

服务于资源建设的会议预报系统构建*

□ 杨代庆 / 中国科学技术信息研究所 北京 100038

中国科学院文献情报中心 北京 100190

王星 常迎春 / 中国科学技术信息研究所 北京 100038

摘要: 文章通过对国外1000多个有会议信息的学协会网站进行分析,制定了会议信息采集策略,以及会议信息元数据集,并给出了人工规范会议信息的原则以及机器辅助抽取会议信息元数据的方法及步骤。最后,介绍了开发的会议预报系统的主要功能。

关键词: 会议预报, 信息采集, 信息筛选, 会议信息系统

DOI: 10.3772/j.issn.1673—2286.2012.08.004

1 介绍

随着科研人员对科技文献的需求不断增长,对国外会议录等灰色文献产生的需求日益强烈。目前,国外会议文献的保障程度是不均衡的,大多数的图书馆和科技信息机构比较重视国外图书、期刊等文献品种的采购,而对价格高、采集难度大的会议录等文献的采购重视不够,使这些文献品种的保障水平明显低于期刊等常规文献品种。以国家科技图书文献中心为例,目前订购国外期刊近17000种,而采购国外会议、科技报告文献仅4000余种,数量和保障水平差异明显。

因此,调查研究会议录等特种文献的出版状况和发行规律,加强对会议的学术评价和资源建设力度,对图书馆丰富和完善馆藏文献

资源品种结构、进一步提高文献保障能力具有重要价值。

目前,每年国际召开的科技会议达数万次,正式发行的各种专业会议文献过万种。一些历史悠久的国际性学术组织,如美国石油工程师学会(SPE)、地球物理勘探学会(SEG)、美国电气和电子工程师协会(IEEE)、美国航空航天学会(AIAA)、国际光学工程学会(SPIE)等在国际科技学术研究领域有巨大影响的学协会举行的学术会议对科技发展起着重要作用。正是学术会议文献在时效性、新颖性、权威性等方面的特点,决定了会议文献不可替代的地位。然而对于图书馆资源建设来说,会议文献的采集却不可避免地会遇到一些问题:

- 出版周期灵活

会议文献一般一年一期,较少一年两期,有的两年一期,甚至不定期;

- 出版形式多样

有预印本、会议录、系列丛书、选集、会议专刊、会刊等等,来源渠道复杂不容易获得。

正因如此,急需通过对国外学协会会议召开信息的采集、监控来帮助图书馆进行会议文献资源建设。目前,国内对会议信息、尤其是将要召开的会议信息的集中揭示系统较少,其中成规模的有教育部科技发展中心主办的“中国学术会议在线”平台^[1],但主要揭示了在中国境内召开的会议信息,对国外学会主办的会议涉及较少。国外学会会议信息报道已有诸如ISI Proceedings^[2]、OCLC Proceedings^[3]、技术会议信息中心

* 本文系中国科学技术信息研究所预研项目“一种并行高维主题聚类方法的研究与实现”(编号:YY-201107)研究成果之一。

[4]等网站或工具,但这些资源或者只收录会议录信息及已召开会议的简单信息,对于将要召开的会议预报缺乏相关详细报道;或者数据库为商业数据库,不能免费提供服务。此外,从图书馆会议文献建设角度来说,这些数据库中包含的会议信息基本已包含在图书馆的既有遴选监控系统中,所以对于图书馆提高会议文献覆盖度帮助有限。为此,直接对国外学协会网站进行会议预报信息采集、监控,成为促进会议文献建设的重要渠道。

2 会议信息的采集

通过调研,收集到约4000个国外学协会网站信息,通过对这些网站进行分析,发现有会议预报信息的网站为1647个。1647个学协会网站中,有的学协会会议预报信息以固定栏目呈现,但也有相当多的会议预报信息是以不固定的形式呈现的,例如直接在页面中嵌入链接,或在某个相关页面中提供几行文字,或直接是一个会议召开信息的PDF文件等。在这些学协会网站中,大部分网站都没有提供对以往会议的回溯检索查询,而只有近期的会议预报。

国外学协会的会议信息数量相差巨大,例如:电气电子工程师学会(IEEE)提供的会议信息约6000条,而美国神经化学学会(ASN)仅提供会议信息3条。

2.1 会议信息采集问题

国外会议信息采集过程中主要面临如下问题:

(1) 需要采集的网站数量众多

信息采集过程中首先面临的问题是定向采集量大。如果全部人工进行采集,后续持续性的监控采集工作量过大。而用机器采集,需要分析网站页面源代码,编制采集程序或定制采集工具,初始工作量巨大。

