模块支持对议会数据的各个字段进行检索。可供检索的字段随不同国家的议会数据类型而改变。本模块主要依赖于 MySQL 的 B+ 树查找算法进行检索匹配。采用 B+ 树结构进行数据存储可以保证数据稳定有序，减少

查询过程中的时间复杂度。

在实际进行查询时，首先利用 SQL 语言构建后台检索语句框架，构建 Servlet 组件用于接收用户查询式。然后将用户查询式转换为对应的 SQL 语句进行查询。在进行查询时，B+ 树使用自顶向下逐层查找的方式匹配检索式和叶子节点。由于 B+ 树的中间节点不包含数据，所以同样大小的磁盘页可以容纳更多的节点数据，提高查询效率。最终将匹配到的查询结果返回给 jsp 结果界面，显示查询结果。

2.3.2 关联网络可视化

关联网络可视化模块提供以下核心功能：

（1）国家抽取

基于爬取到的 XML/PDF/DOC/CSV 等多种格式数据，为实现国家关联网络的挖掘，需要将原始数据文件进行处理，从全文、段落、句子多个层面进行细粒度实体抽取。团队基于 XML 树结构，逐层遍历，结合文本内容特征筛选、获取有效节点，主要采用 python 标准库、xml 及 pdfminer 等的第三方库实现，以加拿大议会文件的实体抽取为例，抽取出会议时间、发言人姓名、职位、发言主题、发言内容、发言人所属政党等实体，图 3 展示了部分抽取结果。对于涉及的相关国家，依赖取自中华人民共和国外交部官方网站[①]的全球国家列表，根据数据文件内部标签特征匹配国家集合并写入数据库保存。

ID	speechdate	speaker_position	speech_text	speaker_party	speaker_name	involve_country
160	1948-03-12	Minister of Finance an	Yesterday in my unavoidableabsence the hon. m	Liberal	Douglas Charles Ab	Canada,United State
171	1948-03-12	Minister of Labour	It is enough to have one loud-voiced person in t	Liberal	Humphrey Mitchell	Canada,Japan
175	1948-03-12		Mr. Speaker, has the Secretary of State for Exte	Progressive Conse	Gordon Graydon	Canada
209	1948-03-12	Minister of Justice anc	So I believe. The next is P.C. 20/6173 of 1945, ch	Liberal	James Lorimer Ilsle	Canada,Japan
223	1948-03-12	Whip of the Co-opera	With respect to the list the minister is placing on	Co-operative Com	Stanley Howard Kn	Japan
229	1948-03-12	Whip of the Co-opera	-in the act that is now on the statute book. I can	Co-operative Com	Stanley Howard Kn	Japan
230	1948-03-12	Minister of Justice anc	I have not consulted with the Chairman or with t	Liberal	James Lorimer Ilsle	Japan
232	1948-03-12	stagedirection	To the Continuation of Transitional Measures Ac			Canada
234	1948-03-12	stagedirection	To the Continuation of Transitional Measures Ac			Japan,Canada
238	1948-03-12	stagedirection	Transitional Measures ActP.C. 9870, 17/12/41-Au			Japan,Canada
273	1948-03-12	Parliamentary Assistar	Much as I would like tosee these cheques autho	Liberal	Walter Adam Tucke	Canada
364	1948-03-12		Mr. Speaker, I wish to discuss this bill at this tim	Progressive Conse	James Arthur Ross	Canada,United King
365	1948-03-12	Minister of Agriculture	That is not what I put on the record. I put on the	Liberal	James Garfield Gar	United States
376	1948-03-12		I would hope that it would be, but I do not knov	Progressive Conse	James Arthur Ross	United States,Canad
413	1948-03-12	Parliamentary Assistar	-you would have heard me saying that I had ask	Liberal	Walter Adam Tucke	Canada
417	1948-03-12	Parliamentary Assistar	-of feeds until such time as it may be embodied	Liberal	Walter Adam Tucke	Columbia,Canada