(2) 需要采集的网站结构各异

各个网站的会议信息发布栏目名称及深度不固定,呈现方式多样。例如:有的采用表格形式,有的采用无结构文本,而有的采用附件形式。并且各个网站提供的会议信息元数据相差巨大,例如:AIAA从日程到联系方式、会议相关的展示会、培训等信息都提供,而有的只提供会议名称及时间、地点。

(3) 需要采集的网站更新速度不一

各个网站的更新速度、更新量差别很大,若监控时的定向采集频率过大,会增加工作量;而采集频率过小,又不能及时采集到会议信息,导致会议信息报道延迟甚至延误。

2.2 采集策略

针对上述问题,对所有需定向采集监控的网站进行了人工筛选,筛选出适宜于人工采集的网站和适宜于机器自动采集的网站;针对机器采集网站,又划分为采集工具定制采集和编制程序采集两类。

- 在确定人工采集类的网站时,采用的规则是:会议信息数量较少;会议信息结构不规范;会议信息以附件形式发布。例如美国化学协会AOCs的会议预报页面。

- 确定采用采集工具定制采集的策略主要是:页面结构比较简单,元数据结构固定,便于元数据位置的定位,翻页标识明确。例如空气与废物管理协会(AWMA)的会议预报页面。

- 确定使用自行开发工具进行采集的策略是:该学会网站会议信息较多,且格式变化较多,值得专门开发下载工具。例如:电气电子工程师学会(IEEE),每个会议都有唯一的编号,并且格式规整固定。



图1 美国化学协会AOCs会议页面

Date	Event Type	Event Title	Event Location
Aug 26, 2010	Webinar	Carbon: Sustainability, Strategy, and Solutions	online
Aug 30 - Sep 2, 2010	Conference	Power Plant Air Pollutant Control MEQA Symposium	Baltimore, MD
Sep 12 - 16, 2010	Conference	15th IUAPPA World Clean Air Congress: Achieving Environmental Sustainability in a Resource Hungry World	Vancouver, BC, Canada
Sep 28, 2010	Course	AIR-207: Design Considerations for the Mitigation of Vapor Intrusion	Chicago, IL
Sep 28, 2010	Course	AIR-206: Sampling and Analysis Methods for Vapor Intrusion	Chicago, IL
Sep 28, 2010	Course	AIR-268: Data Evaluation for Vapor Intrusion Studies	Chicago, IL
Sep 28, 2010	Course	AIR-274: Vapor Intrusion Pathway Modeling: Development and Application	Chicago, IL
Sep 29 - 30, 2010	Conference	Vapor Intrusion 2010	Chicago, IL
Oct 7, 2010	Webinar	Carbon: A Rising Climate Lifts All Participants	online
Nov 2 - 4, 2010	Conference	Symposium on Air Quality Measurement Methods and Technology	Los Angeles, CA
Dec 7 - 9, 2010	Conference	35th Annual A&WMA/EPA Information Exchange	Research Triangle Park, North Carolina
Mar 14 - 16, 2011	Conference	Opportunities in Air Quality-Climate Research in Asian-Pacific Region	Kona, Hawaii
May 10 - 12, 2011	Conference	International Conference on Thermal Treatment Technologies & Hazardous Waste Combustors	Jacksonville, FL
Jun 21 - 24, 2011	Conference	104th Annual Conference and Exhibition	Orlando, FL
Jun 19 - 22, 2012	Conference	105th Annual Conference and Exhibition	San Antonio, TX

图2 空气与废物管理协会 (AWMA) 会议页面

3 会议信息的规范方法

3.1 元数据集

不同的学会网站提供不同的会议信息元数据，且元数据的呈现方式也不相同。为此，在会议预报建设中，需要制定会议采集信息核心元数据集和扩展元数据集，制定的元数据集描述如下：