图 3 加拿大议会文本实体抽取结果示例

（2）主题抽取

在收集到的各国原始议会文件中，已经有部分文件提供缺省的主题、发言人等描述字段，为方便用户迅速了解每个文件或段落的主题，对这些字段进行了抽取建库。在对数据文件进行初步解析的基础上，依赖其内容标识特征遍历提取相关层的主标题及副标题，对其进行直观的标引。图 4 展示了加拿大议会文本段落层次部分主题抽取结果，包含主题（main_topic）、子主题（sub_topic）、段落内容（speech_text）和发言人（speaker_name）字段信息。

① https://www.fmprc.gov.cn/web

doi:10.3772/j.issn.2095-915x.2020.02.004

ID	main_topic	sub_topic	speech_text	speaker_name
3	OATS AND BARLEY	EXTENSION OF SUPPORT PRICES TO	Mr. Speaker, I wish to make a statement that will be c	Clarence Decatur H
4	OATS AND BARLEY	EXTENSION OF SUPPORT PRICES TO	May I ask a question arising out of the announcemen	John George Diefer
5	OATS AND BARLEY	EXTENSION OF SUPPORT PRICES TO	I can advise my hon. friend that the same support pri	Clarence Decatur H
6	OATS AND BARLEY	EXTENSION OF SUPPORT PRICES TO	May I ask a supplementary question? What is the star	John George Diefer
7	OATS AND BARLEY	EXTENSION OF SUPPORT PRICES TO	It is my understanding that the firms who bought barl	Clarence Decatur H
14	HONG KONG	DUFF COMMISSION-QUESTION AS T	The house will recall tha't the day I was requested to	William Lyon Macke
28	QUESTIONS AFFECTING MEN	SHORTAGE OF RAILWAY CARS FOR T	With reference to the shortage of railway cars for the	Lionel Chevrier
31	QUESTIONS AFFECTING MEN	FLAXSEED	I wish to direct a question to the Minister of Trade an	Percy Ellis Wright
32	QUESTIONS AFFECTING MEN	FLAXSEED	The price of flax for the 1948 crop year is being studi	Clarence Decatur H
35	QUESTIONS AFFECTING MEN	INDUSTRIAL DEFENCE BOARD	I wish to address to the Minister of Trade and Comm	Edward George Mc
36	QUESTIONS AFFECTING MEN	INDUSTRIAL DEFENCE BOARD	I would call the attention of my hon. friend to the fact	Clarence Decatur H
37	QUESTIONS AFFECTING MEN	INDUSTRIAL DEFENCE BOARD	The question was passed on to me since I came into	Brooke Claxton
40	QUESTIONS AFFECTING MEN	VICTORY OF HUMBOLDT RED INDIAI	May I direct a question to the Minister of Mines and I	Joseph William Bur
43	QUESTIONS AFFECTING MEN	TOURIST INDUSTRY	I shall be glad to inquire about such reports, and if th	Clarence Decatur H
46	QUESTIONS AFFECTING MEN	GOVERNOR GENERAL'S SPEECH	Mr. Speaker, I rise to speak for two reasons: first, to	Wilbert Ross Aylesv
47	QUESTIONS AFFECTING MEN	GOVERNOR GENERAL'S SPEECH	for them, and it is also good business for the farmers	John Horne Blackm
48	QUESTIONS AFFECTING MEN	GOVERNOR GENERAL'S SPEECH	What does the hon. member mean by $3.41 a month:	Humphrey Mitchell
50	QUESTIONS AFFECTING MEN	GOVERNOR GENERAL'S SPEECH	You have the Minister of Labour on the jump.	Joseph William Bur

图 **4** 加拿大议会文本主题抽取结果示例

（3）关联网络分析与可视化

根据抽取出的国家、主题、时间、人员等实体的语义关系构建知识关联网络，在此基础上进行国家关联关系分析、议题分析、议员分析、议会提及国家的演化分析等，是系统功能的重要组成部分之一，提供了对数据潜在研究价值的挖掘和展示。

以国家关联网络构建为例，本文在提取的国家集合及其共现关系的基础上，基于共现分析理论和社会网络分析方法进行国家共现网络构建和共现指标计算并实现对核心国家群的网络拓扑结构可视化。具体的实现步骤为：首先抽取目标文件所涉及的国家列表，然后对国家之间的共现关系强度进行计算并建立共现网络，提取该网络的最大连通子图，得到国际关系网络，该网络代表该数据集视角下的核心国家群。接着，运用 Louvain 算法 [27] 对国际关系网络进行社区划分，便于直观理解世界各国的集团划分状况，最后对各个主题社区进行可视化展示，通过横向网络指标比较和可视化展示来定位社区视角下的核心国家，分析国家关联关系特征与结构洞，挖掘有价值的模式。图 5 展示了某议题下多个国家之间的关联网络。

3 讨论

3.1 系统评价

（1）系统功能评价

该系统从数据集构建、平台建设与数据分析等层面对多国议会数据进行组织与管理，为英、美、加拿大、俄罗斯以及欧盟的议会数据跨国家与地区的综合利用提供了可行的工具与手段，符合当前议会数据集与数据开放平台研究中对多数据集建立连接关系以实现价值深入挖掘的趋势。

（2）系统优势分析

经文献调研与实践，笔者认为相较于已有的议会数据开放平台，该系统的优势可以体现在以下几点：第一，提供跨国家的检索和针对议会文件主题的揭示和可视化展示，实现议会

数据跨越国家和地区的综合利用。第二，整合零散分布的议会数据，形成多国议会数据集并提供一手资料，具有较强的针对性、及时性且情报含量高。第三，对原始数据文件进行了二次加工，对文本中蕴含的主题、国家、议员等信息进行了细粒度抽取，极大提高用户信息利用的便捷性和效率。第四，根据抽取出的国家、主题、议员等实体的语义关系构建知识关联网络，挖掘数据的潜在价值，可以为政府和相关部门提供决策支持。