- 核心元数据集包括：会议名称、会议时间、会议地点、举办单位；
- 扩展元数据集有：会议网址、会议范围、会议主题、联系方

式、会议附件、会议赞助商、会议年、会议界次、会议级别、会议关键词、会议语种、会议主办国、会议日程、会议交通、会议培训信息、会议主持人。

3.2 规范方法

对采集到的会议预报信息需要进行规范组织后，才能进入系统服务。规范时，主要采用了人工规范方法，对于部分会议信息相对固定的学协会，采取了机器辅助下的人工规范方法。

(1) 人工规范内容及原则

1) 学会信息的“等级名称”规范方式为：从大地名到小地名，中间用加号连接（世界、国际、洲名称、东欧、南亚等名称、国家、国家名称+州/省等名称）；

例如：香港地区会议：亚洲+中国+香港；

国际、亚洲会议：国际+亚洲；

2) 会议学科分类为单分类，选择最为贴切的一个；

3) 人工挑选2-3个词作为会议主题词；

4) 时间规范格式可以为如下中的任意一种：英文月-数字日-数字年；数字日-英文月-数字年；数字日-数字月-数字年；数字月-数字日-数字年；

5) 使用国别规范表、语种规范表、分类规范表对国别、语种和分类进行规范。

(2) 机器辅助的会议信息规范方法

采集到的数据为HTML格式，通过对采集到的文档进行清洗后（去掉广告、导航等信息），再利用机器自动抽取及识别文档的会议元数据以及元数据关系，得到诸如会议名称、会议地点、会议时间、主办者、会议等级等元数据以及会议与会议时间之间的“召开时间”关系，会议与会议地点之间的“召开地点关系”。机器辅助处理过的元数据，再通过人工进行检查，以确定其信息的准确性。

在机器辅助规范整合过程主要需要解决的问题有：如何在保证召回率的基础上以尽可能高的准确率对会议预报元数据及关系进行识别抽取；抽取规则如何定制等。具体实施操作为：

- 1) 会议详细页面抓取
- 2) 页面清洗：

① 通过基于可视区块分析算法VIPS^[5] (Vision-based page segmentation) 对采集到的页面进行可视化区块分析, 得到页面的可视区块属性及嵌套结构(见图4)。

② 采用竞争型神经网络SOFM (Self-Organizing Feature Maps) 实现内容页面的主可视区块提取(会议信息部分)。

本文中提取的页面区块特征为: 可视区块(VB)宽度, VB高度, 左上角坐标, 右下角坐标。

③ 采用链接文字比作为启发规则, 对得到的获胜区块中的内容进行再次清洗, 得到将要抽取会议

元数据的文本串。

3) 实体及关系抽取

实体及关系抽取的方法及工具有很多, 大致分为基于规则、基于统计或混合策略的抽取方法。其中基于规则类的工具有GATE^[6], 基于统计的有Stanford的基于CRF的实体抽取工具^[7]。本文在对清洗后的文本串进行实体关系抽取时使用了Thomson Reuters提供的Calais Web Service^[8]来实现。

通过对图3中的AIAA会议预报页面进行清洗, 采用CALAIS抽取元数据及关系后, 并形成图5的RDF过程如下:

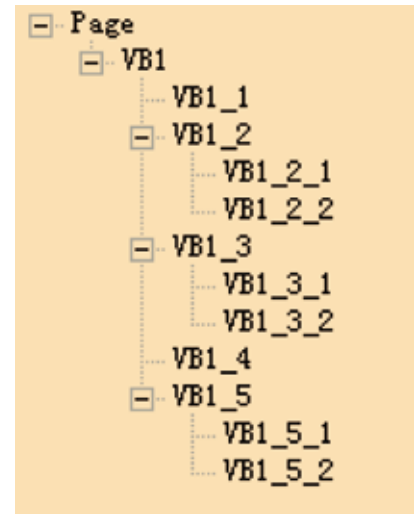


图4 图3中根据VIPS算法的区块结构

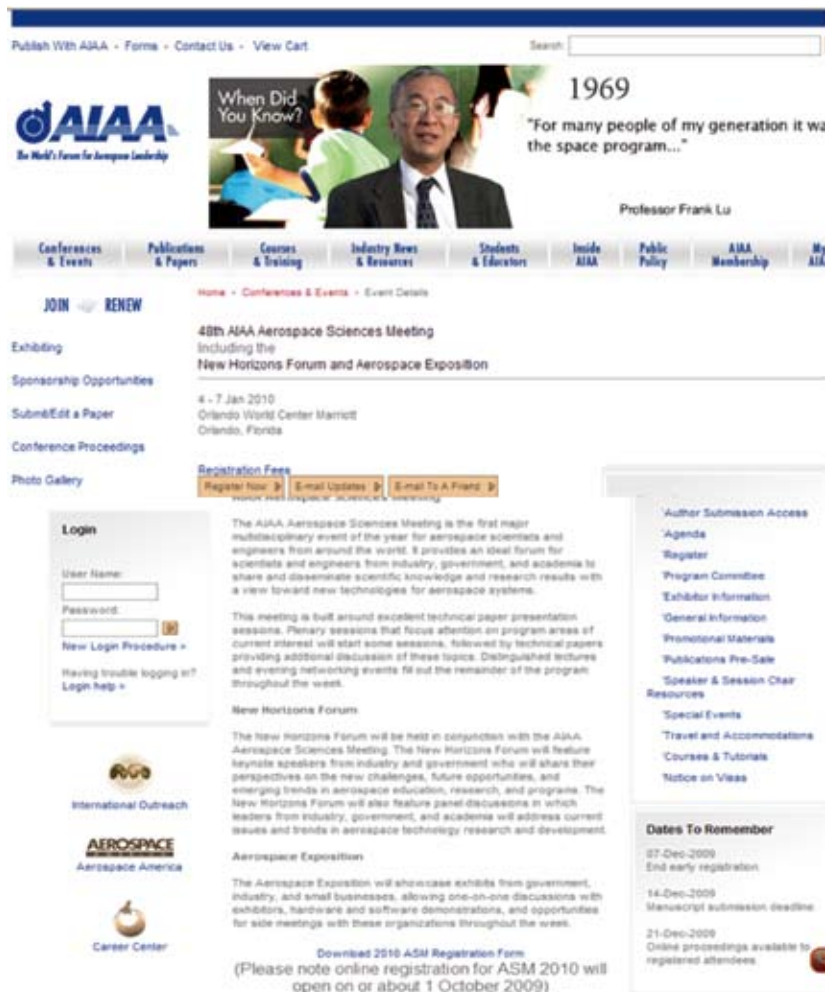


图3 AIAA会议预报页面

4 在线会议预报系统

对国外学协会会议信息基础数据进行采集、规范和处理, 对大量会议信息的标记, 对会议信息的管理、检索、排序等都需要开发一套集数据采集、规范、发布于一体的平台系统, 以保证数据处理高效性和发布的及时性。会议预报系统平台的主要功能包括:

(1) 基础数据的采集和整合功能

对国外学协会及其会议信息进行人工和机器辅助相结合的整理、汇总、规范、查重、合并、比对、标记和整合, 形成会议信息数据库。

(2) 分面导航及馆藏链接

按照会议级别、学科分类、会议召开时间等对会议信息数据进行分面导航, 并将已经召开过的会议与我馆已收藏的会议文献进行挂接(例如IEEE系列会议)。

(3) 数据检索查询

平台系统整合的国外学协会会



图5 图3中的会议抽取元数据后的RDF图

会议信息主要包括了数量庞大的基本数据，既可用于会议文献的查找和获取，也可用于会议文献的评价研究。因此，平台系统提供了灵活的数据检索功能，以求多方位地发挥数据的价值，实现更大程度的数据共享。

5 结语

学术会议信息的监控具有来源点多、每个来源点信息量有限，但总量大、时限性强等特点，仅靠人力难以完成。一方面要进一步研究使用机器进行自动抓取及元数据辅助抽取的方法，进一步减少人工操作；另一方面同一个会议信息随着时间的临近，是变化动态的信息（例如刚开始的会议仅是会议通知，而后再增加会议日程等），这些是需要进一步研究并解决的问题。



图6 会议预报系统浏览页面



图7 会议预报详细页面

参考文献

- [1] 中国学术会议在线[EB/OL]. [2011-05-10]. <http://www.meeting.edu.cn>.
[2] ISI Proceedings [EB/OL]. [2012-03-10]. <http://isiknowledge.com>.
[3] OCLC Proceedings [EB/OL]. [2012-03-15]. <http://firstsearch.oclc.org/FSI>.
[4] 技术会议信息中心[EB/OL]. [2012-03-15]. <http://www.techexpo.com/events/evnts-p1.html>.
[5] VIPS: a Vision-based Page Segmentation Algorithm [EB/OL]. [2011-07-22]. <http://research.microsoft.com/apps/pubs/default.aspx?id=70027>.
[6] GATE [EB/OL]. [2011-08-18]. <http://gate.ac.uk/>.
[7] Named Entity Recognition (NER) and Information Extraction (IE) [EB/OL]. [2011-08-18]. <http://nlp.stanford.edu/ner/index.shtml>.
[8] OpenCalais [EB/OL]. [2011-08-25]. <http://www.opencalais.com/>.

作者简介

杨代庆 (1975-), 男, 中国科学技术信息研究所高级工程师, 中国科学院文献情报中心博士生。E-mail: yangdq@istic.ac.cn

王星 (1977-), 男, 工程师, 中国科学技术信息研究所。

常迎春 (1982-), 女, 工程师。

Information System Construction of Meeting Calendar Served for Resource Construction of Library

Yang Daiqing / Institute of Scientific & Technical Information of China, Beijing, 100038

/ National Science Library, Chinese Academy of Sciences, Beijing, 100190

Wang Xing, Chang Yingchun / Institute of Scientific & Technical Information of China, Beijing, 100038

Abstract: In this paper, the collection strategies of meeting information on the web more than 1,000 academic associations and the metadata set for the meeting are given. Meanwhile, the machine-aided cleaning methods of collected meeting information are guided.

Keywords: Meeting calendar, Information acquisition, Information filtering, Meeting information system

(收稿日期: 2012-04-18)