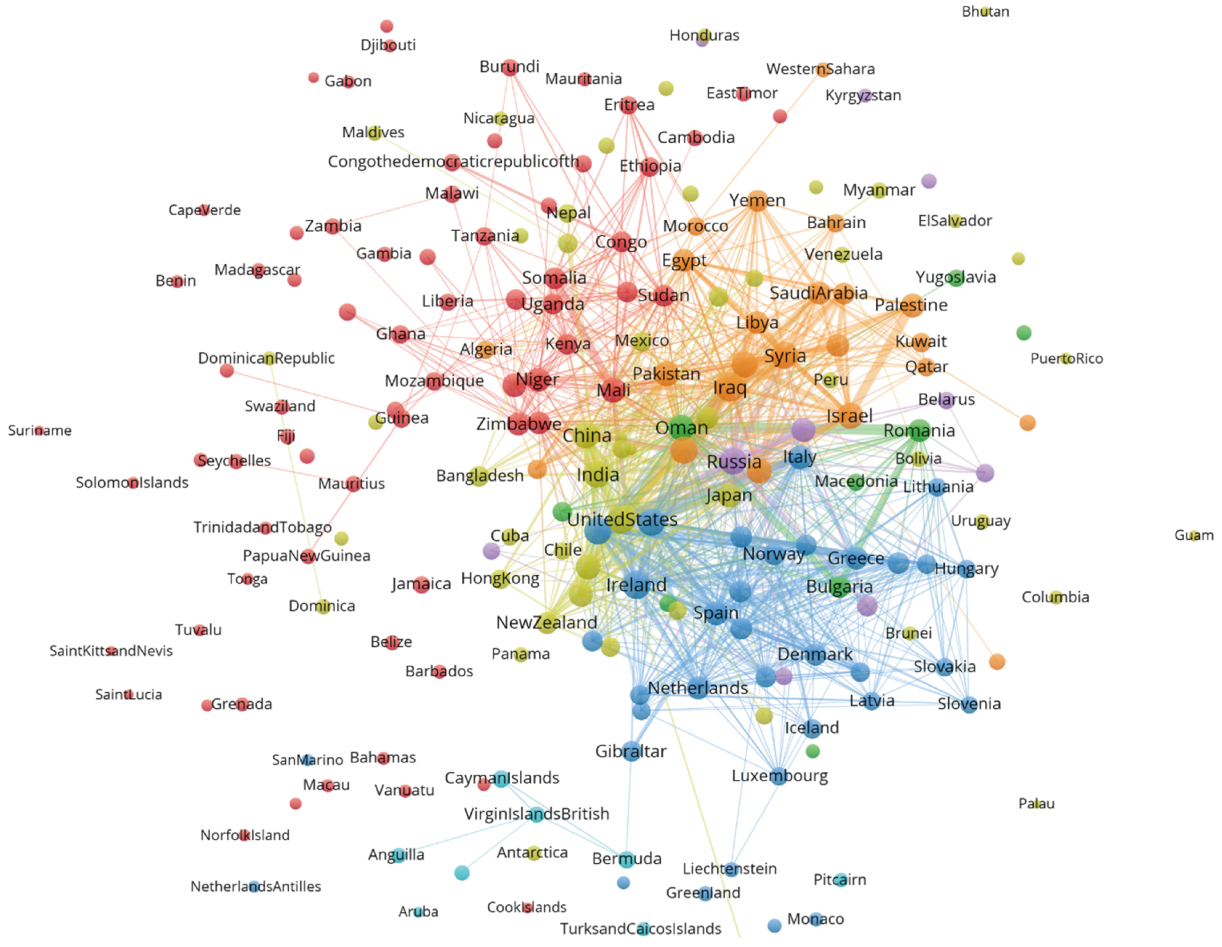

图 5 国家关联网络可视化示例

（3）系统改进方向

目前该系统的设计仅仅是对多国议会数据集利用的初步探讨，该数据集蕴含的价值有待进一步的发掘。在未来的研究中还需进一步改进数据集的组织形式，丰富数据集内容，完善功能设计，改良用户界面，提升用户体验，帮助用户全面、高效地利用议会数据。

3.2 启示

在理论层面上，议会大数据平台的开发人员从定量的视角出发，采取科学计量的手段，对议会数据进行主题抽取与关联网络计算与可视化，揭示议会文本主题分布规律、内在联系及各国关心的议题间的关联。这将有助于理解国家之间关系的特征与内涵，为国际关系研究

doi:10.3772/j.issn.2095-915x.2020.02.004

提供了一个新的定量视角和方法论的补充。

在现实层面上，议会大数据平台中存储的历史发言记录可以针对议员进行历史跟踪，用户可以以此了解议员的背景、动机、历史态度等信息，从而形成多国议会“政治基因库”。议会大数据平台提供的海量原生信息可以满足用户紧跟国际热点、及时了解国外政治态度的变化的需求，对外交政策制定、企业市场环境分析等都有重要意义。

4 结论

本研究综合应用信息检索、自然语言处理等方法和技术，构建了英国、美国、俄罗斯、加拿大和欧盟 5 个国家或组织的议会数据集，在此基础上开发了议会大数据平台。详细介绍了数据集的收集渠道、构成、数据处理过程，重点介绍了数据平台的检索和下载功能，并以国家关联网络为例对数据进行分析和展示。最后，对系统的功能进行了评价，并指出了系统的优势和改进方向。该数据集和平台对科研人员和政府机构相关工作的开展具有较高的实际价值。

在后续的研究中，本团队将会进一步扩展数据集，收集更多国家的开放议会数据，形成更加完备的议会数据集。此外，我们将更深入的对议会文本数据进行分析，挖掘更多的潜在价值，提供功能更加齐全的议会数据服务平台。

参考文献

[1] Kullaa R. The European parliament and the UK parliament: a relationship in foreign policy[J]. Parliamentary History, 2016, 35(1): 67-79.

[2] Kai O, Spencer A. Thinking alike? salience and metaphor analysis as cognitive approaches to foreign policy analysis[J]. Foreign Policy Analysis, 2013, 9(1): 39-56.

[3] Jenichen A. Human rights vs. security? The EU’s secular international identity from a transatlantic perspective[C]. Proceedings of the 14th European Union Studies Association International Biennial Conference, 2015.

[4] Lee S. Construction of China and India’s national interests: The Tibet question[D]. University of Westminster, 2016.

[5] Barnett G A, Xu W W, Chu J, et al. Measuring international relations in social media conversations[J]. Government Information Quarterly, 2017, 34(1): 37-44.

[6] Gravelle T B, Reifler J, Scotto T J. The structure of foreign policy attitudes in transatlantic perspective: comparing the United States, United Kingdom, France and Germany[J]. European Journal of Political Research, 2017, 56(1): 757-776.

[7] Ihalainen P, Matikainen S. The British parliament and foreign policy in the 20th century: towards increasing parliamentarisation?[J]. Parliamentary History, 2016, 35(1): 1-14.

[8] OpeningParliament.org. Declaration on Parliamentary Openness[EB/OL]. [2019-12-20]. http://www.openingpa- rliament.org/declaration.

[9] Berntzen L, El-gazzar R, Johannessen M R. Parliamentary open big data: a case study of the Norwegian parliament’s open data platform[C]. Proceedings of the 15th European, Mediterranean, and Middle Eastern Conference on Information Systems. 2018: 91-105.

[10] Berntzen L, Johannessen M R, Andersen K N, et al. Parliamentary open data in Scandinavia[J]. Computers, 2019(8):65.

[11] Seaton J. The Scottish parliament and e-democracy[J]. Aslib Proceedings, 2005, 57(4): 333–337.

[12] Fernandez-luna J M, Huete J F, Gomez M, et al. Development of the XML digital library from the

parliament of Andalucía for intelligent structured retrieval[C]. Proceedings of the International Symposium on Methodologies for Intelligent Systems. 2008: 417-423.

[13] Van Aggelen A, Hollink L, Kemman M, et al. The debates of the European parliament as linked open data[J]. Semantic Web, 2017, 8(2): 271-281.

[14] Faria C, Rehbein M. Open parliament policy applied to the Brazilian chamber of deputies[J]. The Journal of Legislative Studies, 2016, 22(4): 559-578.

[15] Beelen K, Thijm T A, Cochrane C, et al. Digitization of the Canadian parliamentary debates[J]. Canadian Journal of Political Science/Revue canadienne de science politique, 2017, 50(3): 849-864.

[16] Rohit S V, Singh N. Analysis of speeches in Indian parliamentary debates[J]. arXiv:1808.06834, 2018.

[17] De Campos L M, Fernandez-luna J M, Huete J F, et al. Positive unlabeled learning for building recommender systems in a parliamentary setting[J]. Information Sciences, 2018(433-434): 221-232.

[18] Zhang J, Spirling A, DANESCU-NICULESCU-MIZIL C. Asking too much? the rhetorical role of questions in political discourse[C]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1558-1572.

[19] Rheault L, Beelen K, Cochrane C, et al. Measuring emotion in parliamentary debates with automated textual analysis[J]. Plos one, 2016, 11(12): e0168843.

[20] Przybyla P, Teisseyre P. Analysing utterances in Polish parliament to predict speaker’s background[J]. Journal of Quantitative Linguistics, 2014, 21(4): 350-376.

[21] Kaptein R, Marx M. Focused retrieval and result aggregation with political data[J]. Information Retrieval, 2010, 13(5): 412-433.

[22] Jungermann F, Morik K. Enhanced Services for Targeted Information Retrieval by Event Extraction and Data Mining[C]. Proceedings of the 13th International Conference on Natural Language and Information Systems: Application of Natural Language to Information Systems, 2008: 335-336.

[23] Shadbolt N, O’hara K, Berners-lee T, et al. Linked open government data: lessons from Data.gov.uk[J]. IEEE Intelligent Systems, 2012, 27(3):16-24.

[24] Bulut A T. Measuring political agenda setting and representation in Turkey: introducing a new approach and data set[J]. Party Politics, 2017, 23(6): 717-730.

[25] Galiotou E, Fragkou P. Applying linked data technologies to Greek open government data: a case study[J]. Procedia-social and behavioral sciences, 2013(73): 479-486.

[26] Gielissen T, Marx M. The design of PoliDocs: a web information system for the disclosure of Dutch parliamentary publications[C]. Proceedings of the 6th International Workshop on Web Information Systems Modeling, WISM, 2009.

[27] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 8(10): P10008.